① 21.1.08
מערכות
לומדות

סוף התרגילים מאחורינו.

הוא מתכנן עוד 7-8 שיעורים לאחר מכן שיעורי חזרה.

לא יהיה מבחן בסמסטר. כל נושא הבחינה עדיין בחשיבה. <u>כנראה</u>

שמגני יכתוב את השאלות על החומר שלו והמרצה יכתוב את השאלות

על החומר של ההרצאות. בתקווה השאלות לא יהיו קשות.

אם תשלחו לי אי מייל ל- dinazil@notes-heaven.com

עם בקשה לחזרה על חומר בשיעורי החזרה

המטרה שלנו היום היא לתת מבט מלמעלה על הדברים.

מאחורי למידה חישובית יושבת התפלגות. יש את מרחב המדגם X

של דוגמאות x ושאיתן תווית y מתוך מרחב תוויות Y.

אומרים שיש התפלגות D מעל $X \times Y$ וממנה לוקחים הדוגמאות.

אם הייתה נתונה לנו ההתפלגות אז לא ממש הייתה בעיית למידה.

נניח ש- $X=\{a_1,\dots,a_n\}$; $Y=\{-1,1\}$.

ונניח שיש לנו טבלה של הסתברויות:

$y \backslash x$	a_1	a_2	a_3	---	a_n
1					
-1					

$\Sigma p_{i1} = p(y=1)$

$\Sigma p_{i,-1} = p(y=-1)$

הסתברות מותנית - $p(x|y)=\frac{p(x,y)}{p(y)}$. זה נותן לנו דרך לנרמל את ההתפלגות כך שסכום כל שורה יהיה 1:

$$\sum_x p(x|y) = \sum_x \frac{p(x,y)}{p(y)} = \frac{\Sigma p(x,y)}{p(y)} = \frac{p(y)}{p(y)} = 1$$

בדרך דומה אפשר לנרמל כך שהסכום בכל עמודה יהיה 1 - $p(y|x)$.

עכשיו $p(x|y=1)$ זו פונקציית הסתברות כי הסכום שלה הוא 1.

זה אומר מה הסיכוי של דוגמה x בהינתן שידוע לנו שהדוגמה היא חיובית.

בעיית ההחלטה שלנו היא למעשה להחליט מה יותר גדול - $p(y=1|x)$ או $p(y=-1|x)$ ואם יש לנו טבלה כזו אין בעיה!

נוסחת בייס - $p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$

לפי הנוסחה הזאת רצוי להשוות את $p(y=-1|x) \;?\; p(y=1|x)$

מספיק להשוות את -

$$p(x|y=-1)\,p(y=-1) \;?\; p(x|y=1)\,p(y=1)$$

בד"כ בבעיות למידה יודעים את $p(y=1)$ ו- $p(y=-1)$ ואם לא יודעים אז בד"כ מניחים שהם שווים - כלומר ההתפלגות אחידה.

אז מה שצריך להשוות זה את הגודל שנקרא נראות -

$$p(x|y=-1) \;?\; p(x|y=1)$$

כפי שכבר אמרנו, ט זה כמובן לא נחוץ אם נתונה לנו ההתפלגות.

אז כל מה שעושים בתחום הלמידה החישובית הוא לתת אלגוריתמים למקרה שלא נתונה לנו ההתפלגות.

PAC הוא המודל היחיד בתחום הלמידה החישובית שבו לא עושים שום הנחה על ההתפלגות ולא משערכים אותה. (בשאר הדברים שלמדנו, לא אמרנו את זה במפורש אבל יש הנחה של ההתפלגות (נורמלית))

מודל PAC אומר את הדבר הבא: יש מחלקת היפותזות C ונותנים דיוק ε וביטחון δ ועל סמך הדוגמאות שניתנו לנו אנחנו רוצים למצוא היפותזה h שהשגיאה שלה מקיימת

$$P(err(h) \le \varepsilon) \ge 1-\delta$$

מספר הדוגמאות בשביל להשיג את זה הוא $m(\varepsilon,\delta)$ והוא תלוי אך ורק ב-ε וב-δ ולא בהתפלגות.

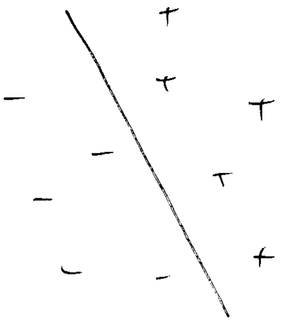
לא תמיד אפשר להשתמש ב- PAC. אם המחלקה C סופית אז צריך $m \ge \frac{1}{\varepsilon}\log\frac{|C|}{\delta}$ ואז אם C אינסופית אבל ה- dimVC שלה סופי d אז גם אפשר ללמוד! $m \ge \frac{1}{\varepsilon}\log\frac{1}{\delta} + \frac{d}{\varepsilon}\log\frac{1}{\varepsilon}$.

אם מימד VC של C אינסופי אז אי אפשר ללמוד PAC -

אי אפשר ללמוד ללא שיערוך של ההתפלגות. יש רק משפחה מצומצמת של בעיות שבהן לא צריך לשערך התפלגות. אבל כשזה כן קורה, זה קסום!

למשל, למחלקת ההיפותזות של מפרידים לינאריים ($f(x)=sgn(w^t x - b)$) יש מימד VC סופי ולכן אפשר ללמוד אותה.

(2)



אפשר לראות מהחסם של m שאם המימד VC קטן יותר אז אפשר לקחת פחות דוגמאות כדי ללמוד לפי ε ו-δ נתונים.

אז במקרה של מפרידים נסתכל על מפרידים עם שוליים בגודל γ - C_γ. אפשר לראות ש- $VCD(C_\gamma)=\min\{\frac{1}{\gamma^2},n\}+1$ אבל $VCD(C)=n+1$

לכן אם γ גדול הקטנו את המימד VC וזה מאפשר לנו לקחת פחות דוגמאות. אז גם המטרה של אלגוריתם SVM היא למצוא מפריד עם שוליים כמה שיותר גדולים.

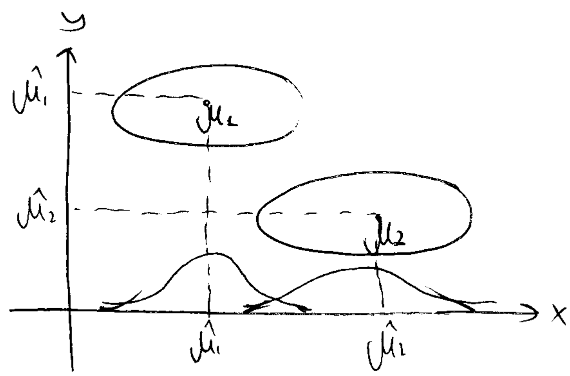
אפשר לסדר את העניינים כך ש- $\gamma=\frac{1}{\|w\|}$ ולכן יש לנו בעיית אופטימיזציה: $\min \frac{1}{2}\|w\| + \lambda\sum\varepsilon_i$ s.t $y_i(w^t x_i - b)\geq 1-\varepsilon_i$

הבעיה הדואלית היא

$$\max \sum\alpha_i - \frac{1}{2}\sum\alpha_i\alpha_j y_i y_j x_i^t x_j$$
s.t
$$0\leq\alpha_i\leq\lambda$$
$$\sum\alpha_i y_i=0$$

למה עברנו לבעיה הדואלית? כי שם אפשר להשתמש ב- Kernel trick ועכשיו אפילו אם במרחב הנמוך הנקודות לא ניתנות להפרדה ע"י יריעה לא לינארית אז אפשר לעבור למרחב גבוה ובו אולי אפשר לעשות הפרדה לינארית.

אם מניחים שהנקודות שנתונות הם מהתפלגות נורמלית אז יש דרך בעזרת אלגברה לינארית למצוא על-מישור מפריד.



נניח שיש לנו שתי קבוצות - דוגמאות חיוביות עם מרכז μ_1 ודוגמאות שליליות עם מרכז μ_2.

אנחנו רוצים למצוא מישור w שעליו נטיל את הדוגמאות רק שנוכל להפריד אותן - אז אנחנו רוצים שהמרכזים יהיו רחוקים כמה שיותר זה מזה אחרי ההטלה. למשל ציר ה-x בציור מקיים את זה. אבל ציר ה-x לא מפריד טוב כי יש לנו שם נקודות חופפות. אם נסתכל דווקא על ציר ה-y במקרה זה אמנם המרחק בין המרכזים לא הכי גדולים אבל אין חפיפה של הנקודות

אז מה שאנחנו רוצים לפתור זה את בעיית האופטימיזציה $\max \frac{\hat{\mu}_1-\hat{\mu}_2}{\hat{S}_1^2+\hat{S}_2^2}$
כאשר S_i^2 זה איזה מדד לפיזור. ואת זה אפשר להעביר לבעיה מהצורה
$Aw=\lambda Bw$ כאשר w ו"ע של $B^{-1}A$.

גם ב-PCA יש הנחה של התפלגות נורמלית. הרעיון היה שיש מטריצה שבה יש יתירות ואנחנו רוצים "לרווח" אותה. את $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ רוצים להעביר למרחב קטן יותר $\begin{pmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_q \end{pmatrix}$ ככה שאין תלות סטטיסטית בין הכניסות.
מטריצת ה covariance של X היא לא אלכסונית. לאחר ההתמרה מטריצת ה covariance של $\hat{X}$ היא אלכסונית. כעת, מטריצת Covariance אלכסונית זה שקול לאי-תלות סטטיסטית רק אם ההתפלגות הם נורמליות.

אם אנחנו מניחים ש- $X\sim N(\mu,E)$ אז

$$p(x)=\frac{1}{\sqrt{2\pi|E|}}e^{-(x-\mu)^tE^{-1}(x-\mu)} \qquad \theta=(\mu,E)$$

אם נתונים דוגמאות חיוביות ודוגמאות שליליות אז משערים את ההתפלגות הנורמלית של החיוביות ואת ההתפלגות הנורמלית של השליליות:
אם נניח $y=-1$; $S=\{x_1,\dots,x_m\}$ iid אז

$$\max_\theta p(S|\theta)=\max_\theta\prod_i p(x_i|\theta)$$

$$\iff \max\log p(S|\theta)=\max\sum\log p(x_i|\theta)$$

אז $\arg\max_\theta p(S|\theta)=\arg\max_\theta\sum\log p(x_i|\theta)$ וזו בעיה שאפשר לפתור.
הבעיה היא שאין כמעט בעיות אידאליות שעומדות בתנאים הדרושים.

נניח שרוצים למצוא משטח מפריד בין שתי התפלגויות נורמליות ששונות בתוחלת

$$p(x|h_1)=N(\mu_1,E)$$
$$p(x|h_2)=N(\mu_2,E)$$

המשטח המפריד הוא מקום ה-x-ים שנמצאים על הגבול $p(h_1|x)-p(h_2|x)=0$.
אם $p(h_1)=p(h_2)$ אז המשטח המפריד הוא $p(x|h_1)-p(x|h_2)=0$.

גישה יותר מתקדמת מניחה שההתפלגות היא קומבינציה לינארית של התפלגויות יפניות. זה נשמע קצת יותר מציאותי מההנחה שההתפלגות נורמלית.

נניח ש- $x \in \mathbb{R}^n$. ו- $Y=\{y_1,\dots,y_k\}$. אז

$$p(x)=\sum_y p(x|y)p(y)=\sum_{j=1}^{k}\underbrace{p(x_1,\dots,x_n|y=y_j)}_{\text{נניח שכל התפלגות ידועה.}}p(y=y_j)$$

כדי לשערך את $p(x)$ צריך לשערך את כל ה $p(x_1,\dots,x_n|y=y_j)$ ואת $p(y=y_j)$.

אם ה-x_i-ים בלתי תלויים אז $p(x_1,\dots,x_n|y)=p(x_1|y)\cdots p(x_n|y)$

<u>דוגמא:</u> יש שני מטבעות. ראשון סיכוי p ל-0 ושני סיכוי q ל-0 .. אנחנו בוחרים מטבע בסיכוי λ לקבל את הראשון, ואז מטילים אותו 3 פעמים.

נניח שנתונות לנו תוצאות של ניסויים כאלה והמטרה היא למצוא את λ, q, p.

יש כאן משתנה חבוי והוא המטבע שנבחר באותו ניסוי $y_i\in\{1,2\}$.

נסמן ב-x_i תוצאת ניסוי - כלומר תוצאת 3 הטלות כנ"ל. אז

$$p(x_i)=p(x_i|y_i=1)\overbrace{p(y_i=1)}^{\lambda}+p(x_i|y_i=2)\overbrace{p(y_i=2)}^{1-\lambda}$$

נסמן ב-m_i את מספר האפסים שיש ב-x_i. אז

$$p(x_i)=\lambda p^{m_i}(1-p)^{3-m_i}+(1-\lambda)q^{m_i}(1-q)^{3-m_i}$$

איך מגדירים את זה כבעיית אופטימיזציה? אם $\theta=(p,q,\lambda)$ ו- $S=(x_1,\dots,x_m)$ אז מה שאנחנו רוצים הוא

$$\max_\theta \log p(S|\theta)=\max_\theta \log\prod_i p(x_i|\theta)=$$
$$=\max_\theta \sum \log\left(\lambda p^{m_i}(1-p)^{3-m_i}+(1-\lambda)q^{m_i}(1-q)^{3-m_i}\right)$$

זאת בעיה <u>קשה</u>. לא מובטח לנו למצוא θ אופטימלי, אבל קיים אלגוריתם שמוצא θ שנותן אופטימום לוקאלי, לא גלובלי. אנחנו עוד נראה אותו.

אם x,y התפלגויות מגדירים את האנטרופיה היחסית ביניהן ע"י

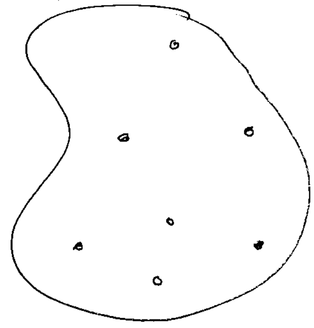
$$D(x\|y)=\sum x_i\log\frac{x_i}{y_i}$$

נסמן ב- $\hat{P}$ את ההתפלגות האמפירית של הדגימות. בדוגמא למעלה

$$\hat{P}_{i_1,i_2,i_3}=\frac{1}{m}\left|\{x_i=(i_1,i_2,i_3)\}\right|$$

אנחנו מחפשים את ההתפלגות הפרמטרית P הכי קרובה להתפלגות האמפירית. כלומר צריך לפתור את הבעיה $\min_P D(\hat{P}\|P)$

דוגמה: יש לנו דגימות של קינון ציפורים במצה בשדה.



ומה שאנחנו רוצים זה לעשות אינטרפולציה להתפלגות - למצוא את ההסתברות לקינון ציפורים בכל נקודה בשדה.

דוגמה קצת יותר פשוטה היא בהטלה של קוביה. נניח שאומרים לנו שתוחלת ההטלה היא 5 ושואלים מהי ההתפלגות. אז אנחנו צריכים למצוא את $P=\begin{pmatrix} p_1 \\ \vdots \\ p_6 \end{pmatrix}$

תחת האילוצים $p_i \ge 0$ $\quad \sum p_i = 1 \quad$ $\sum_{i=1}^{6} i p_i = 5$

יש 6 משתנים ורק 2 משוואות. זה לא מספיק כדי למצוא פתרון יחיד.

עקרון לפלאס אומר למצוא את הפתרון עם האנטרופיה המקסימלית.

כאשר האנטרופיה היא $H(p)=\sum p_i \log p_i$. מקסום האנטרופיה זו למעשה התקרבות להתפלגות אחידה.

אם רוצים להתקרב להתפלגות אחרת P_0 אז מה שאנחנו צריכים זה למצוא את המינימום של האנטרופיה היחסית $D(P_0 \| P)$ בכפוף לאילוצים.

לפעמים אפשר לעשות פקטוריזציה של פונקציית ההתפלגות למכפלה של התפלגויות על תת קבוצות של משתנים. וזה מביא אותנו למה שנקרא מודלים גרפיים.

(4) 28.1.08
אג'וומפור

הסקה בייסיאנית

נניח שיש לנו $X=\{x_1,\dots,x_n\}$. $p(x)$ פונקציית הסתברות

אם $p(x_i)\ge 0$ לכל $1\le i\le n$

ו- $\sum_{i=1}^{n} p(x_i)=1$

אפשר להסתכל על $p(x)$ כווקטור $p(x)=\begin{pmatrix} p(x_1)\\ \vdots \\ p(x_n)\end{pmatrix}$.

לכן וקטור הסתברות.

בצורה דומה אפשר להגדיר וקטור הסתברות מותנית $p(X|y)=\frac{p(X,y)}{p(y)}$

כלומר $p(X=x_i|Y=y_j)=\frac{p(X=x_i,Y=y_j)}{p(Y=y_j)}$ ו- $p(Y=y_j)=\sum_x p(X,Y=y_j)$

הנרמול ב- $p(y)$ נדרש כדי שהווקטור $p(X,Y=y_j)$ יהיה וקטור הסתברות. ואכן יוצא $1=\frac{p(y)}{p(y)}=\sum_{x_i}\frac{p(x_i,y)}{p(y)}=\sum_x p(X|y)$.

מכאן אנחנו מקבלים את

<u>נוסחת בייס</u>: $p(X|y)\,p(y)=p(X,y)=p(y|X)\,p(X)$

או

$$\underbrace{p(y|X)}_{\text{posterior}}=\frac{\overbrace{p(X|y)}^{\text{likelihood}}\ \overbrace{p(y)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}}$$

לא מייד ברור למה זה טוב. לכאורה לא אמרנו שום דבר עמוק אבל זה מעט נותן לנו מוטיבציה.

<u>דוגמה</u>: יש ערכה שמקבלת דגימת דם ונותנת תוצאה חיובית או שלילית.

הדם הוא או בריא או חולה סרטן $X=\{C,H\}$

הדגימה נותנת 2 תשובות $Y=\{+,-\}$.

נתונים הנתונים הבאים:

$p(+|C)=0.98$

$p(-|H)=0.97$

$p(C)=0.01$

המספרים האלה נראים מרשימים, אבל מה שמעניין אותנו הוא למעשה $p(C|+)$

וכאן מגיעה נוסחת בייס:

$$p(C|+)=\frac{p(+|C)\cdot p(C)}{p(+)}=$$

$$\underset{\text{נוסחת ההסתברות השלמה}}{=}\frac{p(+|C)\,p(C)}{p(+|C)\,p(C)+p(+|H)\,p(H)}=\frac{0.98\times 0.01}{0.98\times0.01+0.02\times0.99}=0.266$$

אז עם הערכה הזאת יהיה עדיף להסתמך על המזל...
מה שגורם לזה זה למשל שהסיכוי מלכתחילה להיות מסרטן קטן - ההתפלגות האפריורית - היא מאוד רחוקה מלהיות אחידה, ואז זה משפיע על ההתפלגות הפוסטריורית.

הרבה פעמים לא נתונה לנו ההתפלגות האפריורית ואז מניחים שהיא אחידה.
אז ראינו שלהנחה הזו יש השפעות מרחיקות לכת...

לפעמים טבעי למדוד את $p(x|y)$ ואת $p(y)$ ואז לחשב את $p(y|x)$.

<u>דוגמה:</u> X וקטור תסמינים
Y דיאגנוזה
אז $p(X|y)$ זה מה הסיכוי לתסמינים כאלה בהינתן המחלה.
זה משהו מאוד טבעי למדוד והוא יחסית קל. וגם קל למדוד את הסיכוי של המחלה $p(y)$.
קצת יותר מסובך למדוד את $p(y|x)$ - הסיכוי למחלה בהינתן התסמינים. אבל זה בדיוק משהו שרוצים לדעת כי אם Y יכול לקבל שני ערכים למשל כי אם בא בן אדם עם תסמין x היינו רוצים לדעת מה יותר גבוה - $p(Y=y_1|x)$ או $p(Y=y_2|x)$.
משהו לבדוק את זה זה כמו לבדוק

$$p(x|Y=y_1)p(Y=y_1) \overset{?}{} p(x|y=y_2)p(Y=y_2)$$

אם אנחנו מניחים שההתפלגות האפריורית היא אחידה מספיק רק להשוות את ה likelihood (ראו).

<u>MAP</u>

$$y^* = \arg\max_y p(y|x)$$

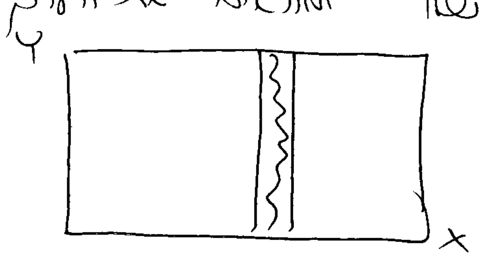
זה המקסימום הפוסטריורי - הנקודה שנותנת את המקסימום בהתפלגות הפוסטריורית.

נחשב:

$$y^* = \arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)p(y)}{p(x)} = \arg\max p(x|y)p(y) =$$

$$= \arg\max p(x,y)$$

אז אם יש לנו טבלה של ההתפלגות המשותפת פשוט קובעים
את X, מסתכלים בעמודה שלו ומוצאים את הערך המקסימלי.



הבעיה היא שלא
תמיד נתונה לנו ההתפלגות
המשותפת.

$$y^* = \arg\max_y p(x|y) \qquad \underline{ML}$$

<u>דוגמא</u> הטלת מטבע $X=\{0,1\}$ ומטבע הטיה $Y=q$

q זה פשוט הסיכוי לקבל 0.

$$p(X=0|Y=q) = q$$
$$p(X=1|Y=q) = 1-q$$

עכשיו נניח שיש לנו k הטלות ב"ת (iid) $X_1,\dots,X_k$
נרצה למצוא את ההטיה q שעבורה הסיכוי לקבל את ההטלות $(x_1,\dots,x_k)$ היא מקסימלי. כלומר רוצים למצוא את

$$\max_q p(x_1,\dots,x_k|q) = \max_q \prod_{i=1}^{k} p(x_i|q)$$

$$\equiv$$

נוסיף $\log$ משני האגפים ונקבל

$$\max_q \log p(x_1\dots x_k|q) = \max \sum_{i=1}^{k} \log p(x_i|q)$$

פה יש שוויון בין הארגומנטים שמביאים למקסימום ומסמנים ב: $\equiv$

יהי λ מספר הפעמים שקיבלנו $x_i=0$:

$$\lambda = |\{x_i=0 : i=1,\dots,k\}|$$

$$\Rightarrow q^* = \arg\max\{\lambda\log q + (k-\lambda)\log(1-q)\}$$

(משים ש- $p(x|q)=\begin{cases} q & x=0 \\ 1-q & x=1\end{cases}$ ואז עושים סכום).

נפתור את בעיית האופטימיזציה. נגזור את הפונקציה לפי q:

$$\frac{\lambda}{q} - \frac{k-\lambda}{1-q} = 0$$
$$\Rightarrow \lambda(1-q) - q(k-\lambda) = 0$$
$$\Rightarrow \lambda - qk = 0 \Rightarrow q = \frac{1}{k}\lambda$$

קיבלנו משהו הגיוני: ההטיה היא המספר היחסי של ההטלות שיצאו 0.

דוגמה: נסתכל על התפלגות נורמלית $\mathcal{N}(\mu, E)$.

μ - התוחלת

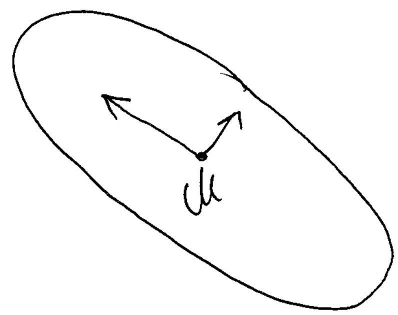
E - מטריצת שונות משותפת. איברי האלכסון הם השונות ומחוץ לאלכסון זו השונות המשותפת.

ההסתברות לקבל וקטור x בהינתן שהוא נלקח מהתפלגות נורמלית $\theta = (\mu, E)$ היא

$$p(x \mid \theta) = \frac{1}{(2\pi)^{n/2} |E|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^t E^{-1}(x-\mu)}$$

n להגודל של המטריצה.

בדו-מימד למעשה יש אליפסואיד עם מרכז ושונות בכל כיוון



נתונה סדרה $x_1, \dots, x_k \in \mathbb{R}^n$ iid והמטרה היא למצוא

$\theta^* = (\mu^*, E^*)$ כך ש- $\theta^* = \arg\max_\theta P(x_1, \dots, x_k \mid \theta)$

אם היה לנו את $p(\theta)$ אז היינו מחשבים את

$$\max p(x_1, \dots, x_k \mid \theta) p(\theta) \equiv \max p(\theta \mid x_1, \dots, x_k)$$

זה הדבר הנכון לעשות. אבל אם אין לנו את ה prior מניחים התפלגות אחידה ואז אומרים ש- MAP ⟶ ML.

אז מה שאנחנו רוצים למצוא זה

$$\theta^* = \arg\max_\theta p(x_1, \dots, x_k \mid \theta) = \arg\max_\theta \prod_{i=1}^{k} p(x_i \mid \theta) =$$
$$= \arg\max_\theta \underbrace{\sum_{i=1}^{k} \log p(x_i \mid \theta)}_{L(\theta)}$$

$$L(\theta) = \frac{1}{2} k n \log(2\pi) - \frac{1}{2} k \log|E| - \frac{1}{2} \sum_{i=1}^{k} (x_i - \mu)^t E^{-1} (x_i - \mu)$$

עכשיו נרצה לגזור ולהשוות $\frac{\partial L}{\partial \sigma} = 0$, $\frac{\partial L}{\partial \mu} = 0$.

אבל ראשית נסתכל על המקרה הפרטי ש- $E = \sigma^2 I$.

אז $E^{-1} = \frac{1}{\sigma^2} I$ ו $|E| = \sigma^{2n}$.

$\frac{1}{2}kn\log(2\pi)$ לא תלוי ב-μ וגם לא ב-σ ולכן

$$L(\theta) \equiv -kn\log\sigma - \frac{1}{2}\sum_{i=1}^{k}\frac{\|x_i-\mu\|^2}{\sigma^2}$$

$$\Rightarrow \frac{\partial L}{\partial \sigma} = -\frac{kn}{\sigma} + \sigma^{-3}\sum_{i=1}^{k}\|x_i-\mu\|^2 = 0$$

$$\Rightarrow -kn + \frac{1}{\sigma^2}\sum_{i=1}^{k}\|x_i-\mu\|^2 = 0$$

$$\Rightarrow \sigma^2 = \frac{1}{kn}\sum_{i=1}^{k}\|x_i-\mu\|^2 \quad (=\sigma^{*2})$$

למה יצא לנו ראו חלוקה ב-n? נניח ש-

$$E = \begin{pmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_n^2\end{pmatrix} \qquad \sigma_1=\sigma_2=\dots=\sigma_n=\sigma$$

אז

$$\frac{1}{k}\sum\|x_i-\mu\|^2 = \sum\sigma_i^2 = n\sigma^2$$

$$\Rightarrow \frac{1}{n}\cdot\frac{1}{k}\sum\|x_i-\mu\|^2 = \sigma^2$$

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{k}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{k}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{k}x_i - k\mu = 0$$

$$\Rightarrow \mu = \frac{1}{k}\sum x_i \quad (=\mu^*)$$

אז קיבלנו ראו את המצבים שאנחנו מכירים - תוחלת ושונות
כפי שהכרנו אותם מראש!
ברשימות באותו אופן המקרה הכללי ושם מתקבל $E = \frac{1}{k}\sum(x_i-\mu)(x_i-\mu)^t$

<u>דוגמה:</u> נניח ש- $Y=\{h_1,h_2\}$. נחפש את

$$h^* = \arg\max_{h_1,h_2} p(h|x)$$

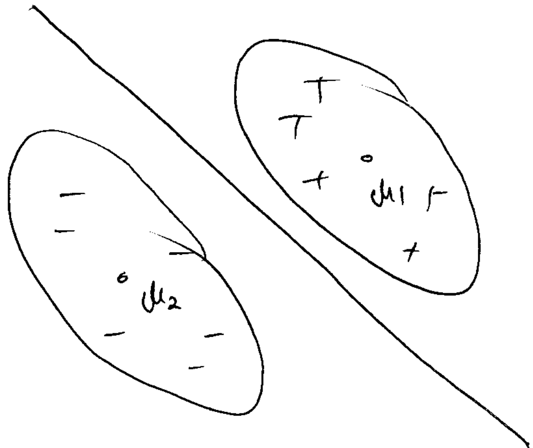
המשטח המפריד כפונקציה של x הוא

$$p(h_1|x) - p(h_2|x) = 0$$

אם נניח ש- $p(h_1)=p(h_2)$ אז המשטח הוא $p(x|h_1) - p(x|h_2) = 0$

נניח ש-
$$p(x|h_1) \sim N(\mu_1, E)$$
$$p(x|h_2) \sim N(\mu_2, E)$$

נמצא מה שאנחנו מקבלים לגבי את המשטח המפריד



$$p(x|h_1) - p(x|h_2) = 0 \Rightarrow p(x|h_1) = p(x|h_2)$$
$$\Rightarrow \frac{p(x|h_1)}{p(x|h_2)} = 1 \Rightarrow \log p(x|h_1) - \log p(x|h_2) = 0$$

$$\Rightarrow -\frac{1}{2}(x-\mu_1)^t E^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^t E^{-1}(x-\mu_2) = 0$$
$$\Rightarrow x^t E^{-1}(\mu_1-\mu_2) \underbrace{-\frac{1}{2}\mu_1^t E^{-1}\mu_1 + \frac{1}{2}\mu_2^t E^{-1}\mu_2}_{-\frac{1}{2}(\mu_1+\mu_2)E^{-1}(\mu_1-\mu_2)} = 0$$

$$\Rightarrow x^t \underbrace{E^{-1}(\mu_1-\mu_2)}_{\omega} - \underbrace{\frac{1}{2}(\mu_1+\mu_2)}_{\mu} \underbrace{E^{-1}(\mu_1-\mu_2)}_{\omega} = 0$$

$$\Rightarrow \omega^t(x-\mu) = 0$$

קיבלנו שהמשטח המפריד הוא על מישור! כלומר המפריד הוא ישר ממומש הפרמטרים שלו הם בדיוק הפרמטרים של LDA שבא בכלל בכיוון אחר בחיפוש.

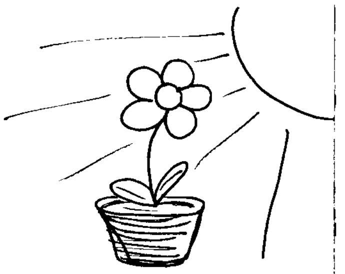
מה עם SVM? הוא אופטימלי מבחינה הרבה כי הוא לא מניח שום דבר על ההתפלגות והוא לא משערך אותה. אבל אנחנו משלמים על זה מחיר שצריך מלא דאטות.

מחר אין שיעור. שבוע הבא אין שיעורים.
בשבוע לאחר מכן לומדים גם בשני וגם בשישי.

7

12.02.08

מע' לומדות



נראות מירבית, אנטרופיה יחסית, Gibbs Max Entropy

נראה שהתפלגות אמפירית של מדגם בת"ל היא הקרובה ביותר, במובן ML, להתפלגות האמיתית.

(הטענה הזו לחלוטין בלתי תלויה במה שעשינו קודם)

נניח ש- $X=\{a_1,\dots,a_n\}$; $p(a|\theta)$ ההסתברות לדגום $a\in X$. יהי מדגם $x_1,\dots,x_m\in X$ iid.

- ספירת מופע של a הוא מספר הפעמים ש-a הופיע במדגם

$$f(a)=|\{i : x_i=a\}|$$

- ההתפלגות האמפירית של a היא

$$\hat{p}(a)=\frac{1}{\sum_{b\in X} f(b)}f(a)=\frac{1}{m}f(a)$$

מה הקשר בין $f(a)$ ו- $p(a)$?

$$p(x_1\dots x_m|\theta)=\prod p(x_i|\theta)=\prod_{a\in X}p(a|\theta)^{f(a)}$$

$$\Rightarrow \max_{p\in Q} p(x_1\dots x_m|\theta)=\max_{p\in Q}\prod_{a\in X}p(a|\theta)^{f(a)}$$

כאשר $Q=\{q\in\mathbb{R}^n : q\ge 0 \quad \sum q_i=1\}$ מרחב וקטורי הסתברות.

נסמן $f_i=f(a_i)$ ו- $p_i=p(a_i|\theta)$.

אז אנחנו צריכים לפתור בעית אופטימיזציה -

$$\max_{p_1\dots p_n}\prod_{i=1}^n p_i^{f_i} \quad s.t.\ p_i\ge 0,\ \sum p_i=1$$

$$\max_{\substack{p_1\dots p_n\\ s.t.\ p_i\ge 0\\ \sum p_i=1}}\prod p_i^{f_i} \equiv \max_{\substack{p_1\dots p_n\\ s.t.\ p_i\ge 0\\ \sum p_i=1}}\sum f_i\log p_i$$

נשתמש בלגרנז'יאן אבל נתעלם מהאילוץ $p_i\ge0$ ונראה שנקבל מה שצריך

$$\mathcal{L}(p_i,\lambda,\mu)=\sum f_i\log p_i-\lambda\left(\sum p_i-1\right)$$

$$\frac{\partial\mathcal{L}}{\partial p_i}=\frac{f_i}{p_i}-\lambda=0 \Rightarrow p_i=\frac{1}{\lambda}f_i$$

$$\Rightarrow 1=\sum p_i=\frac{1}{\lambda}\sum f_i \Rightarrow p_i=\frac{1}{\sum f_j}f_i=\hat{p}_i$$

מכאן ש- $f_i \ge 0$ קיבלנו בחינם ש- $p_i \ge 0$.
עיקרון ה- ML אומר שההסתברות האמפירית היא למעשה מידע
נאמן של הנראות האמיתית.

<u>אנטרופיה יחסית</u> - x,y וקטורים. האנטרופיה היחסית
בינהם היא

$$D(x\|y) = \sum_i x_i \ln\frac{x_i}{y_i} - \sum x_i + \sum y_i$$

כאשר x,y וקטורי התפלגות אז $\sum y_i = \sum x_i = 1$ ואז
$D(x\|y) = \sum x_i \ln\frac{x_i}{y_i}$ וזה נקרא KL-Divergence.
D זו לא מטריקה! קודם כל, D לא סימטרית. גם א-
שוויון המשולש לא מתקיים. הדבר היחיד שמתקיים הוא חיוביות.

<u>טענה:</u> $D(x\|y) \ge 0$.

<u>הוכחה:</u> נשתמש בשני אי-שוויונים:

log-sum inequality: $\sum x_i \ln\frac{x_i}{y_i} \ge \left(\sum x_i\right)\ln\frac{\sum x_i}{\sum y_i} = \bar{x}\cdot\ln\frac{\bar{x}}{\bar{y}}$

$$\ln(x+1) > \frac{x}{1+x} \qquad x>-1,\ x\neq 0$$

מכאן נקבל ש-

$$D(x\|y) \ge \bar{x}\ln\frac{\bar{x}}{\bar{y}} - \bar{x} + \bar{y}$$

מתי $\bar{x}\ln\frac{\bar{x}}{\bar{y}} \ge \bar{x} - \bar{y}$?

$$a\ln\frac{a}{b} \ge a-b \iff \ln\underbrace{\frac{a}{b}}_{x} \ge 1 - \underbrace{\frac{b}{a}}_{1/x} \iff \ln x \ge 1 - \frac{1}{x}$$

$$\ln(x+1) > \frac{x}{x+1},\quad y = x+1 \quad \ln y > \frac{y-1}{y} = -\frac{1}{y}$$
$$x>-1 \qquad\qquad y>0$$

$\Leftarrow$ מאחר ש- $\bar{x}, \bar{y}$ אי שליליים זה מתקיים. ☺

$$\min_p \underbrace{D(f\|p)}_{\sum f_i \ln f_i - \sum f_i \ln p_i - \sum f_i + 1} \quad \text{s.t } p\ge 0,\ \sum p_i = 1$$

$$\equiv \max \sum f_i \ln p_i \quad \text{s.t } p_i \ge 0\ \sum p_i = 1 \implies p_i^* = \frac{1}{\|f\|_1} f_i = \hat{p}_i$$

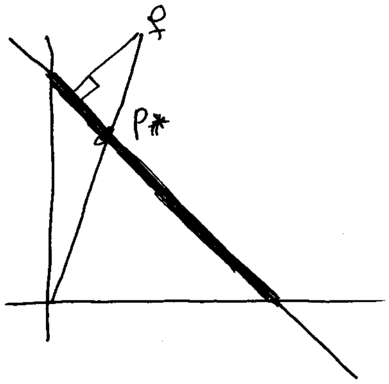
נקודה אחת p אי שלילי. נקודה שני p יושב על על-מישור כך

$\sum p_i = 1$ כי משוואה לינארית

נשים לב ש- p_i^* כי פשוט איזה scaling

של f_i. אם לא נקבל שזה יצא אי-שלילי

שהרי f_i אי-שליליים.



$\sum p_i = 1$

טיפלנו ב- $\min_p D(f \| p)$ שומר משמעות האופטימיזציה

נקודה ימין. נשאל מה אם p נקודה שמאל? מכיוון ש- D

לא סימטרית זה לא יוצא אותו דבר.

נזכור שהאנטרופיה היא $H(x) = -\sum x_i \ln x_i$

$$D(p \| \tfrac{1}{n}\cdot\vec{1}) = \sum p_i \ln p_i - \sum p_i \ln \tfrac{1}{n} =$$

$$= \sum p_i \ln p_i - \cancel{\sum p_i \ln 1} + \sum p_i \ln n =$$

$$= \sum p_i \ln p_i + \ln n \left(\cancel{\sum p_i}\right) = \ln n - H(p)$$

$$\Rightarrow \arg\min_p D(p \| \tfrac{1}{n}\cdot\vec{1}) = \arg\max H(p)$$

עקרון לפלאס אומר שאם אין שום מידע אז צריך לשאוף להתפלגות

אחידה. זה אומר שצריך לחפש אנטרופיה מקסימלית.

אילוץ לינארי

נניח שיש לנו קוביה ושואלים מה הסיכוי שהקוביה תצא i מנחת

$$\max_p H(p) \text{ s.t } p \geq 0 \quad \sum p_i = 1 \quad \sum \alpha_i p_i = \beta$$

$$\mathcal{L}(p, \lambda, \mu) = -\sum p_i \ln p_i + \lambda\left(\sum p_i - 1\right) + \mu\left(\sum \alpha_i p_i - \beta\right)$$

$$\frac{\partial \mathcal{L}}{\partial p_i} = -(\ln p_i + 1) + \lambda + \mu \alpha_i = 0$$

$$\Rightarrow \ln p_i = (1+\lambda) + \mu \alpha_i$$

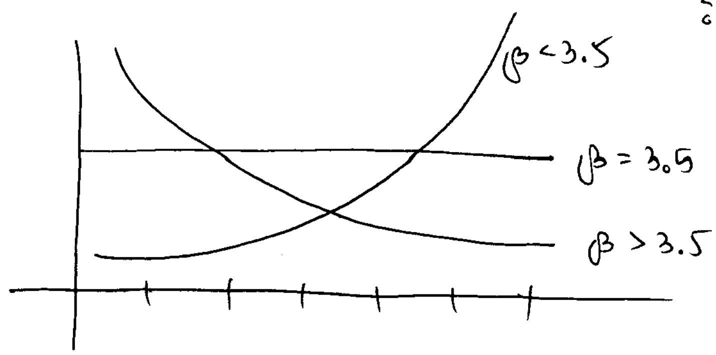
$$\Rightarrow p_i = e^{1+\lambda} \cdot e^{\mu \alpha_i}$$

$$\Rightarrow 1 = \sum p_i = e^{1+\lambda} \sum e^{\mu \alpha_j}$$

$$\Rightarrow \boxed{p_i = \frac{1}{Z} e^{\mu \alpha_i}}$$ Boltzman-Gibbs Distribution

$$Z = \frac{1}{e^{1+\lambda}} = \sum e^{\mu \alpha_j}$$

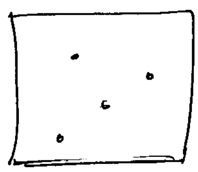
איך זה נראה?



באופן כללי -

$$\min_p D(p\|p_0) \quad s.t \quad p_i\ge 0 \quad \sum p_i=1 \quad \sum \alpha_i p_i=\beta$$
$$\Rightarrow p_i^* = \frac{1}{Z}p_{0i}e^{\mu\alpha_i}$$

$$\min_p D(p\|p_0) \quad s.t \quad p_i\ge 0 \quad \sum p_i=1 \quad \sum_i f_{ij}p_i=b_j \quad j=1,\dots,k$$
$$\Rightarrow \underbrace{p_i^* = \frac{1}{Z}p_{0i}e^{\sum_{j=1}^{k}\mu_j f_{ij}}}_{\text{זה המקרה הכללי של התפלגות בולצמן - גיבס.}}$$

נניח שיש לנו מרחב  ובו נתונות כמה נקודות בתוספת ההסתברות שלהן. אנחנו רוצים לעשות אינטרפולציה. אפשר לעשות את זה עם שיקולים מקומיים כאשר הנקודה לסביבה. אבל מה אם הנקודות בודדות מאוד? ויש לנו אילוצים שמתקיימים ($\sum f_{ij}p_i=b_j$) ואנחנו נחפש את האינטרפולציה שהכי קרובה להתפלגות אחידה. זה בדיוק מה שעשינו למעלה!

נסמן את $P=\{p\in R^n : p\ge 0,\ \sum p_i=1,\ \mathbb{E}[f_j]=\hat{\mathbb{E}}[f_j]\ \ 1\le j\le k\}$

$$\mathbb{E}[f_j]=\sum f_{ij}p_j \qquad \hat{\mathbb{E}}[f_j]=b_j$$

$$Q=\{q\in R^n : q \text{ Gibbs-Distribution}\}$$

$$\min_{q\in Q} D(\hat{p}\|q) \quad - ML$$

אחד האלגוריתמים הנפוצים לפתרון הבעיה הנ"ל הוא Iterated Scaling.

שיעור הבא ביום שלישי.

9 19.2.08
מג'ויומדור

Expectation Maximization (EM)

נתזכר קודם של ML :

נניח שיש לנו עולם $X=\{a_1,\dots,a_n\}$ קבוצת תווים.

ויש לנו מדגם $x_1,\dots,x_m$ כאשר $x_i\in X^d$

כ דוגמה זה וקטור עם d כניסות וכל כניסה אחד מ-n התווים.

נגדיר את תדירות המופע:

$$\bar{P}_{i_1,\dots,i_d} = \left|\{x_i=(a_{i_1},\dots,a_{i_d}) : i=1,\dots,m\}\right|$$

$1\le i_j\le n$

מספר הפעמים ש-$(a_{i_1},\dots,a_{i_d})$ הופיע במדגם.

ב-ML רוצים לפתור את הבעיה

$$p^* = \underset{p\in Q}{\arg\min}\, D(\bar{P}\|p)$$

כאשר $Q=\{q : q\ge 0, \|q\|=1\}$ - קבוצת וקטורי ההסתברות האפשריים.

ראינו שהפתרון היה $p^*=\frac{1}{m}\bar{P}=\hat{p}$.

נסתכל על $Q=\{\sum_{j=1}^{k}\lambda_j p^j \ ; \ \lambda\ge 0, \sum\lambda_j=1 \ \ p^j\in H(\theta_j)\}$

משפחה של התפלגויות עם פרמטרים θ_j

אז Q זו משפחה של קומבינציות לינאריות של התפלגויות שונות.

לדוגמה יכול להיות ש-$H=\{\mathcal{N}(\mu,\Sigma)\}$ $\theta=(\mu,\Sigma)$ ואז אנחנו מסתכלים על צירופים של התפלגויות נורמליות.

עוד דוגמה הכונה $\theta=u$ $H(\theta)=\{u^{\otimes d} \mid u\ge 0 \ u\in\mathbb{R}^n, \|u\|_1=1\}$

זה נקרא התפלגות ברנולי.

מכפלה חיצונית - מכפלת u בעצמו d פעמים:

$$(u^{\otimes d})_{i_1\dots i_d} = u_{i_1}\cdot u_{i_2}\cdots u_{i_d}$$

למשל במקרה ש-$d=2$

אז $u^{\otimes 2}=u\otimes u=uu^t$

ואז $u^{\otimes d}$ זה פשוט המכפלה הזאת d פעמים.

מוטיבציה: Latent Class model

נניח ש- $x_i=(x_1,\dots,x_d)$. אז

$$p(x_1,\dots,x_d)=\sum_y p(x_1\dots x_d|y)p(y) =$$

$$=\sum_{j=1}^{k} p(x_1\dots x_d|y=\alpha_j)p(y=\alpha_j)$$

$y\in\{\alpha_1\dots\alpha_k\}$

את Y אין לנו והוא נקרא משתנה חבוי. אפשר לומר ש- $p(y=\alpha_j)=\lambda_j$; $P(x_1\dots x_d|y=\alpha_j)\in H(\theta_j)$ אז

$$p(x_1\dots x_d)=\sum_{j=1}^{k} p(x_1\dots x_d|y=\alpha_j)\,p(y=\alpha_j) =$$

$$=\sum_{j=1}^{k}\lambda_j p^j \qquad \left(p^j=p(x_1\dots x_d)\right)$$

אז מכאן מגיע העניין שלנו ב- Q - מהסתכלות ברדפים של דגימות שבאים מכל מיני התפלגויות.

דוגמה: $H=\mathcal{U}^{\otimes 3}$. יש מטבעות.

p - ההטיה של מטבע 1 (הסיכוי שייצא אפס)

q - ההטיה של מטבע 2

λ - סיכוי לבחור מטבע 1.

אנחנו בוחרים מטבע ואז מטילים אותו 3 פעמים.

אז $X\in\{0,1\}^3$ מייצג הטלת מטבע שנבחר 3 פעמים.

את זה אנחנו עושים m פעמים ומקבלים מדגם

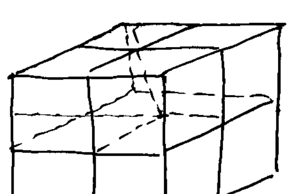
$\{x_1\dots x_m\}$

עכשיו נניח שכל מה שיש לנו זה המדגם הזה והמטרה שלנו היא למצוא את λ, q, p.

ההתפלגות האמפירית היא

$$\hat{p}_{i_1,i_2,i_3}=\frac{1}{m}\left|\{x_i=(i_1,i_2,i_3)\ ;\ i=1,\dots,m\}\right|$$

$\hat{p}$ הוא מערך בגודל $2\times2\times2$ -



יש קוביה ... פקודה

שם קוביה קטנה פה מייצגת דירוג (i_1,i_2,i_3) ובתוכה רשום כמה פעמים הדירוג הזה הופיע בדגימה שלנו.

נשאל מה ההסתברות $P(X=(i_1,i_2,i_3)) = ?$

נגדיר $Y=\{1,2\}$ משתנה שמציין את המטבע הנבחר.

$$\Rightarrow P(X=(i_1,i_2,i_3)) = \underbrace{P(X|y=1)}_{p^{\#0}(1-p)^{3-\#0}}\underbrace{P(y=1)}_{\lambda} + \underbrace{P(X|y=2)}_{q^{\#0}(1-q)^{3-\#0}}\underbrace{P(y=2)}_{1-\lambda} =$$

$$= \lambda p^{\#0}(1-p)^{3-\#0} + (1-\lambda)q^{\#0}(1-q)^{3-\#0}$$

מהו Q? למעשה Q היא הקבוצה הבאה:

$$Q = \left\{ \lambda \begin{pmatrix} p \\ 1-p \end{pmatrix}^{\otimes 3} + (1-\lambda)\begin{pmatrix} q \\ 1-q \end{pmatrix}^{\otimes 3} : 0 \le p,q,\lambda \le 1 \right\}$$

מה ה-ML?

$$P^* = \underset{0\le q,p,\lambda \le 1}{\arg\min}\, D(\hat{P} \| Q)$$

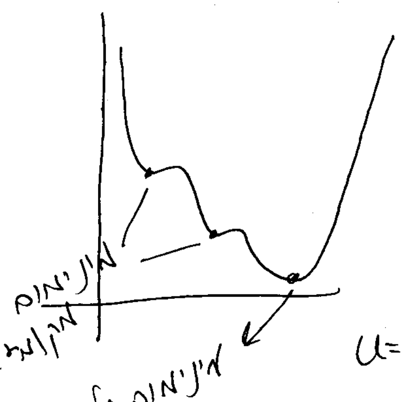
$$\underset{y}{\arg\min}\, D(x\|y) = \arg\max \sum_i x_i \log y_i$$

$$\Rightarrow \underset{0\le p,q,\lambda\le 1}{\arg\min}\, D(\hat{P}\|Q) = \underset{0\le p,q,\lambda\le 1}{\arg\max} \sum_{0\le i_1,i_2,i_3\le 1} \hat{P}_{i_1i_2i_3} \lg\Big(\lambda p^{\overbrace{3-(i_1+i_2+i_3)}^{\#1}}(1-p)^{i_1+i_2+i_3} + (1-\lambda)q^{3-(i_1+i_2+i_3)}(1-q)^{i_1+i_2+i_3}\Big)$$

את הבעיה הזאת לא נפתור. היא קשה מפני כי יש פה לוג של סכום!

מטרת האלגוריתם EM היא לעבור מלוג של סכום לסכום של לוגים שזה קל לטפל.

גם הבעיות מהצורה הזאת הן NP-קשות. אנחנו לא נדע למצוא את הפתרון האופטימלי. מה שניתן למצוא זה פתרון לוקאלי - יהיה לנו אלגוריתם שמתכנס לנקודת קיצון מקומית.



ניתן תחושה למה הבעיה הזו קשה.

$$H = \{ u^{\otimes d} : u \in \mathbb{R}^n,\ u \ge 0,\ \|u\|_1 = 1 \}$$

$$Q = \{ \sum \lambda_j u_j^{\otimes d} : \lambda \ge 0,\ \|\lambda\|_1 = 1,\ \|u_j\| = 1,\ u_j \ge 0 \}$$

אם $d=2$ אז $\underset{n\times n}{Q} = UDU^t$ כאשר $U = [u_1, \dots, u_k]_{n\times k}$

$$D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{bmatrix}_{k\times k}$$

כזכור $UU^T = \sum_{i=1}^{k} u_i u_i^t \Leftarrow UDU^t = \sum_{i=1}^{k} \lambda_i u_i u_i^t$

הדרגה של Q היא לכל היותר k. כלומר במקרה ש- $d=2$ מחפשים מטריצה שדרגתו לכל היותר 2 והיא הכי קרובה ל- $\hat{P}$. אם האינדקסים היו ייחודיים היה אפשר לפתור את זה עם PCA. עבור $d>2$ הדברים האלה נקראים טנזורים. וכדי לפתור את הבעיה צריך לדעת למצוא את דרגת הטנזור (דרגה של טנזור זה מספר הטנזורים מדרגה 1 שסכומם אותו טנזור. טנזור מדרגה 1 זה $u^{\otimes d}$), והוכיחו שזו בעיה קשה.

עד כאן ההקדמה.

נקודת מבט סטטיסטית - פיתוח קלאסי שלי של EM

ML

$$\theta^* = \operatorname{argmax} \log P(x_1 \dots x_m \mid \theta)$$

יש דגימות $x_1,\dots,x_m$ שהגיעו ממשפחת התפלגויות עם פרמטר θ ומחפשים את ה התפלגות θ שנותנת את הנראות המקסימלית.

לקבוצה $\{x_1,\dots,x_m\}$ נקרא D. אז

$$\theta^* = \operatorname{argmax} \underbrace{\log \underbrace{\sum_y P(D,y \mid \theta)}_{P(D\mid\theta)}}_{L(\theta)}$$

$$\max L(\theta) = ?$$

$q(y \mid D,\theta)$ - התפלגות שרירותית. כלומר $\sum_y q(y \mid D,\theta) = 1$

נגדיר פונקציה $Q(q,\theta)$ שתקיים

$$L(\theta) \ge Q(q,\theta) \quad \forall q$$

$$L(\theta) = \log\Big(\sum_y p(D,y \mid \theta)\Big) = \log\Big[\sum_y q(y \mid D,\theta)\frac{P(D,y \mid \theta)}{q(y \mid D,\theta)}\Big] \ge$$

יש פה הכפלה סמויה
שהיא מתפרקת לזהות ממש מאופס

[אי שוויון Jensen אומר ש- $\log(\sum w_j x_j) \ge \sum w_j \log x_j$ כאשר $\sum w_j = 1$]

$$\ge \sum_y q(y \mid D,\theta) \log\Big(\frac{P(D,y \mid \theta)}{q(y \mid D,\theta)}\Big) =: Q(q,\theta)$$

אנחנו נטען שלעשות אופטימיזציה על Q לה קל יותר מאשר על L ושיש התאמה בין המקסימום של Q למקסימום של L.

אז יש לנו $L(\theta) \ge Q(q,\theta)$ לכל q. אנחנו הולכים
לעשות איזה תהליך איטרטיבי במעבר של ה-θ-ות.
בהינתן $\theta = \theta^{(t)}$ מהו q^* כך ש- $L(\theta^{(t)}) = Q(q^*, \theta^{(t)})$
(ז"א, $q^* = \max_q Q(q, \theta^{(t)})$)

<u>טענה:</u> $q^* = p(y|D, \theta^{(t)})$.

<u>הוכחה:</u>

$$Q(P(y|D,\theta^{(t)}), \theta^{(t)}) =$$

$$\overset{\text{הגדרה}}{=} \sum_y p(y|D,\theta^{(t)}) \log \frac{P(D,y|\theta^{(t)})}{p(y|D,\theta^{(t)})} =$$

$$= \sum_y p(y|D,\theta^{(t)}) \log \frac{\cancel{P(y|D,\theta^{(t)})}\, p(D|\theta^{(t)})}{\cancel{P(y|D,\theta^{(t)})}} =$$

$$= \log P(D|\theta^{(t)}) \cdot \underbrace{\sum_y p(y|D,\theta^{(t)})}_{=1} =$$

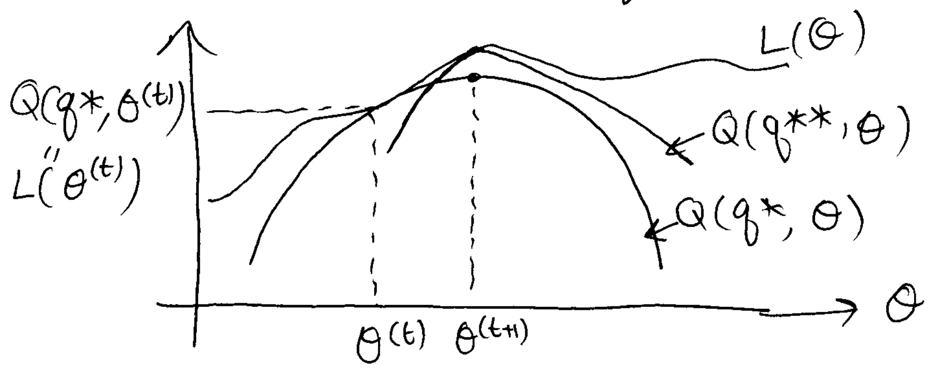
$\log P(D,\theta^{(t)})$ לא תלוי ב-y

$$= \log P(D|\theta^{(t)}) = L(\theta)$$

ועכשיו גם ברור שזה מקסימום כי מאחר ש- $L(\theta) \ge Q(q,\theta)$
והגענו לשוויון - השוויון הזה בהכרח במקסימום.

שוב, יש לנו $\forall q$ $L(\theta) \ge Q(q,\theta)$ ובהינתן $\theta = \theta^{(t)}$

$$L(\theta^{(t)}) = Q(\underbrace{p(y|D,\theta^{(t)})}_{q^*}, \theta^{(t)})$$



עכשיו שאנחנו נמצאים ב- האופטימלי הזה אז עושים אופטימיזציה
של Q לפי θ אנחנו עולים וזה בהכרח גורם לעליה גם ב- $L(\theta)$.
נניח שהגענו ל- $\theta^{(t+1)}$. עכשיו נמצא את q^{**} לפי הנוסחה הקודמת
וככה נמשיך איטרטיבית.

אז האלגוריתם הוא כזה:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(P(y|D,\theta^{(t)}),\theta) =$$

$$= \arg\max_{\theta} \sum_{y} P(y|D,\theta^{(t)}) \log P(D,y|\theta) =$$

$$= \mathop{E}_{Y\sim P(y|D,\theta^{(t)})}[\log P(D,y|\theta)]$$

אז האלגוריתם מוצא תוחלת ואז ממקסם אותה.

בשיעור הבא נראה מה קורה כאשר $x_1 \dots x_m$ הם iid ואז הבעיה הזה יהיה פשוט יותר ונוכל לפתור את בעית האופטימיזציה.

שבוע הבא נפגשים ביום שלישי.

(12) 26.02.08

מעין ומפור.

EM

התחלנו עם $D=\{x_1,\dots,x_m\}\subseteq \mathbb{R}^d$

$$\theta^* = \arg\max_\theta \log P(D|\theta) = \arg\max \log \sum P(D,y|\theta)$$

משתנה חבוי ←

וסיימנו עם תהליך איטרטיבי

$$\theta^{(t+1)} = \arg\max_\theta \sum_y P(y|D,\theta^{(t)}) \log P(D,y|\theta)$$

זה עוד לא אלגוריתם. זה רק מסגרת כללית.

כדי להפוך את זה לאלגוריתם של ממש אנחנו נניח ש- $x_1,\dots,x_m$ הם iid.

תחת ההנחה הזאת החיים שלנו יהפכו להרבה יותר קלים.

$$P(D|\theta) = \prod_{i=1}^m P(x_i|\theta)$$

$$P(D,y|\theta) = \prod_{i=1}^m P(x_i,y_i|\theta) \longrightarrow$$

לכל דוגמה יש את ה-y שלו.
ומשם לדוגמה זו המטבע
לכל דוגמה y אומר מהו המטבע הנבחר.

$$P(y|D,\theta) = \prod_{i=1}^m P(y_i|x_i,\theta)$$

$$\Rightarrow \theta^{(t+1)} = \arg\max_\theta \sum_{y_1\dots y_m}\Big[\underbrace{\prod_i P(y_i|x_i,\theta^{(t)})}_{P(y|D,\theta^{(t)})} \sum_i \underbrace{\log P(x_i,y_i|\theta)}_{\alpha(y_i)}\Big] =$$

$$= \arg\max \sum_{i=1}^m \sum_{y_1\dots y_m} \alpha(y_i) P(y_1|x_1,\theta^{(t)})\cdot\ldots\cdot P(y_m|x_m,\theta^{(t)}) =$$

$$= \arg\max \sum_{i=1}^m \Big[\underbrace{\sum_{y_1} P(y_1|x_1,\theta^{(t)})}_{=1} \underbrace{\sum_{y_2} P(y_2|x_2,\theta^{(t)})}_{=1} \ldots \sum_{y_i}\alpha(y_i)P(y_i|x_i,\theta^{(t)}) \ldots \underbrace{\sum_{y_m} P(y_m|x_m,\theta^{(t)})}_{=1}\Big]$$

$$\left(\text{כי } \sum_x P(x|y)=1\right)$$

$$= \arg\max \sum_{i=1}^m \Big[\sum_{y_i} \alpha(y_i) P(y_i|x_i,\theta^{(t)})\Big] =$$

$$= \arg\max \sum_{i=1}^m \sum_{j=1}^k P(y_i=j|x_i,\theta^{(t)}) \log P(x_i,y_i=j|\theta)$$

נניח $y_i\in\{1,\dots,k\}$

הצלחנו לפשט את הסכום הענק לסכום קטן יותר.

נשתמש בזה כדי לפתור את בעיית המטבעות. במקרה זה $k=2$ (נסמן $w_i^{(t)} = P(y_i=1|x_i,\theta^{(t)})$: $\Leftarrow$ $P(y_i=2|x_i,\theta^{(t)}) = 1-w_i^t$. מחפשים

$$\max_{\lambda,p,q} \sum_{i=1}^m \Big(w_i^{(t)} \log\big[\underbrace{P(x_i|y_i=1,\theta)}_{p^{n_i}(1-p)^{3-n_i}} \underbrace{P(y_i=1|\theta)}_{=\lambda}\big] + (1-w_i^{(t)})\log\big[\underbrace{P(x_i|y_i=2,\theta)}_{q^{n_i}(1-q)^{3-n_i}} \underbrace{P(y_i=2|\theta)}_{=1-\lambda}\big]\Big)$$

כאשר n_i מספר האפסים. (הסתברות λ)

(הסיכוי לבחור את המטבע השני)

$$= \max_{\lambda,p,q} \overbrace{\sum_{i=1}^{m}\left(\omega_i^{(t)} \log\left[\lambda p^{n_i}(1-p)^{3-n_i}\right] + (1-\omega_i^{(t)}) \log\left[(1-\lambda) q^{n_i}(1-q)^{3-n_i}\right]\right)}^{(*)}$$

לפי נוסחת בייס:

$$p(y_i=j \mid x_i, \theta^{(t)}) = \frac{p(x_i \mid y_i=j, \theta^{(t)})\, p(y_i=j \mid \theta^{(t)})}{p(x_i \mid \theta^{(t)})}$$

$$p(x_i \mid \theta^{(t)}) = \sum_j p(x_i, y_i=j \mid \theta^{(t)}) = \sum_j p(x_i \mid y_i=j, \theta) \cdot p(y_i=j \mid \theta)$$

המכנה הוא רק גורם נרמול שאומר לנו שבסוף הסכום של כולם יהיה 1, אז למעשה מספיק לנו למצוא רק את המונה. אבל אותו אנחנו יודעים לחשב.
ולמשל אם $j=1$ במקרה של המטבע

$$(**)\quad w_i^{(t)} = \frac{\lambda^{(t)} {p^{(t)}}^{n_i}(1-p^{(t)})^{3-n_i}}{\lambda^{(t)} {p^{(t)}}^{n_i}(1-p^{(t)})^{3-n_i} + (1-\lambda^{(t)}) {q^{(t)}}^{n_i}(1-q^{(t)})^{3-n_i}}$$

נגזור את $(*)$:

$$\frac{\partial(*)}{\partial p} = \frac{\sum_i w_i n_i}{p} - \frac{\sum_i w_i(3-n_i)}{1-p} = 0$$

$$\Rightarrow (1-p)\sum w_i n_i - p\left[\sum w_i(3-n_i)\right] = 0$$

$$\Rightarrow \sum w_i n_i - p\Big[\underbrace{\sum w_i n_i + \sum w_i(3-n_i)}_{3\sum w_i}\Big] = 0$$

$$\Rightarrow p^{(t+1)} = \frac{1}{\sum w_i^{(t)}} \sum_i w_i^{(t)} \cdot \frac{n_i}{3}$$

$$q^{(t+1)} = \frac{1}{\sum (1-w_i^{(t)})} \sum_i (1-w_i^{(t)}) \frac{n_i}{3}$$ באופן דומה

$$\frac{\partial(*)}{\partial \lambda} = \frac{\sum_i w_i}{\lambda} - \frac{\sum_i (1-w_i)}{1-\lambda} = 0$$

$$\Rightarrow (1-\lambda)\sum w_i - \lambda \sum(1-w_i) = 0$$

$$\Rightarrow \lambda^{(t+1)} = \frac{1}{m}\sum_i w_i^{(t)}$$

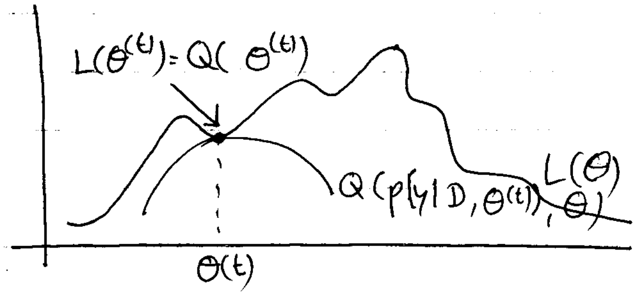
עכשיו מהנוסחה $w_i^{(t)}$ יש לנו נוסחאות עדכון ל- λ, p, q. איך מעדכנים את w_i? בנוסחה $(**)$ פשוט מציבים את ה λ, p, q המעודכנים.

מתחילים את התהליך מאיזשהו ניחוש התחלתי. $q^{(0)}, p^{(0)}, \lambda^{(0)}$ למשל $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$.
מפה נגזר $w_i^{(0)}$ וכך מתחילים לעדכן לפי הנוסחאות.

התחלנו מ- $\log P(D|\theta) = L(\theta)$ ומצאנו $Q(q,\theta)$ רק ש-

$L(\theta) \geq Q(q,\theta)$ לכל q. !

$$p(y|D,\theta) = q^* = \arg\max_q Q(q,\theta)$$
$$L(\theta) = Q(p(y|D,\theta), \theta)$$



יש משפט שאומר שהנקודות הסטציונריות של $L(\theta)$ הן גם הנקודות הסטציונריות של $Q(q,\theta)$ והמדויק אומר שמציאת המקסימום של Q נותנת את המקסימום של L.

Gaussian Mixture

אנחנו מתעניינים במקרה הפרטי של גאוסיאנים -

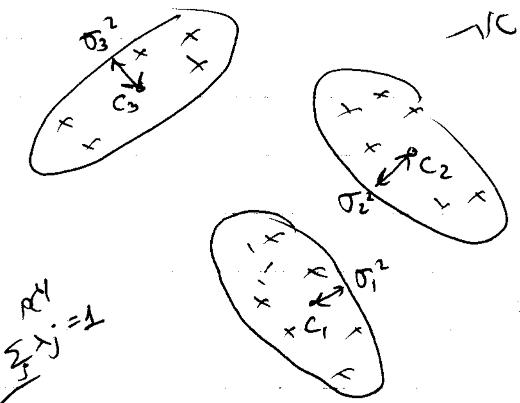
$x \in \mathbb{R}^d$

$E_j = \sigma_j^2 I$

$$p(x) = \sum_{j=1}^{k} p(x|y=j)\overbrace{p(y=j)}^{\lambda}$$

$$N(c_j, \sigma_j^2) = \frac{1}{(2\pi)^{d/2}\sigma_j^d} e^{-\frac{\|x-c_j\|^2}{2\sigma_j^2}}$$

המטרה היא בהינתן מדגם $\{x_1,\dots,x_m\} \subseteq \mathbb{R}^d$ למצוא את הפרמטרים $\theta = (\theta_1,\dots,\theta_k)$ $\theta_j = (c_j, \sigma_j^2)$.



זה מעניין לגשת ב- clustering - אם נמצא את הפרמטרים של הגאוסיאנים נוכל לקבוע לכל נקודה לאיזה גאוסיאן היא שייכת (יותר מדויק - לאיזה גאוסיאן היא שייכת בהסתברות הכי גבוהה)

נשתמש ב-EM כדי לפתור את המקרה שהגאוסיאנים הם עיגולים ולא אליפסות (

נגדיר $w_{ij}^{(t)} = p(y_i=j \mid x, \theta^{(t)})$ ונזכור שרוצים $\sum_j w_{ij} = 1$

$\lambda_j = p(y=j|\theta)$

$\sum_j \lambda_j = 1$ גם

אז
$$\theta^{(t+1)} = \arg\max_{\substack{\lambda_1 \dots \lambda_k \\ c_1 \dots c_k \\ \sigma_1^2 \dots \sigma_k^2}} \sum_{i=1}^{m}\sum_{j=1}^{k} w_{ij}^{(t)} \log\left[\lambda_j P(x_i|y_i=j,\theta)\right]$$

s.t. $\sum \lambda_j = 1$

ה ביטוי מה עמוד הקודם

הלגרנז'יאן הוא

$$\mathcal{L}(\lambda, c, \sigma, \mu) = \sum_i \sum_j [\ ---\ ] - \mu\left(\sum \lambda_j - 1\right)$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial \lambda_j} = \frac{\sum_i w_{ij}}{\lambda_j} - \mu = 0 \Rightarrow \lambda_j \mu = \sum_{i=1}^{m} w_{ij}$$

$$\Rightarrow \sum_j \lambda_j \mu = \sum_j \sum_i w_{ij}$$

$$\Rightarrow \mu \underbrace{\sum_j \lambda_j}_{=1} = \sum_i \underbrace{\sum_j w_{ij}}_{=1}$$

$$\Rightarrow \mu = m$$

$$\Rightarrow \boxed{\lambda_j^{(t+1)} = \frac{1}{m} \sum_{i=1}^{m} w_{ij}^{(t)}}$$

אין סיכוי שבמבחן
לא נקבל משהו כזה.
אנחנו צריכים לשיר
דברים כאלה מאמצע הלוח!

לפני שנמשיך לגזור לפי שאר המשתנים, נפשט את הביטוי.

$$\log P(X \mid Y=j;\theta) \equiv -d \log \sigma_j - \frac{\|x - c_j\|^2}{2\sigma_j^2}$$

משאירים רק את הפרמטרים
שרלוונטיים לגזירה

$$\frac{\partial \mathcal{L}}{\partial c_j} = \sum_i \frac{w_{ij}}{\sigma_j^2}(x_i - c_j) = 0$$

$$\Rightarrow \sum_i w_{ij} - \sum_i w_{ij} c_j = 0$$

$$\Rightarrow \boxed{c_j^{(t+1)} = \frac{1}{\sum_i w_{ij}^{(t)}} \sum_i w_{ij}^{(t)} x_i}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_j} = -\sum_i w_{ij} d \frac{1}{\sigma_j} + \frac{1}{\sigma_j^3} \sum_i w_{ij} \|x_i - c_j^2\| = 0$$

$$\Rightarrow \frac{1}{\sigma_j^2} \sum_i w_{ij} \|x_i - c_j\|^2 - d \sum_i w_{ij} = 0$$

$$\Rightarrow \boxed{\sigma_j^{2(t+1)} = \frac{1}{d \sum_i w_{ij}^{(t)}} \sum_i w_{ij} \|x_i - c_j\|^2}$$

ואיפה c_j זה ה c_j שמצאנו קודם למטה

$$w_{ij}^{(t)} = P(y_i = j \mid x_i, \theta^{(t)}) = \frac{\overbrace{P(x_i \mid y_i = j)}^{N(c_j^{(t)}, \sigma_j^{(t)})} \overbrace{P(y_i = j)}^{\lambda_j^{(t)}}}{P(x_i)}$$

זה נרמול. אפשר לחשב קודם את המונים
ואז לחלק בסכום.

נחזור לאנטרופיה היחסית. רצינו לפתור את הבעיה

$$\min \underbrace{D\left(\hat{p} \,\|\, \sum_j \lambda_j p^j\right)}_{L(\theta)} \quad \text{s.t} \quad \lambda \ge 0,\ \sum \lambda_j = 1,\ p^j \in H$$

לכל שתי מטריצות A,B נגדיר את $A \circ B$ להיות המכפלה שלהן איבר-איבר.

נגדיר בעיה חדשה

$$\min \overbrace{\sum_{j=1}^{k} D\left(w^j \circ \hat{p} \,\|\, \lambda_j p^j\right)}^{Q(w,\theta)} \quad \text{s.t} \quad \lambda \ge 0 \quad \sum \lambda_j = 1 \quad p^j = H$$

כאשר $w^1, \dots, w^k$ מערכים בגודל של $\hat{p}$.

אפשר להראות שמתקיים $L(\theta) \le Q(w,\theta)$ לכל $w \ge 0$ רק ש- $\sum w^j = 1$.

$$w^* = \arg\min_w Q(w,\theta)$$

$$(*) \qquad w_s^{*j} = \lambda_j p_s^j \,/\, \text{נרמול}$$

$$\Rightarrow \quad L(\theta) = Q(w^*,\theta)$$

$$\Rightarrow \quad \lambda^* = \arg\min_\lambda Q(w,\theta) \Rightarrow \lambda^* = \begin{pmatrix} \|w^1 \circ \hat{p}\|_1 \\ \vdots \\ \|w^k \circ \hat{p}\|_1 \end{pmatrix} \qquad (**)$$

EM

$t = 1, 2, \dots$

1) $w^{(t+1)} = (*)$

2) $\lambda^{(t+1)} = (**)$

3) פותר k בעיות נפרדות

$$\theta_j^{(t+1)} = \arg\min_{\theta_j} D\left(w^j \circ \hat{p} \,\|\, \lambda_j^{(t+1)} \cdot p^j\right) \quad \text{s.t.} \quad p^j \in H$$

העיקרון של EM הוא שמוסיפים משתנים חדשים וזה מאפשר להוציא את הסכימה החוצה ואז צריך לפתור k בעיות נפרדות וזה קל יותר.

שבוע הבא אין שיעור

עוד שבועיים ניפגש ביום שני בבוקר. - clustering

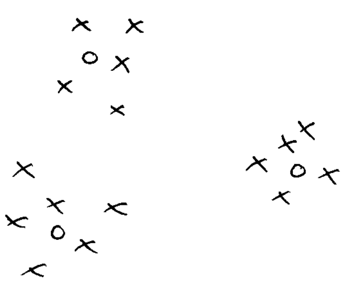
15 10.03.08
מערכות לומדות.

אשכול - Clustering

עסקנו בנושא הזה קודם כשדיברנו על נקודות שלקוחות מאוסף גאוסיאנים. בעזרת EM הצלחנו ללמוד את הפרמטרים שלהם ואפילו כשניתן נקודה חדשה יכולנו לשער מאיזה גאוסיאן היא הגיעה.

K-means

זה האלגוריתם הפשוט ביותר לאשכול. יש אוסף נקודות ומניחים שיש K אשכולות ואנחנו יודעים את K (זו גם חולשה של האלגוריתם כי במקרים רבים לא טריוויאלי שיודעים את מספר האשכולות). מניחים שלכל אשכול יש מרכז ואנחנו רוצים למצוא קבוצות קומפקטיות סביב המרכזים האלה.



הקלט של הבעיה: K - מספר האשכולות

$x_1,\dots,x_m\in\mathbb{R}^n$ - הנקודות

הפלט של הבעיה: $y_1,\dots,y_m\in\{1,\dots,K\}$ תיוגים של הנקודות.

למעשה יש לנו פה בעיית אופטימיזציה:

$$\min_{\substack{y_1,\dots,y_m\\ c_1,\dots,c_K}} \sum_{j=1}^{K} \sum_{y_i=j} \frac{1}{2}\|x_i-c_j\|^2$$

אופטימיזציה לסירוגין (alternating) -

1) נניח ש- $c_1,\dots,c_K$ ידועים. אז כמובן

$$y_i=\arg\min_j \|x_i-c_j\|^2$$

2) נניח ש- $y_1,\dots,y_m$ ידועים. לכל $S\subseteq\{1,\dots,m\}$

$$\arg\min_c \sum_{x_i\in S}\frac{1}{2}\|x_i-c\|^2 = \frac{1}{|S|}\sum_{x_i\in S}x_i$$

שכן

$$\frac{1}{2}\|x_i-c\|^2 = \frac{1}{2}(x_i-c)^t(x_i-c) =$$
$$= \frac{1}{2}x_i^t x_i - x_i^t c + \frac{1}{2}c^t c$$

אז אם נגזור ונשווה לאפס נקבל

$$\sum_i (x_i - c) = |S|c - \sum_i x_i = 0$$
$$\Rightarrow c = \frac{1}{|S|}\sum_i x_i$$

זה מגדיר לנו אלגוריתם. מתחילים מניחוש התחלתי של מרכזים. אז אפשר לחשב לפי (1) את התוויות. ואז לפי (2) אפשר לשנות שוב את המרכזים וככה לחזור על השלבים לסירוגין.
האלגוריתם הזה תמיד מתכנס <u>למינימום לוקאלי</u> של פונקציית האנרגיה.

<u>Spectral Clustering</u>

אנחנו הולכים להשתמש באותה פונקצית אנרגיה אלא שננסה למצוא לה פיתרון יותר טוב.
נרשום את הבעיה קצת אחרת. נגדיר את $\Psi_1,\dots,\Psi_k$ להיות הקבוצות שאנחנו מעוניינים בהן.

$$\min_{\substack{c_1,\dots,c_k\\ \Psi_1,\dots,\Psi_k}} \sum_{j=1}^{k}\sum_{x_i\in\Psi_j}\frac{1}{2}\|x_i - c_j\|^2$$

אנחנו יודעים ש- $c_j = \frac{1}{l_j}\sum_{x_i\in\Psi_j} x_i$ כאשר $l_j = |\Psi_j|$. נציב ונקבל

$$\min_{\substack{\Psi_1,\dots,\Psi_k\\ c_1,\dots,c_k}} \sum_{j=1}^{k}\sum_{x_i\in\Psi_j}(x_i^t x_i - 2x_i^t c_j + c_j^t c_j)$$

$$\equiv \min_{\Psi_1,\dots,\Psi_k}\sum_{j=1}^{k} l_j\cdot\frac{1}{l_j^2}\sum_{r,s\in\Psi_j} x_r^t x_s - 2\sum_{j=1}^{k}\frac{1}{l_j}\sum_{r,s\in\Psi_j}x_r^t x_s =$$

אפשר להשמיט את $x_i^t x_i$ כי הוא לא משפיע על המינימום.

$$= \min_{\Psi_1,\dots,\Psi_k} -\sum_{j=1}^{k}\frac{1}{l_j}\sum_{r,s\in\Psi_j}x_r^t x_s$$

$$\equiv \max_{\Psi_1,\dots,\Psi_k}\sum_{j=1}^{k}\frac{1}{l_j}\sum_{r,s\in\Psi_j}x_r^t x_s$$

קיבלנו משהו טוב! במקום מרחקים אוקלידיים קיבלנו ביטוי שתלוי רק במכפלות פנימיות. אז אפשר להשתמש ב- Kernel Trick.
נגדיר מטריצה $K = K(x_i, x_j) = \phi(x_i)^t\phi(x_j)$.
נגדיר מטריצה

$$F_{ij} = \begin{cases} \frac{1}{l_r} & x_i, x_j\in\Psi_r \\ 0 & \text{אחרת} \end{cases}$$

אם נמיין את הנקודות לפי השיוך לאשכולות אז F נראית למעשה כך:

$$F = \begin{bmatrix} F_1 & & 0 \\ & \ddots & \\ 0 & & F_k \end{bmatrix}$$

כאשר $F_r = \frac{1}{l_r}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ \vdots & & & & \vdots \\ 1 & \cdots & & & 1 \end{bmatrix}$

בגודל $l_r \times l_r$.

נשים לב ש- $\sum_{i,j} K_{ij} F_{ij} = \sum_{j=1}^{k} \frac{1}{l_j} \sum_{r,s} x_r^t x_s$

לכן אנחנו צריכים לפתור את

$$\max_F \sum_{i,j} K_{ij} F_{ij} = \max_F tr(KF)$$

מהם התנאים ההכרחיים ל- F?

נגדיר:

$$G_{ij} = \begin{cases} \frac{1}{\sqrt{l_j}} & x_i \in \psi_j \\ 0 & \text{אחרת} \end{cases}$$

$$G = [g_1, \dots, g_k]_{m \times k} \qquad g_r = \frac{1}{\sqrt{l_j}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Big\} \psi_r$$

$$\Rightarrow g_r g_r^t = \frac{1}{l_r} \begin{bmatrix} 0 & & \\ & 11^t & \\ & & 0 \end{bmatrix} = \begin{bmatrix} 0 & & \\ & F_r & \\ & & 0 \end{bmatrix}$$

$$\Rightarrow \sum_{r=1}^{k} g_r g_r^t = F$$

אז התנאים הם:

1) $F = GG^t$

2) $G \geq 0$

3) $F \cdot 1 = F^t 1 = 1 \Rightarrow GG^t 1 = 1$, כלומר F סטוכסטית כפולה - סכום השורות וסכום העמודות הוא 1.

4) $G^t G = I$

אז נכתוב שוב את הבעיה שלנו:

$$tr(KF) = tr(KGG^t) \underset{tr(AB) = tr(BA)}{=} tr(G^t K G)$$

⇐ צריך לפתור את הבעיה

$$\max_G tr(G^t K G) \quad s.t \quad \begin{array}{l} G \geq 0 \\ GG^t 1 = 1 \\ G^t G = I \end{array}$$

נשים לב שהבעיה הזו שקולה לבעיה של K-means. מסה"כ עשינו כאן מניפולציות אלגבריות על פונקציית האנרגיה.

תנאים מספיקים: G המטריצות המקיימות את התנאים $G \geq 0$, $GG^t 1 = 1$, $G^t G = I$ הן מהצורה

$$G_{ij} = \begin{cases} \frac{1}{\sqrt{l_j}} & x_i \in \Psi_j \\ 0 & \end{cases}$$

אנחנו נרצה למצוא אלגוריתם שנותן קירוב לפונקציית האנרגיה הנ"ל ומבוסס על ניתוח ספקטרלי.

נסביר את האילוצים:

- $G^t G = I$ פירושו שכל נקודה שייכת לאשכול אחד בלבד
- $GG^t 1 = 1$ נותן איזשהו "איזון" שנבין יותר בהמשך.

מה קורה אם דורשים ש-G סתם סטוכסטית ולא סטוכסטית כפולה?

$$\max_G tr(G^t K G) \quad s.t. \quad G \geq 0 \quad \underbrace{G^t G = D}_{\text{אורתוגונלית}} \quad \underbrace{G \cdot 1 = 1}_{\text{G סטוכסטית}}$$

<u>טענה:</u> אוסף המטריצות G המקיימות $G \geq 0$, $G1 = 1$, $g_i^t g_j = 0$ ל- $i \neq j$ הן מהצורה

$$G_{ij} = \begin{cases} 1 & x_i \in \Psi_j \\ 0 & \end{cases}$$

<u>הוכחה:</u>

$$0 = \sum_i G_{ir} G_{is} = g_r^t g_s = 0$$

⇐ בכל שורה של G יש רק כניסה אחת שונה אי-שלי מאפס.

מהתנאי ש- $G1 = 1$ מקבלים שיש בדיוק אחת כזאת. ☺

$$\Leftarrow \quad GG^t = \begin{bmatrix} 11^t & & 0 \\ & \ddots & \\ 0 & & 11^t \end{bmatrix}$$

נניח ש- $K=2$. אז

$$tr(G^tKG) = tr(KGG^t)$$

$$\Rightarrow \max tr(G^tKG) = \max \sum_{(i,j)\in\Psi_1} K_{ij} + \sum_{(i,j)\in\Psi_2} K_{ij}$$

$$\equiv \min \sum_{\substack{i\in\Psi_1 \\ j\in\Psi_2}} K_{ij}$$

זאת בדיוק בעיית min-cut - יש קודקודים בגרף וצריך למצוא את החתך המינימלי. הבעיה כאן היא שיכול להיות שהחתך מאוד לא מאוזן ולכן האלגוריתם יכול לתת לנו אשכולות לא מאוזנים. לכן היינו רוצים לקחת את הבעיה הזאת ולהכניס לה איזון מסוים וזה בדיוק מה שנעשה בבעיה המקורית שבה יש דרישה נוספת.

כפי שאמרנו, נוסיף את התנאי $G^t1 = \frac{m}{K}1$ זה אומר שבכל אשכול יש בדיוק $\frac{m}{K}$ נקודות (זה קצת קיצוני...) ויחד עם $G1=1$ זה נותן $GG^t1 = \frac{m}{K}1$.

מהתנאי $G^tG=D$ מקבלים $D=\frac{m}{K}I$ שכן:

$$G^tG = D \Rightarrow G^t\underbrace{G\cdot 1}_{=1} = D\cdot 1 \Rightarrow \underbrace{G^t 1}_{\frac{m}{K}1} = D\cdot 1 \Rightarrow D = \frac{m}{K}1$$

$$\Rightarrow \max_G tr(G^tKG) \text{ s.t } G\ge 0, GG^t1 = \frac{m}{K}1, G^tG = \frac{m}{K}1$$

כיוון ש- $\frac{m}{K}$ קבוע הוא לא משנה את האופטימיזציה ובדיוק יוצא

$$\max_G tr(G^tKG) \text{ s.t } G\ge 0, GG^t1 = 1, G^tG = 1$$

הבעיה הזו קצת מסובכת מדי, אז מחלקים אותה לשתי בעיות:

1) מצא K' סטוכסטית כפולה הקרובה ביותר ל- K (צריך להגדיר מה זה "קרוב ביותר")

2) פתור את הבעיה $\max_G tr(G^tK'G) \text{ s.t } G^tG = I$

וזאת זה אנחנו יודעים לפתור באמצעות ניתוח ספקטרלי.

למה האסטרטגיה הזאת טובה? אנחנו רוצים $\max tr(KGG^t)$
ולמעשה יש פה מין מכפלה פנימית. GG^t תהיה קרובה ל-K במובן של
מכפלה פנימית. אז אם נחליף את K במטריצה דומה שהיא
סטוכסטית כפולה K' והתוצאה תהיה GG^t "שדומה" ל-K' אז
GG^t "דומה" לסטוכסטית כפולה.
האילוץ $G \geq 0$ מסבך לנו את החיים, אז נתעלם ממנו ונקבל משהו
שאנחנו יודעים לפתור. ב-$K=2$ זה בעצם מספיק להתעלם מהאילוץ
הזה והיתרון של בעיה ספקטרלית הוא שאנחנו לא תלויים בניחוש
התחלתי. אם $K>2$ יש G מיני דרכים להתמודד עם זה.

Ratio - Cuts

נגדיר $D = diag(K1)$. (אם $V=(v_1,\dots,v_n)$ אז $diag\,v = \begin{pmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_n \end{pmatrix}$)

טענה: $K-D+I$ סטוכסטית כפולה הקרובה ביותר ל-K בשגיאת L_1 (נורמה $\|x\| = \sum |x_i|$).

הוכחה: $r = \min_F \|K-F\|_1 \quad s.t \quad F1=1,\ F^t 1 = 1$

לכל מטריצה A מתקיים $\|A\|_1 \geq \|A1\|_1$. לכן:

$$\|K-F\|_1 \geq \|(K-F)1\|_1$$

ול-F כלשהו. יתקיים

$$r \geq \|(K-F)1\|_1 = \|D1 - 1\|_1 = \|D-I\|_1$$

נציב $F = K-D+I$ ונקבל

$$K-(K-D+I) = D-I$$

$\Leftarrow \|D-I\|_1 = \|K-F\|_1$ ולכן זה אכן האופטימום. ☺

אם V ו"ע של $\overbrace{D-K}^{laplacian}$ עם ע"ע λ אז V ו"ע של $K-D+I$ עם ע"ע $1-\lambda$.
האלגוריתם אומר לקחת את הו"ע הכי קטן עם ע"ע שאינו 0 של $D-K$
ו"ע שני קטן ביותר של $D-K$ נקרא Fiedler Vector והוא נותן לנו
חלוקה לאשכולות. מחר נסביר הכל...

(18) 11.03.08

מעין זומדור

סכמה כללית

① מצא מטריצה סטוכסטית כפולה K' הקרובה ביותר ל- K

② מצא G: $\max_G tr(G^t K' G)$ s.t. $G \geq 0$ $G^t G = I$

טענה: $K' = K - D + I$ כאשר $D = diag(K1)$ קרובה ביותר ל- K במשמעות L_1.

ננסה לפתור את הבעיה ללא אילוץ החיוביות

$$\max_G tr(G^t K' G) \quad s.t \quad G^t G = I$$

נניח v ו"ע של $D-K$ עם ע"ע λ. אז

$$(K-D+I)v = -(D-K)v + v = -\lambda v + v = (1-\lambda)v$$

$\Leftarrow$ v ע"ע של K' עם ע"ע $1-\lambda$.

הו"ע המובילים של K' הם הו"ע עם הע"ע הקטנים ביותר של $D-K$.

(בגלל 6 מיני שיקולים של אלגברה לינארית)

נתבונן במקרה ש- $k=2$.

ברור ש- $(D-K)1 = D1 - K1 = 0$ $\Leftarrow$ $v_1 = 1$

ו"ע עם ע"ע $0 = \lambda_1$. בגלל ש- G סימטרית הו"ע שלה אורתונורמליים

אז אם v_2 הו"ע השני, אז $v_2^t v_1 = 0$. הו"ע v_2 נקרא Fiedler vector. בגלל ש- $v_2^t v_1 = 0$ ב- v_2 חייבים להיות גם ערכים חיוביים וגם שליליים. $v_2 \in \mathbb{R}^m$ - הקואורדינטות החיוביות קובעות מי מבין $x_m, \dots, x_1$ שייכות ל- Ψ_1 והקואורדינטות השליליות קובעות מי מבין $x_1, \dots, x_m$ שייכות ל- Ψ_2.

—

נרצה למצוא חתך מינימלי - $cut(\Psi_1, \Psi_2) = \sum_{i \in \Psi_1, j \in \Psi_2} K_{ij}$

גישת clustering:

מצא $x \in \mathbb{R}^m$ רק ש-

$$\min_x \frac{1}{2} \sum_{i,j=1}^m (x_i - x_j)^2 K_{ij} \quad s.t \quad x^t x = 1$$

אם נפתור אז הקואורדינטות של x יקבעו את החלוקה לאשכולות.

$$\frac12 \sum_{i,j}(x_i^2+x_j^2+2x_ix_j) = \frac12 \sum_i x_i^2 \sum_j k_{ij} + \frac12 \sum_j x_j^2 \sum_i k_{ij} - x^t K x =$$

$$= x^t D x - x^t K x = x^t(D-K)x$$

$\Leftarrow$ צריך לפתור את

$$\min_x x^t(D-K)x \quad s.t \quad x^t x = 1 \qquad (*)$$

בעיה כזאת כבר ראינו. באופן כללי הפיתרון של

$\min \frac12 x^t A x$ s.t $x^t x = 1$ (A סימטרית)

הוא הו"ע הגדול ביותר של A. (מה? נגזור:

$$Ax - \lambda x = 0 \Rightarrow Ax = \lambda x$$

$$\Rightarrow \frac12 x^t A x = \frac12 x^t \cdot \lambda x = \frac12 \lambda x^t x = \frac12 \lambda$$

אז רוצים את λ הקטן ביותר.

$\Leftarrow$ הפיתרון של (*) הוא הו"ע הכי קטן (כלומר ש"ק ע"ע הכי קטן) של $D-K$. את הו"ע הזה מוצאים עזרת Matlab.

נניח $G=(V,E)$ גרף מכוון. מטריצת תחולה של G הוא

$$B = \begin{bmatrix} 1 & \\ 0 & \\ \vdots & \cdots \\ -1 & \\ 0 & \end{bmatrix}_{|V|\times|E|}$$

1 מסמן קשת נכנסת
-1 מסמן קשת יוצאת

מתקיים $BB^t = D-K$ כאשר K מטריצת השכנויות בגרף.

normalized cuts

נגדיר $K' = D^{-1/2} K D^{-1/2}$ כאשר $D = diag(K1)$

טענה: (ללא הוכחה): נתבונן באלגוריתם האיטרטיבי:

$$K^{(0)} \leftarrow K$$

$$D^{(0)} \leftarrow diag(K)$$

$$K^{(t+1)} \leftarrow (D^{(t)})^{-\frac12} K^{(t)} (D^{(t)})^{-\frac12}$$

$$D^{(t+1)} \leftarrow diag(K^{(t)} 1)$$

אז $K^{(t)} \xrightarrow[t\to\infty]{} K'$ כאשר K' מטריצה סטוכסטית.

והיא הקרובה ביותר ל-K במשמעות אנטרופיה יחסית.

(19)

בתהליך הזה הצעד הראשון הוא הכי דומיננטי ולכן $K' = D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ הוא בקירוב סטוכסטית כפולה.

רוצים לפתור את הבעיה

$$\max_G tr(G^t (D^{-\frac{1}{2}} K D^{-\frac{1}{2}}) G) \quad s.t \quad G^t G = I$$

מה שמעניין אותנו הוא הו"ע השני של $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$

נפתח את האלגוריתם

נגדיר חתך $cut(A,B) = \sum_{i\in A} \sum_{j \in B} K_{ij}$

עבור $A,B = V$, $A \cap B = \emptyset$

$$sum(V_1, V_2) = \sum_{i\in V_1} \sum_{j\in V_2} K_{ij}$$

$$\Rightarrow \quad cut(A, V-A) = sum(A,V) - sum(A,A)$$

$$Ncuts(A,B) = \frac{cut(A,B)}{sum(A,V)} + \frac{cut(A,B)}{sum(B,V)}$$

הנרמול כל מונע אשכולות בגדלים מאוד שונים.

$$Nassoc(A,B) = \frac{sum(A,A)}{sum(A,V)} + \frac{sum(B,B)}{sum(B,V)}$$

$$\Rightarrow \quad Ncuts(A,B) = 2 - Nassoc(A,B)$$

$$\Rightarrow \quad \min Ncuts(A,B) \equiv \max Nassoc(A,B)$$

אם נפתור את הבעיה:

$$\max_{\psi_1,\dots,\psi_k} Nassoc(\psi_1,\dots,\psi_k) = \max \sum_{j=1}^{k} \frac{sum(\psi_j,\psi_j)}{sum(\psi_j,V)}$$

נגדיר $\bar{G} = [g_1,\dots,g_k]$ כאשר

$$g_j = \frac{1}{\sqrt{sum(\psi_j,V)}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \} \psi_j$$

$$\Rightarrow g_j^t K g_j = \frac{1}{sum(\psi_j,V)} \cdot \sum_{r,s\in\psi_j} K_{r,s}$$

$$tr(\bar{G}^t K \bar{G}) = \sum_j g_j^t K g_j = Naccos(\psi_1,\dots,\psi_k)$$

$$g_j^t D g = \frac{1}{sum(\psi_j, V)} \cdot \sum_{r\in\psi_j} d_r = \frac{1}{sum(\psi_j, V)} \cdot \sum_r \sum_s K_{r,s} =$$

$$= \frac{sum(\psi_j, V)}{sum(\psi_j, V)} = 1$$

$$\Rightarrow \bar{G}^t D G = I$$

נגדיר $G = D^{1/2}\bar{G}$ $\Leftarrow$ $G^t G = I$

$\bar{G} = D^{-\frac{1}{2}} G$. נציב את זה בבעיה שלנו ונקבל

$$\max_G \; tr((D^{-1/2}G)^t K (D^{-\frac{1}{2}}G)) \quad s.t \quad G^t G = I$$

במקרה זה v_1 הו"ע הראשון הוא $v_1 = D^{\frac{1}{2}} \cdot 1$ שהרי

$$(D^{-\frac{1}{2}} K D^{-\frac{1}{2}}) D^{\frac{1}{2}} 1 = D^{-\frac{1}{2}} K 1 =$$

$$= \begin{bmatrix} \frac{1}{\sqrt{d_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{d_m}} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix} = D^{-\frac{1}{2}} 1$$

טוען $D^{-\frac{1}{2}}1$ ו"ע עם ע"ע $\lambda = 1$.

הוקטור הזה שוב לא מכיל אינפורמציה. השני, בגלל שהוא אורתוגונלי מכיל ערכים חיוביים ושליליים ולכן זה שימושי אותנו לאשכול.

ישנם יותר משני אשכולות צריך להתמודד עם $G \geq 0$.
יש שתי שיטות להתמודד עם זה. בכל אופן הבעיה הנה NP - קשה!

בשבוע האחרון לסמסטר יהיו שני שיעורי תרגול - ימי שני ושלישי.

בשבועיים הקרובים נמשיך ללמוד על מודלים גרפיים

ניפגש ביום שלישי הבא.

(20) 18.03.08
מערכות
לומדות -

מודלים גרפיים Graphical Models

היום נציג את הבעיה ואת היישומים שלה ולא נאמץ את המוח יותר מדי. בשיעור הבא נדבר על אלגו' כוחות. מהותו של אמון יש קישורים לתחומים מאוד עדכניים בנושא.

עד היום התבוננו בשערוך של התפלגויות. המקרה הפשוט היה של התפלגות אחת. המקרה היותר מורכב של צירוף לינארי של התפלגויות פתרנו את EM.

במודלים גרפיים ידועה ההתפלגות ורוצים להסיק ממנה מסקנות. נניח ש- $x_i\in\{1,\dots,k\}$ ויש לנו $P(x_1,\dots,x_n)$

נניח שרוצים למצוא את $\max\limits_{x_1,\dots,x_n} P(x_1,\dots,x_n)$

גם כאשר יודעים את P יש פה בעיית חיפוש עם n^k מצבים שזה הרבה.

עוד דוגמה היא כשרוצים לחשב את $p(x_i)=\sum\limits_{\{x_1,\dots,x_n\}\setminus\{x_i\}} P(x_1,\dots,x_n)$

זו סכימה של n^{k-1} איברים וזה הרבה.

תזכורת: נניח $R(\hat\theta|x)=\sum\limits_{\theta} l(\theta,\hat\theta)\,p(\theta|x)$ ורוצים למצוא

$$\theta^*=\min_{\hat\theta} R(\hat\theta,x)$$

אם $l(\theta,\hat\theta)=\begin{cases}0 & \theta=\hat\theta\\ 1 & \text{אחרת}\end{cases}$ מקבלים

$$\theta^*_{MAP}=\arg\max_{\theta} P(\theta|x)$$

ואם $l(\theta,\hat\theta)=(\theta-\hat\theta)^2$ מקבלים

$$\theta^*_{MMSE}=\sum_{\theta}\theta\, p(\theta|x)$$

אז זאת המוטיבציה שבבעיה.

אנחנו נתעניין במקרה ש-

$$P(x_1,\dots,x_n)=\frac{1}{Z}\prod_{\alpha}\Psi_\alpha(X_\alpha)$$

כאשר $\alpha\subseteq\{1,\dots,n\}$; $X_\alpha\subseteq\{x_1,\dots,x_n\}$; $\Psi_\alpha\geq 0$

למשל

$$P(x_1,\dots,x_6)=\frac{1}{Z}\Psi_1(x_1)\cdots\Psi_6(x_6)\Psi_{124}(x_1,x_2,x_4)\Psi_{236}(x_2,x_3,x_6)\Psi_{135}(x_1,x_3,x_5)$$

מה שמיוחד כאן כרגע הוא שכל Ψ היא פונקציה של פחות משתנים ממה שהתחלנו וזה פשוט יותר לטיפול.

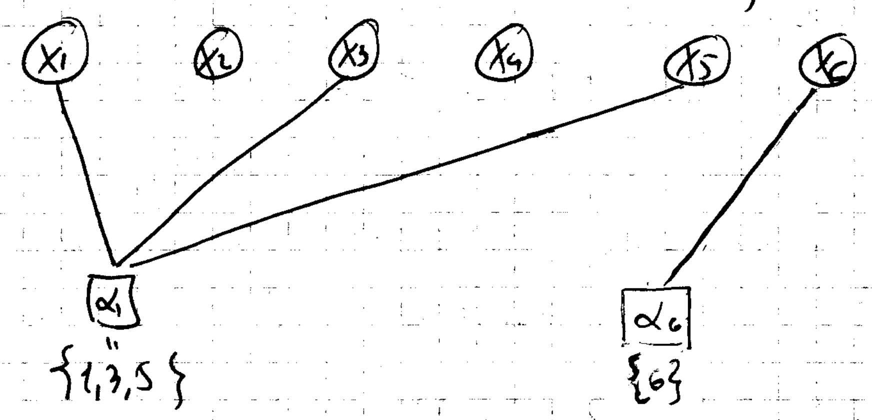
את הדבר הזה אפשר לתאר בגרף.

השיטה הכללית ביותר נקראת factor graph.

זה גרף דו צדדי שיש לו שני סוגי צמתים:

צמתי משתנים (variable nodes) וצמתים שמתאימים ל-α-ות.

במקרה מקודם



יש קשת בין צומת α לצומת x אם α כוללת את x.

במקרה ש- $\alpha=\{i\}$ מסמנים $\Psi_i(x_i)=\phi(x_i)$

ואז ב- $P(x_i|y_i)$ ל- x_i קוראים hidden node

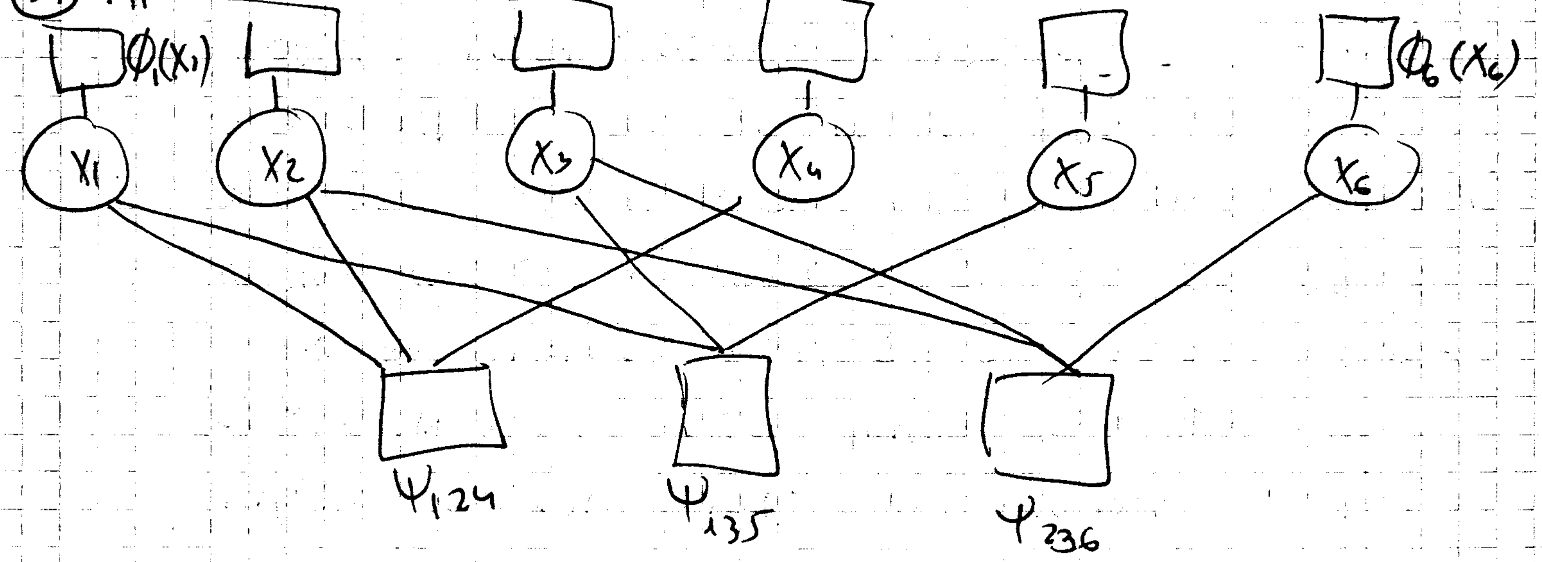
ול- y_i קוראים תצפית (observation node)

אז את $p(x_1,\dots,x_n|y_1\dots y_n)$ אפשר לפרק

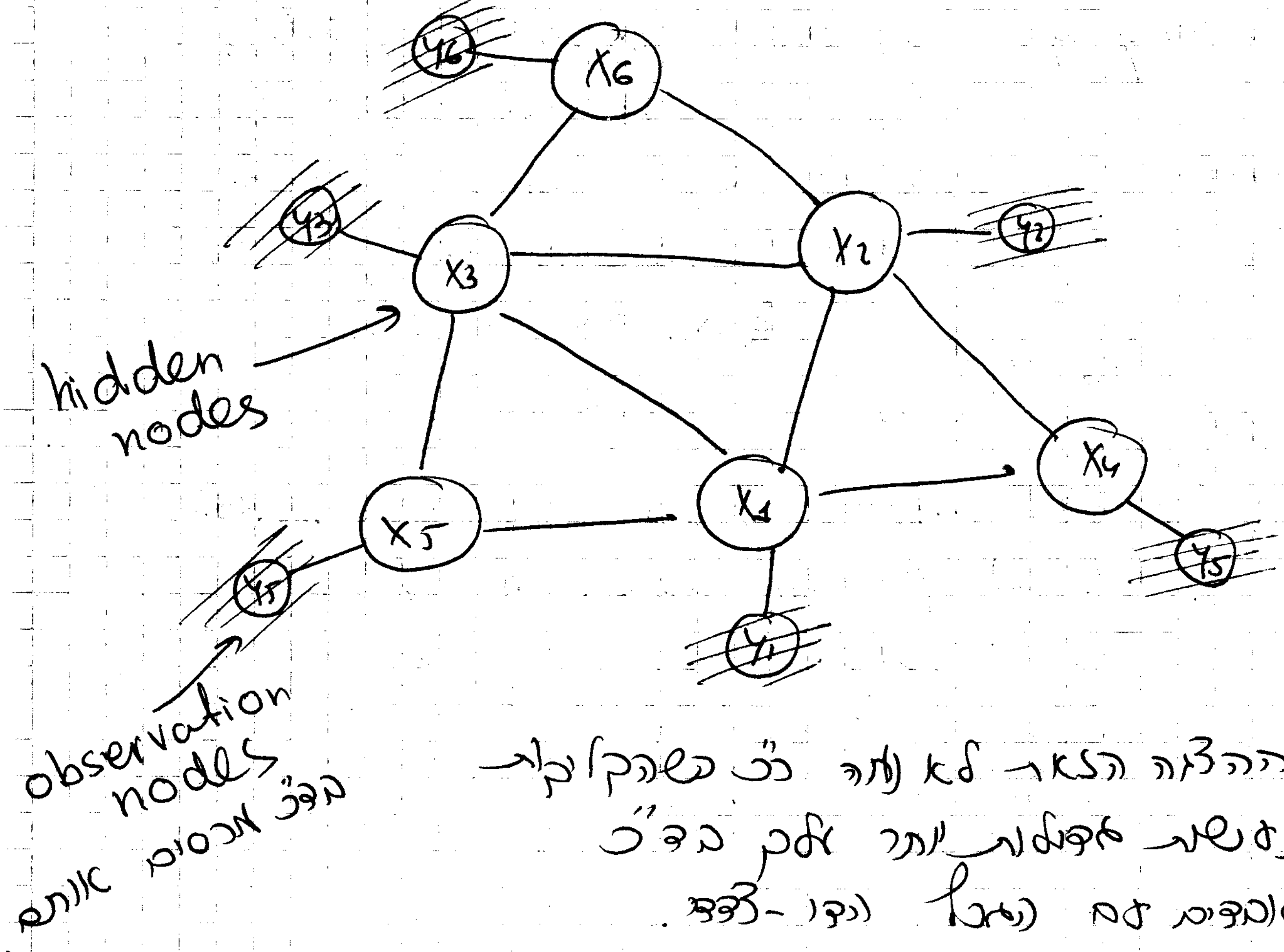
למכפלה של חלקים שיש בהם שני איברים וחלקים עם איבר אחד.

$$p(x_1\dots x_n|y_1\dots y_n)=\underbrace{\prod_i p(x_i|y_i)}_{=\phi_i(x_i)}\cdot\prod_{\alpha}\Psi_\alpha(X_\alpha)\quad(|\alpha|>1)$$

21 $P(x_1|y_1)$



עוד דרך לייצג את זה הוא ע"י גרף (רגיל) שהקליקות שלו מתאימות לקבוצות α. בדוגמא שלנו



ההצגה הזאת לא נוחה כ"כ כשהקליקות נעשות גדולות יותר אבל בד"כ עובדים עם הגרף הזה.

המטרה שלנו בתחום הזה לחשב את $p(x_i)$. ורוצים להוריד את הסיבוכיות מ- 2^n ל- n (גודל הקליקה המקסימלית). טוענים כדומה להשתמש בתכונות של הגרף.

מתברר שבמקרים אפשר לעשות את זה בדרך מאוד יעילה יש אלגוריתם BP שעובד אמ"מ הגרף הוא עץ אין

מעגלים. אפשר להפעיל את האלגוריתם גם על גרף שיש בו מעגלים אבל אז לא מובטח לגביו שום דבר.
אבל מסתבר שרוב הדוגמאות המעניינות הן עם מעגלים. התחום היה די מת במשך כ-10 שנים ואז הוא קם לתחיה עם החקר של משהו שנקרא turbo codes.
זה עורר את התחום מחדש והתחילו לפתח אלגוריתמים מקורבים שמובטח שיתכנסו גם במקרה שיש מעגלים.

עכשיו ניתן מגוון דוגמאות שממחישות למה רצוי לפתור את הבעיה של חישוב $P(x_i)$.

<u>Linear Codes</u> מתחום הזה factor graph
(נקרא tunnel graph).

נניח ש- $X_1 \dots X_n$ אלה הביטים של ה-codeword
Ψ_α הן ה- parity checks:

$$\Psi_\alpha = \begin{cases} 1 & \sum_\alpha X_\alpha = 0 \pmod 2 \\ 0 & \text{אחרת} \end{cases}$$

(123)
בדוגמה שלנו יש 3 ביטים שמיועדים להודעה ! 3
ביטים בשביל ה parity check (456).
אז יש 8 הודעות חוקיות:

$$\begin{array}{l|cccccccc} \text{הודעה} & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline \text{parity} & 0 & 0 & \dots \\ & 0 & 1 \\ & 0 & 1 \end{array}$$

$G = \begin{bmatrix} I_{k\times k} \\ P_{(n-k)\times k} \end{bmatrix}$ מטריצה יוצרת שבהנתן הודעה נותנת את ההודעה + ה parity שלה

למשל, במקרה שלנו $P = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ ואז $G\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$

נניח $r = Gs$ ונשדר s. ההודעה r -
ההודעה הנשלחת. מהודעת r צריך למצוא את s.
הבעיה היא שיש רעש בשליחה. אז למעשה $r = Gs + n$
כאשר n שגיאה.

אז מה שמעניין אותי הוא

$$p(x_1 \dots x_n | y_1 \dots y_n) = \prod p(x_i | y_i) \prod_\alpha \psi_\alpha(x_\alpha)$$

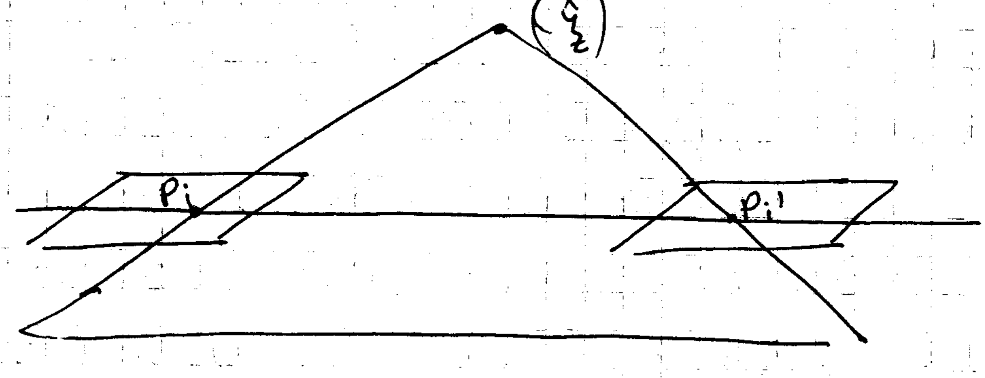
מהנחה שהביטים בלתי תלויים אלה באלה (זו הנחה סבירה). למה כופלים ב- $\psi_\alpha(x_\alpha)$? כי יש הודעות שיכולות להיות שהן לא חוקיות בכלל ואז ה- $\psi_\alpha(x_\alpha)$ מאפס את זה.

אם היינו מחשבים את $p(x_i) = \begin{pmatrix} p(x_i=0) \\ p(x_i=1) \end{pmatrix}$ למעשה היינו מקבלים code-word מתוקן - היינו בוחרים 0 או 1 בהתאם להסתברות הגבוהה יותר.

אפשר גם לחשב את $\arg\max_{x_1 \dots x_n} p(x_1 \dots x_n | y_1 \dots y_n)$ אבל מסתבר שהאלגוריתם שמחשב את זה פחות אמין מזה שמחשב את $p(x_i)$.

מסתבר ש- BP ברוב המקרים מתכנס אל הבעיה הנכונה ונותן תוצאה קרובה מאוד למה שנכון. אבל יכול להיות שה- codeword שנבחר לא חוקי ואז צריך לטפל בהתאם.

<u>Stereopsis</u> יש שתי תמונות מוזזות קצת ורוצים לקבל תמונת עומק.



לכל נקודה p_i מוצאים את הנקודה המתאימה p_i'
והתזוזה היחסית שלהן (disparity) פרופורציונית לעומק.
לכן כדי למצוא את תמונת העומק הצטמצמנו לבעיה של
מציאת ההתאמה $p_i \leftrightarrow p_i'$.

אז $\phi_i(x_i)$ = סיכוי לגודל תזוזה x_i על סמך התאמת תמונות.

בהנחה שההתאמה נמצאת באותו ישר עוברים לאורך הישר הזה ועושים קונבולוציה. נקודת המקסימום של הקונבולוציה מתאימה לגודל התזוזה. הבעיה היא שיש המון המון רעשים שמפריעים לשיטה הזאת לעבוד.

נסמן

עבור i,j שכנים
$$\psi_{ij}(x_i,x_j) = \begin{array}{c} n \\ \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \end{array} \rightarrow \psi_{ij}(x_i=r, x_j=s) = e^{-|r-s|}$$

$$p(x_1 \dots x_n) = \prod_{i=1}^{n} \phi_i(x_i) \prod_{i,j \in N(i)} \psi_{ij}(x_i,x_j)$$

המטריצה ψ_{ij} למעשה מעודדת שפיקסלים שכנים
יהיו דומים זה לזה.

הסיכוי לפיקסל i להיות בכל אחת מ-N ה-disparities
הוא
$$p(x_i) = \begin{pmatrix} p(x_i=1) \\ \vdots \\ p(x_i=N) \end{pmatrix}$$

וכאן משתמשים ב- BP.

Super Resolution יש תמונה. אם מגדילים אותה
מתחילה להיות טשטוש. היינו רוצים למלא את הטשטוש
הזה באינפורמציה. נצלם מלא תמונות שההבדל ביניהן הוא
בהרעדה של המצלמה. כך יודעים שלכל פיקסל יש כמה דגימות
ורק אפשר להקטין את הרעש. זה סופררזולוציה רגילה.
אבל מה אם בכלל יש רק תמונה אחת.

נניח שהתמונה שלנו באה ממשפחה מסויימת של תמונות (למשל, רק נוף או רק פרצופים).

נניח יש לנו תמונה 500x500 שרוצים להגדיל ל- 1000x1000.

נסתכל על מדגם אימון של תמונות בגודל 1000x1000 ונעשה להן down sampling ל- 500x500.

עכשיו נסתכל על פאצ'ים בגודל קטן (נניח 5x5). הדבר הפשוט לעשות הוא להחליף כל חלק קטן 5x5 בתמונה שלנו בפאצ' 10x10 הכי קרוב אליו מבין מאגר התמונות שיש לנו. אלא שזה יכול ליצור אי רציפויות בתמונה.

נניח שהפאצ'ים הם $x_1 \dots x_n$ נניח $x_i \in \{1,\dots,16\}$ (אחד מתוך 16 הקרובים ביותר (מבחינת ערכי אפור)). אנחנו רוצים גם

$$p(x_i) = \begin{pmatrix} p(x_i = 1) \\ \vdots \\ p(x_i = 16) \end{pmatrix}$$

אם ל- $p(x_i)$ ניקח בחשבון עם שכנים כמו רציפות אז נגדיר $\psi_{ij}(x_i, x_j) = e^{-d(x_i, x_j)}$, $j \in N(i)$ ואז

$$p(x_1 \dots x_n) = \prod_i \underbrace{\phi_i(x_i)}_{= e^{-d(x_i, y_i)}} \left(\prod_{i, j \in N(i)} \psi_{ij}(x_i, x_j) \right)$$

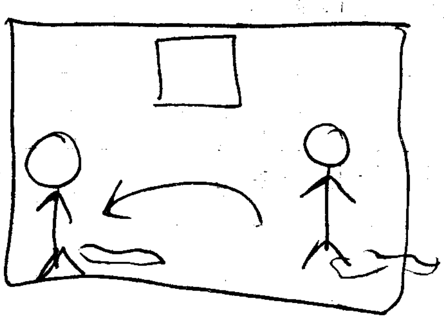
ידע אודות שכנות

גם כאן פותרים בעזרת BP והתוצאות יוצאות ממש טוב!!
(מאמר: Freeman, Jones, Pazstor 2002)

<u>Image Editing</u> (cho, יפה ב- CVPR 08)

יש תמונה שיש בה אובייקט כלשהו ורוצים להעביר אותו ממקום אחד למקום שני. כמובן רוצים שמשהו זה ייראה טוב!

מי שרוצה את הנושא
שישאר (אותו אי-את)



נחלק את התמונה לפאזלים לא חופפים $X_1,\dots,X_n$

לכל $x_i \in \{1,\dots,n\}$, n מיקומים השונים לכל אחד.

נכניס אילוץ של המשתמש – $\theta_i(x_i) = \begin{pmatrix}\theta_i(x_i=1)\\ \vdots \\ \theta_i(x_i=n)\end{pmatrix}$

למשל, אם רוצים להעביר את חלק i למקום k

$$\begin{cases}\theta_i(x_i=k)=1\\ \theta_i(x_i\neq k)=0\\ \theta_j(x_j=l)=e^{-(j-l)^2}\end{cases}$$

כך שהשאר הדברים יישארו פחות או יותר איפה שהיו במקור

וגם דומה לרציפות: $\psi_{i,j}(x_i,x_j) = e^{-\text{מדד מרחק בגבולות הפאזלים}}$. סה"כ

$$p(x_1\dots x_n) = \prod_i \phi_i(x_i) \prod_{i,j\in N(i)} \psi_{ij}(x_i,x_j)\underbrace{E(x_1\dots x_n)}_{\text{מקרבים ע"י } \prod E(x_i)}$$

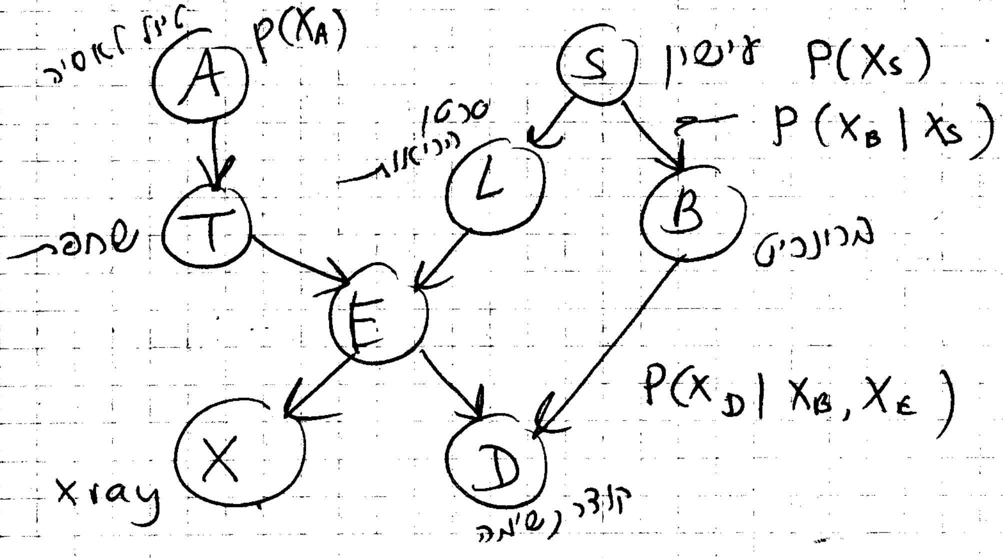
ה-E הזה דואג רק שלא יהיו חורים - שכל חלק עובר למקום אחר.

נשוב רוצים למדוד כאן את $p(x_i)$ כדי להחליט מה לבחור לאן.

רשתות בייסיאניות

דיאגנוסטיקה רפואית - יש סימפטומים, בדיקות ומחלות ועל סמך הקשרים שיש רוצים לתת אבחנות.

דוגמא "שחפת" (Lauritzen 1988)



(24)

לכל קשת רשום $P(X_i \mid par(X_i))$
הסיכוי של X_i בהינתן ההורים שלו.

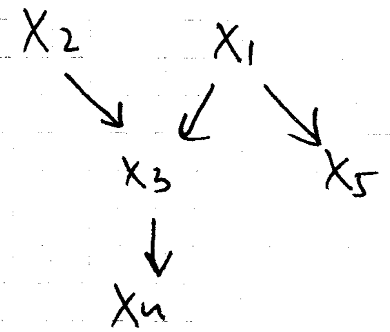
$$P(X_A \dots X_D) = \prod_i P(X_i \mid par(X_i)) =$$

$$= P(X_A)P(X_S)P(X_L \mid X_S)P(X_B \mid X_S)P(X_E \mid X_T, X_L)P(X_D \mid X_B, X_E)$$

ושואלים $P(X_D) = ?$

באופן כללי X_i לא תלוי במשתנים שאינם "ילדים" של X_i בהינתן ההורים שלו. נסמן $X_i \perp Z_i \mid par(X_i)$

למשל



$X_1 \perp X_2$

$X_4 \perp \{X_1, X_2, X_5\} \mid X_3$

$X_5 \perp \{X_2, X_3, X_4\} \mid X_1$

$$P(X_1 \dots X_5) = P(X_1)\underbrace{P(X_2 \mid X_1)}_{=P(X_2)}P(X_3 \mid X_1, X_2)\underbrace{P(X_4 \mid X_1, X_2, X_3)}_{=P(X_4 \mid X_3)}\underbrace{P(X_5 \mid X_1, X_2, X_3, X_4)}_{P(X_5 \mid X_1)}$$

בגלל האי-תלות

הרשת הבייסיאנית מתארת את האי תלויות בין המשתנים ומניחים שיודעים את ההתפלגויות המותנות ומה שרוצים לחשב זה את ה- marginals ($P(X_i)$).

בקיצור סיימנו את המבוא והמוטיבציה -

בשיעור הבא נפתח את BP ולאחר מכן נראה דברים יותר מתוחכמים...

ניפגש ביום שני ב- 8^{00}

מודלים גרפיים (המשך)

ב- notes-heaven יופיעו צילומים מהספר שהשיעור הזה לקוח ממנו.

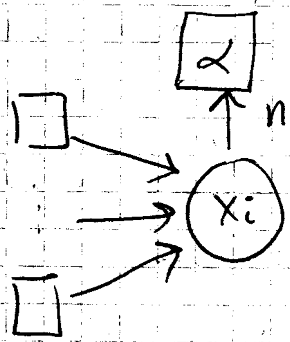
$$P(X_1,\dots,X_n) = \frac{1}{Z}\prod_{\alpha} \Psi_\alpha(X_\alpha)$$

BP

תיארנו את המצב בגרף:

ב- BP עוברות הודעות בין הצמתים.

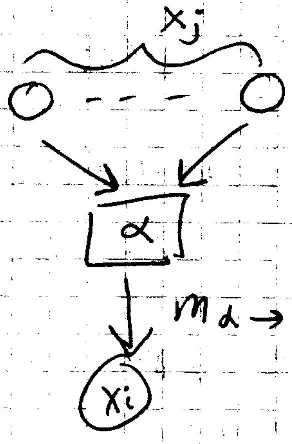
סוג 1:



$$n_{i\to\alpha}(x_i) = \prod_{\beta\in N(i)\setminus\alpha} m_{\beta\to i}(x_i)$$

זו מכפלה של הודעות שנשלחות מהשכנים של i אליו.

סוג 2:



$$m_{\alpha\to i}(x_i) = \sum_{X_\alpha\setminus x_i} \Psi_\alpha(X_\alpha)\prod_{j\in N(\alpha)\setminus i} n_{j\to\alpha}(x_j)$$

ההודעות מתחילות מהעלים של הגרף:

1) אם העלה הוא משתנה אז $n_{j\to\alpha}(x_j) \equiv 1$

2) אם העלה הוא α אז $m_{\alpha\to j}(x_j) = \sum_{X_\alpha\setminus x_j}\Psi_\alpha(X_\alpha)$

ודרשנו כל צומת שולחת הודעה אחרי שקיבלה את כל ההודעות מהשכנים שלא אליהם היא רוצה לשלוח.

אחרי שכל צומת קיבלה את מה שיש לה לקבל ושלחה את כל מה שיש לה לשלוח האלגוריתם גמר. אם מדובר בעץ האלגוריתם יעצור, ואם יש מעגלים אז הוא לא יעצור.

ראינו,

$$P(x_i) \cong \prod_{\alpha \in N(i)} m_{\alpha \to i}(x_i)$$

שווה עד כדי מכפלה בסקלר

$$P(X_\alpha) = \sum_{\{x_1 \dots x_n\} \setminus x_\alpha} P(x_1 \dots x_n) = \Psi_\alpha(X_\alpha) \prod_{i \in N(\alpha)} n_{i \to \alpha}(X_i)$$

אם כל ההודעות התכנסו אין בעיה לחשב את מה שדרוש.

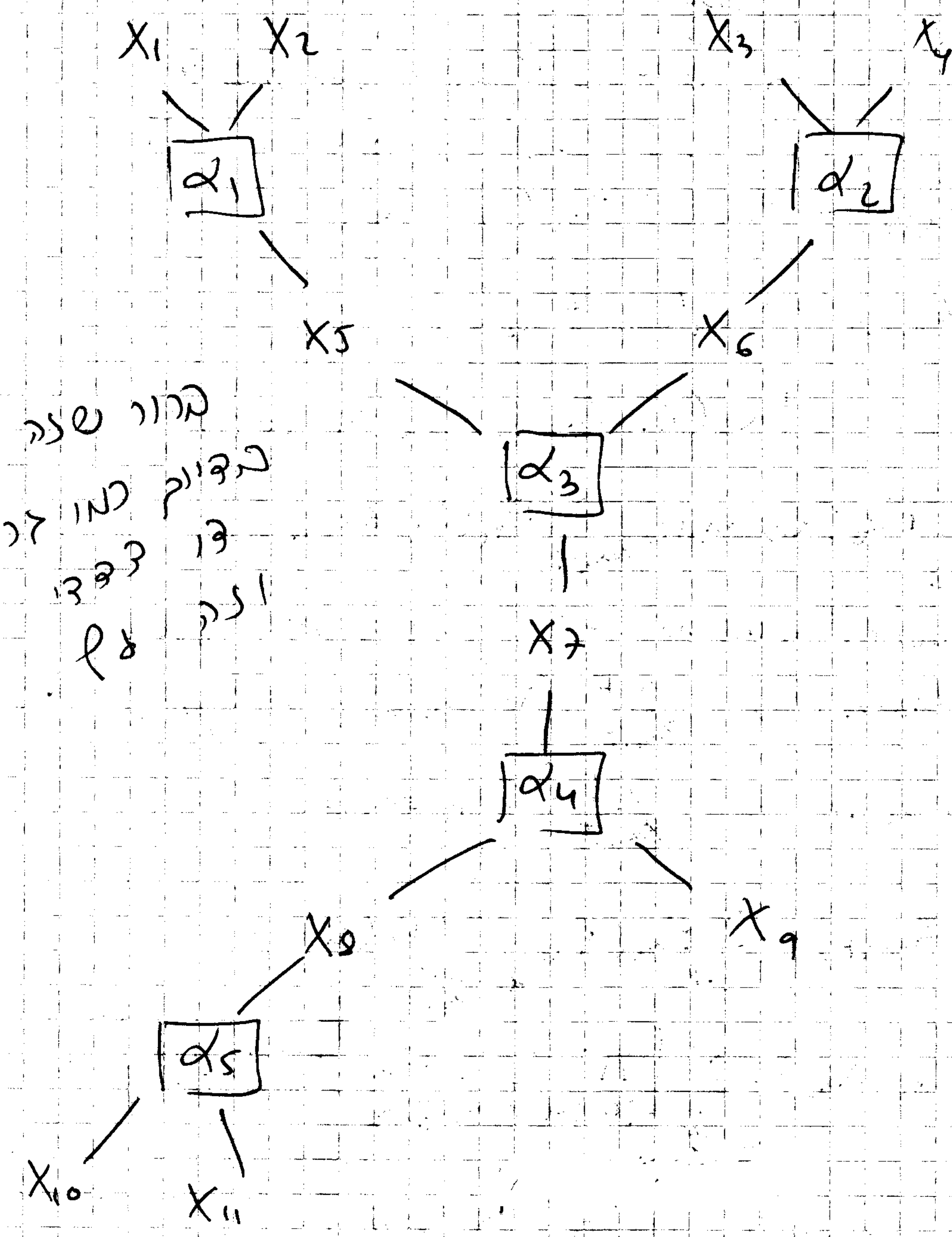
הסיבוכיות תלויה בדרגה של הצמתים. זה אלגוריתם אופטימלי מבחינת סיבוכיות וגם נכון כאשר ורק כאשר הגרף הוא עץ.

מסמנים $P(x_i) = b_i(x_i)$; $P(x_\alpha) = b_\alpha(x_\alpha)$.

אם הגרף לא עץ האלגוריתם יכול לפעול אבל לא ברור מה התוצאה שלו.

זה אלגוריתם טוב ויעיל שקל לתכנת, אלא שברוב המקרים המעניינים יש בגרף מעגלים ואז לא ברור מה קורה.

דוגמה:



ברור שזה
בדיוק כמו גרף
פו
קטפי
והוא עץ.

ההתפלגות המשותפת היא:

$$p(x_1,\dots x_{11}) = \Psi_{\alpha_3}(x_7, x_5, x_6)\,\Psi_{\alpha_1}(x_1, x_2, x_5)\,\Psi_{\alpha_2}(x_3, x_4, x_6)\cdot$$
$$\cdot\,\Psi_{\alpha_4}(x_7, x_8, x_9)\,\Psi_{\alpha_5}(x_{10}, x_{11}, x_8)$$

נחשב את $p(x_7)$. ל-x_7 יש תת עץ שהשורש שלו α_3 ותת עץ שהשורש שלו α_4.

נגדיר משתנים:

$$Z_{\alpha_3|7} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$
$$Z_{\alpha_4|7} = \{x_8, x_9, x_{10}, x_{11}\}$$

קבוצות זרות!
(כי זה עץ!)

נגדיר

$$F_\alpha(x_i, z_{\alpha|i}) = \prod_{\beta\in \ell^{\text{מתחת}}_{\alpha|i}} \psi_\beta(x_\beta)$$

מכפלה של הפונקציות אבל רק אלה שנמצאות מתחת לעץ α מנותקת מ i.

$$\Rightarrow F_{\alpha_3}(x_7, z_{\alpha_3|7}) = \psi_{\alpha_3}(\,)\,\psi_{\alpha_1}(\,)\,\psi_{\alpha_2}(\,)$$

$$F_{\alpha_4}(x_7, z_{\alpha_4|7}) = \psi_{\alpha_4}(\,)\,\psi_{\alpha_5}(\,)$$

$$\Rightarrow P(x_1,\dots,x_{11}) = F_{\alpha_3}(x_7, z_{\alpha_3|7})\cdot F_{\alpha_4}(x_7, z_{\alpha_4|7})$$

$$P(x_7) = \left(\sum_{z_{\alpha_3|7}} F_{\alpha_3}(x_7, z_{\alpha_3|7})\right)\left(\sum_{z_{\alpha_4|7}} F_{\alpha_4}(x_7, z_{\alpha_4|7})\right)$$

בגלל שהקבוצות זרות זה מתפרק למכפלת סכומים.

נסמן:

$$- \; m_{\alpha_3\to 7}(x_7) = \sum_{z_{\alpha_3|7}} F_{\alpha_3}(x_7, z_{\alpha_3|7})$$

$$- \; m_{\alpha_4\to 7}(x_7) = \sum_{z_{\alpha_4|7}} F_{\alpha_4}(x_7, z_{\alpha_4|7})$$

$$\sum_{z_{\alpha_3|7}} F_{\alpha_3}(x_7, z_{\alpha_3|7}) = \sum_{z_{\alpha_3|7}} \psi_{\alpha_3}(x_{\alpha_3}) \prod_{\alpha\in \ell^{\text{מתחת}}_{\alpha_3|7}} \psi_\alpha(x_\alpha) =$$

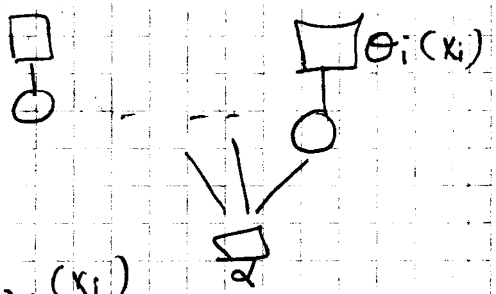
$$= \sum_{x_{\alpha_3}\setminus x_7} \psi_{\alpha_3}(x_{\alpha_3}) \sum_{\text{מה שנשאר}} \left[\prod_{j\in N(\alpha_3)\setminus 7} \left[\prod_{\alpha\in N(j)\setminus\alpha_3} F_\alpha(x_j, z_{\alpha|j})\right]\right] =$$

$$= \sum_{x_{\alpha_3}\setminus x_7} \psi_{\alpha_3}(x_{\alpha_3}) \underbrace{\prod_{j\in N(\alpha_3)\setminus 7} \prod_{\alpha\in N(j)\setminus\alpha_3} \underbrace{\sum_{z_{\alpha|j}} F_\alpha(x_j, z_{\alpha|j})}_{m_{\alpha\to j}(x_j)}}_{n_{j\to\alpha_3}(x_j)}$$

אלה בדיוק ההגדרות של ההודעות מאלגוריתם!
כמובן זה נכון לכל צומת ואז סה"כ מספר ההודעות הוא פעמיים מספר הקשתות בגרף וזהו זה עץ אז הכל טוב ויפה.

(27)

אם $P(x_1 \dots x_n) = \prod_{i=1}^{n} \theta_i(x_i) \prod_\alpha \psi_\alpha(x_\alpha)$



אז

$$n_{i \to \alpha}(x_i) = \prod_{\beta \in N(i) \setminus \alpha} m_{\beta \to i}(x_i)$$

$$m_{\alpha \to i}(x_i) = \sum_{x_\alpha \setminus x_i} \psi_\alpha(x_\alpha) \prod_{j \in N(\alpha) \setminus i} \theta_j(x_j)\, n_{j \to \alpha}(x_j)$$

$$p(x_i) = \prod_{\alpha \in N(i)} m_{\alpha \to i}(x_i)\, \theta_i(x_i)$$

עד כה ראינו נקודת מבט של הסקה בייסיאנית משנות ה-80. עכשיו נעבור לנקודת מבט יותר מודרנית.

הגישה המודרנית יוצאת מנקודת מבט של אופטימיזציה. ננסח הבעיה כבעיית אופטימיזציה.

$$p(x) = \frac{1}{Z} \prod_\alpha \psi_\alpha(x_\alpha)$$

מחפשים התפלגות b^*

$$\min_{\substack{b \\ b \geq 0 \\ \|b\|_1 = 1}} D\left(b \,\|\, \prod \psi_\alpha(x_\alpha)\right) \Rightarrow b^*(x) = P(x) = \frac{1}{Z} \prod_\alpha \psi_\alpha(x_\alpha)$$

זה נרמול שמגיע מהדרישה ש- $\|b\|_1 = 1$

$$D\left(b \,\|\, \prod \psi_\alpha(x_\alpha)\right) = \sum_\alpha \sum_{x_\alpha} \underbrace{E_\alpha(x_\alpha)}_{-\ln \psi_\alpha(x_\alpha)} b_\alpha(x_\alpha) - H(b)$$

זה נקרא free energy.

את המחובר הראשון לא בעיה לחשב באופן יעיל, אבל מסובך לחשב את $H(b)$ זה עדיין אקספוננציאלי.
אז המטרה היא למצוא קירוב לאנטרופיה שקל לחשב אותו והוא מדויק לעל.

טענה 1: בעץ

$$P(x_1,\dots,x_n) = \frac{\prod_\alpha P(x_\alpha)}{\prod_{i=1}^{n} P(x_i)^{d_i-1}}$$

כאשר d_i = דרגת צומת משתנה x_i.

טענה 2:

$$H(P) \underset{\text{במקרה של עץ}}{=} H_{Bethe}(P) = \sum_\alpha \sum_{x_\alpha} P(x_\alpha)\log P(x_\alpha) - \sum_i (1-d_i)\sum_{x_i} P(x_i)\log P(x_i) =$$

$$= \sum_\alpha H(P_\alpha) + \sum_i (1-d_i) H(P_i)$$

ערכי המורכבות הם לא אקספוננציאלית אלא תלויה של הקליקה המקסימאלית.

Bethe free energy

$$\min_b \sum_\alpha \sum_{x_\alpha} E_\alpha(x_\alpha) b_\alpha(x_\alpha) - \sum_\alpha H(b_\alpha) - \sum_i (1-d_i) H(b_i)$$

s.t.

$$\sum_{x_\alpha \setminus x_i} b_\alpha(x_\alpha) = b_i(x_i) \quad \forall \alpha \in N(i)$$

$$b \geq 0$$

$$\sum_{x_\alpha} b_\alpha(x_\alpha) = 1 \quad \forall \alpha$$

טענה (ידידיה, פרימן, וייס 2005): נקודות שבת של Bethe free energy נותנות את $b_\alpha(x_\alpha)$, $b_i(x_i)$ של BP.

מסקנות:

1. בעץ, Bethe free energy האופטימיזציה קמורה בתחום האילוצים. $\Leftarrow$ יש לה מינימום יחיד וזה ה marginals
2. לגרף עם מעגל, Bethe free energy אינו קמורה.

(28)

יש אלגוריתמים שפותרים בעות כללית ומוצאים מינימום לוקאלי (CCCP) אבל הם מאוד מאוד מסובכים אז זה לא תפס.

גם כלל אי אפשר לתאר כסכום של ...פורשים.
אז אולי לפתור את הבעיה לעצים האלה ייתן קירוב של הפיתרון הנכון (tree reweighted)

הכיוון היותר עדכני הוא Convex free energy

שיעור תגבור:

10-14 29/4

שבוע הבא אין שיעורים.

29

29.04.08
מני טומפור

שיעור חזרה למבחן

* החומר של מודלים גרפיים: מדויק מה שמופיע בסיכומים שלי - לא פחות ולא יותר. - ידיעת BP ואת השימושים!

* נושאים להיום:

- שימוש בחוק בייס
- מציאת אומדן ML
- צירוף התפלגויות ל- EM
- מעבר מ- primal ל- dual.

התרגילים שנפתור היום לא יופיעו בבחינה אבל הם דומים!

$\boxed{\text{שאלה 1}}$

קוביה הוטלה 20 פעם עם תוצאות ב... 1 2 1 3 3 4 ...
מצא אומדן ML להסתברות פאות.

$D = x_1, \dots, x_{20}$ וצריך לפתור את

$$\max_\theta \log P(D|\theta) = \max \log \prod_{i=1}^{20} p(x_i|\theta) = \quad s.t\ \sum p_i = 1$$

$$= \max_{p_1 \dots p_6} \underbrace{\log [p_1^6 p_2^4 p_3^3 p_4^3 p_5^2 p_6^2]}_{} \quad s.t.\ \sum p_i = 1$$

$$= 6\log p_1 + \dots + 2\log p_6$$

נגזור את הלגרנז'יאן:

$$\frac{\partial \mathcal{L}}{\partial p_1} = \frac{6}{p_1} - \lambda = 0 \Rightarrow p_1 = \frac{1}{\lambda}\cdot 6$$

$$\vdots$$

$$\frac{\partial \mathcal{L}}{\partial p_6} = \frac{2}{p_6} - \lambda = 0 \Rightarrow p_6 = \frac{1}{\lambda}\cdot 2$$

אבל $\sum p_i = 1 \Leftarrow \lambda = 20$.
וקיבלנו תוצאה הגיונית: ההסתברויות הן יחס הופעת הפאות במדגם.

הערה: בכתה ראינו הכר שזה מה שאמור לצאת. במבחן מצופה מאיתנו לפתח את זה מחדש.

$\boxed{\text{שאלה 2}}$

מצא ML של

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

(זאת התפלגות log-normal והיא די איזוטרית - אז לא נתמקד בה, אלא בטכניקה)

$$\max_\theta \log p(x_1 \dots x_m \mid \theta) =$$

$$= \max_\theta \prod_{i=1}^m f(x_i \mid \mu, \sigma) =$$

$$= \max_{\mu, \theta} \underbrace{\sum_{i=1}^m \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma x_i}\right) + \log\left(e^{-\frac{(\log x_i - \mu)^2}{2\sigma^2}}\right)\right]}$$

$$= -\frac{m}{2}\log 2\pi - m\log\sigma - \sum_{i=1}^m \log x_i - \sum_{i=1}^m \frac{(\ln x_i - \mu)^2}{2\sigma^2}$$

אין אילוצים אז פשוט נגזור ונשווה לאפס:

$$\frac{\partial}{\partial\mu} = \frac{1}{\sigma^2} \cdot \sum_{i=1}^m (\ln x_i - \mu) = 0 \Rightarrow \mu = \frac{1}{m}\sum_{i=1}^m \ln x_i$$

$$\frac{\partial}{\partial\sigma} = -\frac{m}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^m (\ln x_i - \mu)^2 = 0 \Rightarrow \sigma^2 = \frac{1}{m}\sum_{i=1}^m (\ln x_i - \mu)^2$$

שאלה 3

מטבע הוגן ושתי קוביות - אחת הוגנת והשנייה תמיד נופלת על פאה "1".

מבצעים ניסוי 3 פעמים: מטילים מטבע - אם H אז מטילים קוביה הוגנת. אחרת, הקוביה השניה.

נניח שקיבלנו "1 1 1".

מה הסיכוי שהמטבע נפל HHH ?

ההטלות כמובן בלתי תלויות. אזי

$$p(HHH \mid 111) = p(H \mid 1)^3$$

כרגע נשתמש בנוסחת ב"ס -

$$p(H \mid 1) = \frac{p(1 \mid H)\cdot p(H)}{p(1)} = \frac{1/6 \cdot 1/2}{\frac{1}{6}\cdot\frac{1}{2} + 1\cdot\frac{1}{2}} = \frac{1}{7}$$

$$\Rightarrow p(HHH \mid 111) = \frac{1}{7^3}$$

שאלה 4

שאלה חשוב שיש להבין היטב. קשה ממנה גופים במבחן!

$$\min \tfrac{1}{2} w^t w \quad s.t \quad y_i(w^t x_i) \geq 1 \qquad y_i = \pm 1$$

זו בעיית SVM כאשר $b=0$.

אנחנו נפתור את בע"י ridge-regression:

$$\min \tfrac{1}{2} w^t w + \Sigma (y_i - w^t x_i)^2$$

ניתן לרשום זאת באופן שקול:

$$\min \tfrac{1}{2} w^t w + \Sigma \varepsilon_i^2 \quad s.t \quad \varepsilon_i = y_i - w^t x_i$$

וזה משהו שיותר קל לפתור.

נרשום את הלגרנג'יאן:

$$\mathcal{L} = \frac{1}{2} w^t w + \sum_i \varepsilon_i^2 + \sum_i \lambda(y_i - w^t x_i - \varepsilon_i)$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum \lambda_i x_i = 0 \implies w = \sum \lambda_i x_i$$

אז w הוא קומבינציה לינארית של הנקודות $x_1 \dots x_m$

$$\frac{\partial \mathcal{L}}{\varepsilon_i} = 2\varepsilon_i - \lambda_i = 0 \implies \varepsilon_i = \frac{1}{2}\lambda_i$$

נסמן $Q = [x_1 \dots x_m]$. אז $w = Q \cdot \lambda$

נציב את מה שקיבלנו בלגרנג'יאן:

$$\mathcal{L} = \frac{1}{2}\lambda^t Q^t Q \lambda + \frac{1}{4}\sum_i \lambda_i^2 - \frac{1}{2}\sum \lambda_i^2 +$$

$$+ \sum \lambda_i y_i - \lambda^t Q^t Q \lambda =$$

$$= -\frac{1}{2}\lambda^t Q^t Q \lambda - \frac{1}{4}\lambda^t \lambda + \lambda^t y$$

זאת הבעיה הדואלית ואותה אנחנו רוצים למקסם.

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -Q^t Q \lambda - \frac{1}{2}\lambda + y = 0$$

$$\implies \lambda = (Q^t Q + \tfrac{1}{2} I)^{-1} y$$

$$\implies w = Q\lambda = Q(Q^t Q + \tfrac{1}{2} I)^{-1} y$$

בהינתן x נמצא את y:

$$y = w^t x = y^t (Q^t Q + \tfrac{1}{2} I)^{-1} Q^t x$$

מה קיבלנו? נוסחת רגרסיה ששימוש ב-y מה שגיאות הקודמות...

שאלה 5

צירוף קמור של שתי התפלגויות log-normal:

$$g(x) = \lambda f(x|\mu_1,\sigma_1) + (1-\lambda) f(x|\mu_2,\sigma_2)$$

נתונים $x_1,\dots,x_m$ מצא ML עבור ה- $\mu_1, \sigma_1, \mu_2, \sigma_2$

יש פה משתנה חבוי y שאומר אם הדגימה הגיעה מ- (μ_1,σ_1) או מ- (μ_2,σ_2), כלומר

$$p(x_1\dots x_m) = \sum_y p(x_1\dots x_m|y)\,p(y)$$

הדגימות הן iid ונתחיל מנקודה הנחה כזו:

$$\theta^{(t+1)} = \max_\theta \sum_{i=1}^m \underbrace{p(y_i=1|x_i,\theta^{(t)})}_{w_i^{(t)}} \log\big[\underbrace{p(x_i|y_i=1,\theta)}_{f(x_i|\mu_1,\sigma_1)}\underbrace{p(y_i=1|\theta)}_{\lambda}\big]$$
$$+\sum_{i=1}^m \underbrace{p(y_i=2|x_i,\theta^{(t)})}_{1-w_i^{(t)}} \log\big[\underbrace{p(x_i|y_i=2,\theta)}_{f(x_i|\mu_2,\sigma_2)}\underbrace{p(y_i=2|\theta)}_{1-\lambda}\big]$$

נפתח את זה ונתעלם מכל מה שלא קשור לפרמטרים, כי בכל אופן הוא "יפול" מגזירה

אז צריך לפתור:

$$\max_\theta \Big[\sum w_i^{(t)}\big[\log\lambda - \log\sigma_1 - \tfrac{1}{2\sigma_1^2}(\log x_i-\mu_1)^2\big] + \sum(1-w_i^{(t)})\big[\log(1-\lambda) - \log\sigma_2 - \tfrac{1}{2\sigma_2^2}(\log x_i-\mu_2)^2\big]\Big]$$

$$\frac{\partial}{\partial\lambda} = \sum_i\Big[\frac{w_i^{(t)}}{\lambda} - \frac{(1-w_i^{(t)})}{1-\lambda}\Big] = 0$$

הערה חשובה: במקרה הכללי אסור לשכוח את האילוץ $\sum\lambda_i = 1$!

נפתח:

$$\frac{1}{\lambda}\sum_i w_i^{(t)} - \frac{1}{1-\lambda}\sum_i (1-w_i^{(t)}) = 0$$

$$\Rightarrow (1-\lambda)\sum w_i^{(t)} - \lambda \sum (1-w_i^{(t)}) = 0$$

$$\Rightarrow \sum w_i^{(t)} - \lambda\left[\sum w_i^{(t)} + \sum (1-w_i^{(t)})\right] = 0$$

$$\Rightarrow \lambda^{(t+1)} = \frac{1}{m}\sum w_i^{(t)}$$

את w_i מעדכנים בעזרת נוסחת ב"ס:

$$w_i^{(t+1)} = p(y_i=1 \mid x_i, \theta^{(t+1)}) =$$

$$= \frac{p(x_i \mid y_i=1, \theta^{(t+1)})\, p(y_i=1 \mid \theta^{(t+1)})}{p(x_i \mid \theta^{(t+1)})} =$$

$$= \frac{f(x_i \mid \mu_1^{(t+1)}, \sigma_1^{(t+1)})\,\lambda^{(t+1)}}{f(x_i \mid \mu_1^{(t+1)}, \sigma_1^{(t+1)})\,\lambda^{(t+1)} + f(x_i \mid \mu_2^{(t+1)}, \sigma_2^{(t+1)})\,(1-\lambda^{(t+1)})}$$

המבחן קל - אין טריקים!

אם יודעים את החומר מקבלים 100

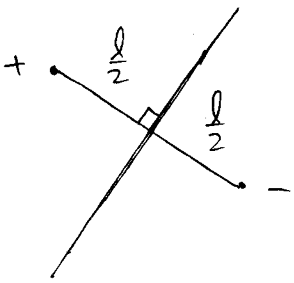
32

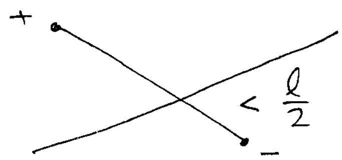
2005 א שאלה 3

(1) מדגם S $\qquad h(x) = \text{sign}(w^t x + b)$

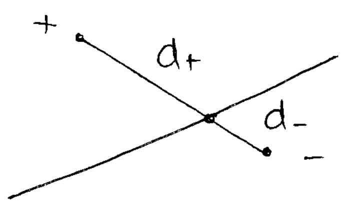
$$\gamma = \min_i \left\{ \frac{y_i(w^t x_i + b)}{\|w\|} \right\}$$

(2) המרחק המקסימלי בין שתי דגימות - חיוביות ושליליות הוא l. נוכיח מראין ש- $\gamma \le \frac{l}{2}$.
כי אם המישור המפריד מאונך לישר בין הנקודות ותחצה אותו אזי אז השוליים שווים $\frac{l}{2}$ ואם הוא לא מאונך אז השוליים קטן מ- $\frac{l}{2}$.





באופן יותר פורמלי:



$$l = d_+ + d_-$$

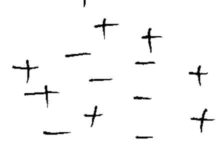
$$\gamma \le \min(d_+, d_-) \le \frac{l}{2}$$

(3) יש מרחב $\mathbb{R}^n$ עם נקודות - $x_1, \dots, x_m$ שלא בהכרח ניתנות להפרדה לינארית.
אנחנו לוקחים את הנקודות למרחב $\mathbb{R}^{n+m}$ ושם אנחנו רוצים שהן תהיינה ניתנות להפרדה.
נגדיר העתקה באופן הבא:

$$x_i \longmapsto (x_i^1 \dots x_i^n \; 0 \dots 0 \; \underset{\text{מקום } n+i}{1} \; 0 \dots 0) = x_i'$$

למה המדגם החדש S' ניתן להפרדה?
כי לכל תיוג ניקח $W = [0 \dots y_1 \dots y_m]$ ואז $W^t x_i' = y_i$

אם זה המפריד שקיבלנו ב- $\mathbb{R}^{n+m}$ ונחזיר חזרה ל- $\mathbb{R}^n$
אין שום הבטחה שזה יעבוד שם. הרי מלכתחילה
זה לא בטוח ניתן להפרדה, למשל:



(4) נגדיר $K(x,y)=x^t y$.

צריך להגדיר K' רק שהפעלת האלגוריתם S' (שמעביר את x מ- $\mathbb{R}^n$ ל- $\mathbb{R}^{n+m}$ ואז מפעיל עליו את S שיודע להפריד נקודות ניתנות להפרדה) עם K תהיה שקולה להפעלת האלגוריתם S עם K'.
אז פשוט נגדיר

$$K'(x_i,x_j)=\begin{cases} x_i^t x_j & i\neq j \\ x_i^t x_j+1 & i=j \end{cases}$$

$\boxed{\text{2005 ב' שאלה 1}}$

יש נקודות $x_1,\dots,x_m\in\mathbb{R}^d$ $m>d$ ומריצים עליהן PCA רק שיוצא $\hat{x}=UU^t x$ כאשר U התקבל מ- PCA. נתון ש- $\|x\|=1$.
נחשב את $\|x-\hat{x}\|^2$.

מתקיים $x=\hat{x}+(x-\hat{x})$. מהאלגוריתם מתקבל $\hat{x}\perp x-\hat{x}$ לכן

$$\|x\|^2=\|\hat{x}\|^2+\|x-\hat{x}\|^2$$

$$\Rightarrow \|x-\hat{x}\|^2=1-\|\hat{x}\|^2=$$

$$=1-\|UU^t x\|^2 \underset{\uparrow}{=} \|U^t x\|^2$$

כי U אורתוגונלית

2007 מועד ב' שאלה 2

1) צ"ל: $VC\text{-}dim(H_1\cap H_2) \le \min(VC\text{-}dim(H_1), VC\text{-}dim(H_2))$

מימד VC הוא המספר המקסימלי d כך שיש d נקודות שניתן להשיג את כל 2^d תיוגים עם היפותזות מ- H.

נניח $d = \min(VC_1, VC_2)$ ונניח בשלילה ש- $VC\text{-}dim(H_1\cap H_2) > d$

בה"כ $d = VC_1$, אז
$VC\text{-}dim(H_1\cap H_2) > VC\text{-}dim(H_1)$

⇐ קיימות d+1 נקודות כך שיש 2^{d+1} תיוגים אפשריים ע"י היפותזות ב- $H_1\cap H_2$. אבל אם איזו היפותזה היא ב- $H_1\cap H_2$ אז בפרט היא ב- H_1 ולכן מתייגת d+1 נקודות בכל הדרכים האפשריות. וזה בסתירה לכך ש- $VC\text{-}dim(H_1) = d$

2) הדוגמאות הטריוויאליות הן כאשר $H_1 = H_2$ או $H_1 \subseteq H_2$.

3) ידוע ש- $\Pi_H(m) \le \left(\frac{em}{d}\right)^d$ כאשר d הוא ממד VC של H. יהיו H_1, H_2 מחלקות היפותזות כך ש- $VC\text{-}dim(H_1) = VC\text{-}dim(H_2) = 1$. צריך להראות ש- $VC\text{-}dim(H_1\cup H_2) < 5$.

מתקיים $\Pi_{H_2}(5), \Pi_{H_1}(5) \le e\cdot 5$

נניח ש- $VC\text{-}dim(H_1\cup H_2) = d$. אז יש d נקודות כך שאפשר לתייג אותן בכל 2^d הדרכים.
H_1 יכולה לתת לכל היותר ed תיוגים וכך גם H_2. לכן סה"כ $H_1\cup H_2$ יכולה לתת לכל היותר

vcd תיוגים ל-d הנקודות ולא בהכרח
$vcd \ge 2^d$. נבדוק את כל ה-d-ים האפשריים.
הראשון שזה לא מתקיים עבורו הוא החסם
למימד.

2007 א' שאלה 6

זאת שאלה קטנה אז התשובה היא פשוט
PCA - בלי פיתוחים.

2005 א' שאלה 2

1. ברור. הגדרנו את err עם $\hat{err}$.

2. צריך לתת דוגמא למרחב מדגם X עם פילוג
$\mathcal{D}$ ונקודות $x_1, \dots, x_m$ רק ש-
$\hat{err}(h) = 0$ $\quad err(h) \ge \frac{1}{2}$
עבור היפותזה $h \in C$.

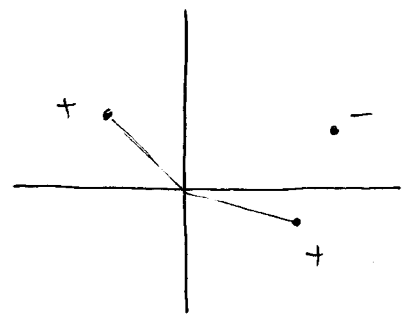
ניקח $X = \mathbb{R}$ ופילוג כלשהו - נניח גאוסיאני
! H - יכולה לתת כל תיוג לכל נקודה ומרגע
שמקבלים איזשהו מדגם עליו נותן את התשובה
הנכונה ולכל X שלא במדגם נותן תשובה נכונה
בהסתברות $\frac{1}{2}$ - כמו הטלת מטבע.

3. דוגמא נגדית לזה ומעשה הסעיף הקודם.
מה לא נכון בהוכחה של הרשלה? ברגע שבוחרים
היפותזה תוך הסתכלות על המדגם משפט הופדינג
הוא דבר לא נכון.

4. הוכחנו בתרגול

2007 א' שאלה 1

(1) אם יש 3 נקודות מהכרח יש שתיים שהזווית ביניהן פחות מ- 180° אז אם נתייג אותן ב-(+) ואת השלישית ב-(-) אי אפשר להפריד:



(2) נגדיר $K(a,b)=(1+ab)^2$ עבור $a,b\in\mathbb{R}$
נראה ש-K פונקצית גרעין: ניקח
$$\varphi(a)=(1,\sqrt{2},a,a^2)$$

2007 ב' שאלה 1

(1) $K_1,\dots,K_n$ גרעינים PSD (positive semi definite)
$\alpha_1,\dots,\alpha_n\ge 0$ שלמים אי שליליים. אז $K=\sum\alpha_i K_i$ גרעין

$$v^tKv=v^t\sum\alpha_i K_i v=\sum\underbrace{\alpha_i}_{\substack{\ge 0\\ \text{נתון}}}\underbrace{v^tK_iv}_{\substack{\ge 0\\ \text{כי PSD}}}\ge 0$$

$\Leftarrow$ זה גרעין.

(2) $w\in\mathbb{R}^n$. יהי H על מישור המתאים מפריד הניצב ל-w
$H=\{z: w^tz=0\}$ נגדיר
$$d(x,H)=\min_{z\in H}\|z-x\|$$
צ"ל: $d(x,H)=\dfrac{|w^tx|}{\|w\|}$

נשתמש במכפלי לגרנג':

$$\min_z \|z-x\|^2 \quad s.t \quad w^t z = 0$$

נכתוב את הלגרנז'יאן של הבעיה:

$$\mathcal{L}(z,\lambda) = (z-x)^t(z-x) + \lambda w^t z$$

$$\frac{\partial \mathcal{L}}{\partial z} = 2z - 2x + \lambda w = 0$$

$$\Rightarrow z = x - \frac{\lambda}{2} w$$

$$\Rightarrow w^t\left(x - \frac{\lambda}{2} w\right) = w^t x - \frac{\lambda}{2} w^t w = 0$$

$$\Rightarrow \lambda = \frac{2w^t x}{\|w\|^2} \quad \Rightarrow z^* = x - \frac{w^t x}{\|w\|^2} w$$

$$\Rightarrow \|z^* - x\| = \left\|x - \frac{w^t x}{\|w\|^2} w - x\right\| =$$

$$= \frac{|w^t x|}{\|w\|^2} \cdot \|w\| = \frac{|w^t x|}{|w|}$$