למידה לא מפוקחת - יורם לוזון, תשפ"ה

רגב יחזקאל אימרה

February 3, 2025

תוכן העניינים

חלק I

מנהלה.

מרצה: יורם לוזון.

מייל מרצה: $louzouy@math.biu.ac.il$.

אין מבחן בקורס, יש עבודה שמהותה היא שאי אפשר לייצר אותה דרך $ChatGPT$. העבודה בזוגות, תהיה כתובה ב-$LaTeX$ ובאנגלית.

חלק II

הקדמה.

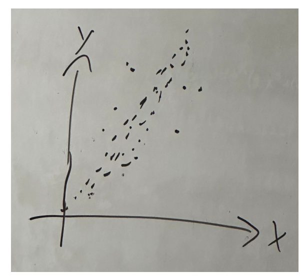
בעיה: יש לנו אוסף תצפיות $\bar{x_i}$ ולכל תצפית יש לנו תווית y_i וכל זוג כאלו מגיעים ממודל $\mathbb{P}\left(\{x_i, y_i\}|\theta\right)$ עם פרמטרים לא ידועים- θ.

<u>בעיית סיווג לא מפוקחת</u> זה בעיה שלא אכפת לי מ-y_i וכל מה שאני רוצה להעריך זה $\mathbb{P}\left(x_i|\hat{\theta}\right)$.

בהסקה סטטיסטית המדד שלנו היה כמה $\hat{\theta}$ קרוב ל-θ.

כאן נניח שאני לא יודע את $\hat{\theta}$ אבל נניח שאני יודע להעריך את $\mathbb{P}\left(x_i|\hat{\theta}\right)$. ככל שאני מעריך את $\mathbb{P}\left(x_i|\hat{\theta}\right)$ יותר נכון ככה אני יותר שמח.

דוגמה 0.1. יש לי נקודות



וכעת יש לי x חדש ואני לא יודע מה הוא y, ואני רוצה לדעת מה ה-y שמתאים ל-x שלי.

אם אני רוצה למצוא קשר בין x ל-y בצורה של $y=f(x)$ זה יהיה בעייתי כי זה יהיה מאוד מסובך למצוא כזה. לכן אני אתקן ואומר כי $y=f(x)+\varepsilon$ כאשר ε הוא רעש שאני יכול להעריך את ההתפלגות שלו אבל לא אותו.

לכן נכתוב $y=f(x)+N(0,\sigma)$ כאשר אני לא יודע גם את σ. כעת, עוד בעיה היא שיש הרבה הרבה פונקציות בעולם ואני לא יכול לעבור על כולן ולראות מי הכי טובה לי, לכן אני רוצה לבחור את $f(x)$ להיות ממשפחת פונקציות מסויימת.

הנחה 1: נניח יש לי בסיס $\Phi_i(x)$ של פונקציות $i\in\{1,...,k\}$ וכן $f(x)=\sum\limits_{i=1}^{k} w_i\Phi_i(x)$. איך אני יודע מהם ה-$w_i$ שאני צריך?

הנחה 2: ערך ה-w שאני רוצה הוא ה-w שעבורו ההסתברות של הנתונים $\{(x_1,y_1),(x_2,y_2),...,(x_k,y_k)\}$ הינה מירבית, כלומר

$$L_{\{(x_i,y_i)\}}(w)=\mathbb{P}(\{(x_i,y_i)\}|w)$$

מקבל מקסימום.

הנחה 3: הדגימות שלי בלתי תלויות ומאותה התפלגות (i.i.d.).

לפיכך

$$L=\mathbb{P}\left(\{(x,y)\}|w\right)=\prod_{i=1}^{k}\mathbb{P}\left(\{(x_i,y_i)\}|w\right)$$

נפתור בעזרת נראות מירבית:

$$\begin{aligned}\log(L)&=\sum_{i=1}^{k}\log\left(\mathbb{P}\left(x_i,y_i|w\right)\right)=\sum_{i=1}^{k}\log\left(\mathbb{P}(y_i|x_i,w)\right)=\sum_{i=1}^{k}\log\left(\mathbb{P}(y_i|x_i,w)\cdot\mathbb{P}(x_i|w)\right)\\&=\sum_{i=1}^{k}\log\left(\mathbb{P}(y_i|x_i,w)\right)+\underbrace{\log\left(\mathbb{P}(x_i)\right)}_{\substack{\text{לא מעניין אותנו}\\ \text{כי גוזרים לפי } w}}=\sum_{i=1}^{k}\underbrace{\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)}_{\substack{\text{לא מעניין אותנו}\\ \text{כי גוזרים לפי } w}}-\frac{(y_i-f(x_i))^2}{2\sigma^2}\end{aligned}$$

לכן ללא קבועים סה"כ יש לי

$$\log(L) = -\sum_{i=1}^{k} (y_i - f(x_i))^2$$

אבל במקום למקסם את $\log(L)$ אני יכול למזער את $-\log(L)$ לכן נגדיר את ה-LOSS (הפסד) להיות $\text{Loss} = \sum_{i=1}^{k} (y_i - f(x_i))^2$ ואני רוצה למזער את ההפסד שלי (לגיטימי סה"כ).

הבעיה היא שלפעמים f היא מאוד קשה לחישוב או שאני לא יודע בכלל את w_i.

בעצם, אני לא רוצה למקסם את $\mathbb{P}(\{(x_i, y_i)\}|w)$, אלא את $\mathbb{P}(w|\{(x_i, y_i)\})$. למזלי הם קשורים אחד לשני לפי חוק בייס:

$$\mathbb{P}(w|\{(x_i, y_i)\}) = \frac{\mathbb{P}(\{(x_i, y_i)\}|w) \cdot \mathbb{P}_0(w)}{\mathbb{P}(\{(x_i, y_i)\})}$$

אבל אני רוצה למקסם את $\log\left(\mathbb{P}(w|\{(x_i, y_i)\})\right)$ ולכן אני ממקסם את

$$\log\left(\mathbb{P}(w|\{(x_i, y_i)\})\right) = \log\left(\mathbb{P}(\{(x_i, y_i)\}|w)\right) + \log\left(\mathbb{P}_0(w)\right) - \log\left(\mathbb{P}(\{(x_i, y_i)\})\right)$$

אבל נזכור שכשאני ממקסם את $\log\left(\mathbb{P}(w|\{(x_i, y_i)\})\right)$ אני גוזר לפי w ולכן אפשר להתעלם מ-$\log\left(\mathbb{P}(\{(x_i, y_i)\})\right)$ (כי זה קבוע ב-w ולכן כשאני אגזור את זה אני אקבל 0).

אפשר להגיד כי $\mathbb{P}_0(w) = \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{\|w\|^2}{2\beta^2}}$ כי אני לא יודע את w, כי הכי נורמלי לנחש שהוא מתפלג נורמלי ☺.

ניקח לזה log ואחרי ניקוי קבועים נקבל

$$\log\left(\mathbb{P}_0(w)\right) = -\frac{\|w\|^2}{2\beta^2}$$

ניזכר גם כי $\log\left(\mathbb{P}(\{(x_i, y_i)\}|w)\right)$ הוא הנראות המירבית שלי ממקודם.

בסופו של יום אני רוצה לעשות

$$\max\left(-\frac{\|w\|^2}{2\beta^2} - \sum_{i=1}^{k} (y_i - f(x_i))^2\right)$$

או באופן שקול

$$\min\left(\frac{\|w\|^2}{2\beta^2} + \sum_{i=1}^{k} (y_i - f(x_i))^2\right)$$

ונקרא ל-$\frac{\|w\|^2}{2\beta^2}$ איבר רגולריזציה. הוא ביטוי מתמטי של ה-$\mathbb{P}_0$.

סיכום מושגים:

מודל: $y = f(x) + N(0, \sigma)$.

רעש: נורמלי.

נראות ($LOSS$): מסיקים מהמודל ומהרעש.

$\mathbb{P}_0$: מסיקים מהנראות.

רגולריזציה: מסיקים מ-$\mathbb{P}_0$.

דוגמה 0.2. אם אני בכיתה ומודד את הגבהים של כל התלמידים ונכנס לי סטודנט חדש ואני רוצה לנחש את הגובה שלו, אני אנחש לפי הנראות המירבית, זה קלי קלות. אבל אם אני רוצה להפריד בין בנים ובנות הפעם אני יודע שההתפלגות שלי צריכה להיות מורכבת מ2 משתנים מקריים נורמלים שהיא $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)$. למודל כזה קוראים "מודל תערובת גאוסיאנים", או פשוט $G.M.M$. אם יש לי דגימה חדשה x אזי

$$\mathbb{P}(x) = \pi_1 N(x|\mu_1, \sigma_1) + \pi_2 N(x|\mu_2, \sigma_2)$$

כאשר π_i זה ההסתברות להיות באשכול ה-i.

$$\log\left(\mathbb{P}(x)\right) = \log\left(\pi_1 \cdot \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \pi_2 \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}\right)$$

איזה באסה, יש לי $+$ בתוך ה-log ובגזירה זה יוצא מזעזע. לגזור ולהשוות ל-0 זה מאוד לא פתיר, לכן חייבת להיות דרך טובה יותר לחשב את זה. הבעיה היא שהכנסנו משתנים סמויים וזה מה שהכניס לנו את הסימן $+$ ל-log.

$$\frac{\partial}{\partial \pi_1} \log\left(\mathbb{P}(x)\right) = \sum_{x \in \text{דגימות}} \frac{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{\pi_1 \cdot \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \pi_2 \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}} = 0$$

בשביל זה הומצא אלגוריתם טוב שעוזר לנו מאוד.

חלק III
אלגוריתמי אשכול ב- $\mathbb{R}^n$.

1 אלגוריתם Expectation maximization.

הרעיון: אם רק הייתי יודע את $\mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2$, הייתי יכול בקלות להעריך את הסיכוי שכל דגימה שייכת לכל אשכול, ואז הייתי יכול לאמוד מחדש את $\mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2$ וחוזר חלילה.

רקע: יש לי אוסף של j מטבעות, לכל מטבע יש סיכוי q_j להיות 1. בכל פעם אני לוקח מטבע מסויים ומטיל אותו n פעמים. לכן הסיכוי לתצפית k_i- מספר הפעמים שקיבלתי עץ בניסוי ה-i היא

$$\mathbb{P}(k_i|j) = \binom{n}{k_i} q_j^{k_i} (1-q_j)^{n-k_i}$$

לכן

$$\mathbb{P}(k_i) = \sum_j \pi_j \mathbb{P}(k_i|j)$$

אני יודע כך הזמן למה k שווה, בלי לדעת מהו j. זה ניסוי אחד מתוך הרבה הרבה ניסויים שעשיתי. לכן הנראות שלי תהיה

$$L = \prod_i \left(\sum_j \pi_j \mathbb{P}(k_i|j) \right)$$

אני עכשיו עשיתי ניסוי ואני רוצה לדעת מהם π_j וכן q_j. נגדיר

$$\log(L) = \log\left(\prod_i \left(\sum_j \pi_j \mathbb{P}(k_i|j) \right) \right) = \sum_i \log\left(\sum_j \pi_j \mathbb{P}(k_i|j) \right)$$

וכמו בשיטת נראות מקסימלית אדרוש

$$\frac{\partial}{\partial q_i} \log(L), \frac{\partial}{\partial q_j} \log(L) = 0$$

הבעיה היא שהנגזרת יוצאת מאוד מסובכת

$$\frac{\partial}{\partial \pi_j} \log(L) = \sum_i \frac{1}{\sum\limits_{j'} \pi_{j'} \mathbb{P}(k_i|j')} \cdot \pi_j \frac{\partial}{\partial q_j} \mathbb{P}(k_i|j) = 0$$

וכן

$$\begin{aligned} \frac{\partial}{\partial q_j} \mathbb{P}(k_i|j) &= \frac{k_i}{q_j} \mathbb{P}(k_i|j) - \frac{n-k_i}{(1-q_j)} \mathbb{P}(k_i|j) \\ &= \underbrace{\left(\frac{k_i}{q_j} - \frac{n-k_i}{(1-q_j)} \right)}_{:=f(k_i,q_j)} \mathbb{P}(k_i|j) \end{aligned}$$

לכן

$$\frac{\partial}{\partial q_i} \log(L) = \sum_i \frac{\pi_j \mathbb{P}(k_i|j) f(k_i,q_j)}{\sum\limits_{j'} \pi_{j'} \mathbb{P}(k_i|j')} = 0$$

וזו רק המשוואה הראשונה.

מהמשוואה השנייה אני מקבל גם אילוץ: $\sum\limits_j \pi_j = 1$ כלומר נפתור

$$\frac{\partial}{\partial \pi_j} \left(\sum_i \log\left(\sum_{j'} \pi_{j'} \mathbb{P}(k_i|j') \right) + \lambda \left(\sum_j \pi_j - 1 \right) \right) = 0$$

ונקבל

$$\sum_i \frac{\mathbb{P}(k_i|j)}{\sum\limits_{j'} \pi_{j'} \mathbb{P}(k_i|j')} + \lambda = 0$$

ומכאן

$$(*)\sum_i \frac{\cancel{\sum_j \pi_j \mathbb{P}(k_i|j)}}{\cancel{\sum_{j'} \pi_{j'} \mathbb{P}(k_i|j')}} + \underbrace{\sum_j \pi_j}_{=1} \lambda = 0$$

כלומר

$$\lambda = -n$$

כעת מהמשוואה $(*)$ נקבל

$$\sum_i \underbrace{\frac{\mathbb{P}(k_i|j)}{\sum_{j'} \pi_{j'} \mathbb{P}(k_i|j')}}_{:=\gamma_{ij}} = \pi_j n$$

לכן עבור π_j קיבלנו $\sum_i \gamma_{ij} = \pi_j n$ ועבור q_i נקבל $\sum_i \gamma_{ij} f(k_i, q_j) = 0$.

מהו γ_{ij}? זה המשקל היחסי של המטבע ה-j בניסוי i.

זה אומר, שאם הייתי מגדיר אינדיקטור z_{ij}- האירוע שבחרתי מטבע j בניסוי i אזי נקבל

$$\gamma_{ij} = \mathbb{P}(z_{ij}) = \mathbb{E}\left[z_{ij}\right]$$

אם רק הייתי יודע את π_j, q_j הייתי יודע גם את γ_{ij} לכל i ולכל j. אם גם הייתי יודע את γ_{ij} הייתי יודע את π_j, q_j.

לכן אני אנחש את π_j, q_j, אחשב את γ_{ij}, מזה אני אחשב את π_j, q_j החדשים ומשם γ_{ij} וכו.

בעיה כללית: בהינתן אוסף של משתנים סמויים z_{ij} כאשר ניסוי i מתבצע עם פרמטרים θ_j ויש לכל מצב j סיכוי π_j, ורוצים לאמוד אומדן נראות מירבית של θ בהינתן סדרת ניסויים $x_1, ..., x_n$.

נכתוב:

$$\mathbb{P}(x|\theta) = \sum_z \mathbb{P}(x, z|\theta)$$

ולכן

$$\log\left(\mathbb{P}(x|\theta)\right) = \log\left(\sum_z \mathbb{P}(x, z|\theta)\right)$$

בעיה! יש לי + בתוך ה-log. אני אגדיר בשביל זה התפלגות חדשה $\mathbb{Q}(z)$ המקיימת $\mathbb{Q}(z) > 0$ וכן $\sum_z \mathbb{Q}(z) = 1$ ונקבל

$$\log\left(\mathbb{P}(x|\theta)\right) = \log\left(\sum_z \frac{\mathbb{P}(x, z|\theta)}{\mathbb{Q}(z)} q(z)\right) \underset{\star}{\geq} \sum_z \mathbb{Q}(z) \log\left(\frac{\mathbb{P}(x, z|\theta)}{\mathbb{Q}(z)}\right)$$

$\star$ לפי אי שוויון ינסן.

נחשב את

$$\begin{aligned}
\triangle &= \sum_z \mathbb{Q}(z) \log\left(\mathbb{P}(x|\theta)\right) - \sum_z \mathbb{Q}(z) \log\left(\frac{\mathbb{P}(x, z|\theta)}{\mathbb{Q}(z)}\right) \\
&= \sum_z \mathbb{Q}(z) \log\left(\frac{\mathbb{P}(x|\theta) q(z)}{\mathbb{P}(x, z|\theta)}\right) \\
&= \sum_z \mathbb{Q}(z) \log\left(\frac{\mathbb{Q}(z)}{\mathbb{P}(z|\theta, x)}\right)
\end{aligned}$$

לכן בהינתן שתי התפלגויות $\mathbb{Q}, \mathbb{P}$ נגדיר

$$D_{KL}(\mathbb{Q}, \mathbb{P}) = \sum_z \mathbb{Q}(z) \log\left(\frac{\mathbb{Q}(z)}{\mathbb{P}(z)}\right)$$

ונוכל לכתוב

$$\triangle = D_{KL}(\mathbb{Q}, \mathbb{P}(z|x, \theta))$$

ונקבל

$$\log\left(\mathbb{P}(x|\theta)\right) = \sum_z \mathbb{Q}(z) \log\left(\frac{\mathbb{P}(x, z|\theta)}{\mathbb{Q}(z)}\right) + D_{KL}(\mathbb{Q}, \mathbb{P}(z|x, \theta))$$

ואף נסמן $L := \sum\limits_z \mathbb{Q}(z)\log\left(\frac{\mathbb{P}(x,z|\theta)}{\mathbb{Q}(z)}\right)$. כעת אני רוצה להעריך את $\log\left(\mathbb{P}(x|\theta)\right)$ אבל זה מסובך לי. למזלי, להעריך את

$$\sum_z \mathbb{Q}(z)\log\left(\frac{\mathbb{P}(x,z|\theta)}{\mathbb{Q}(z)}\right) + D_{KL}(\mathbb{Q},\mathbb{P}(z|x,\theta))$$

זה קל לי.

נאפס את $D_{KL}(\mathbb{Q},\mathbb{P}(z|x,\theta))$, נרצה $\mathbb{Q}(z)=\mathbb{P}(z|x,\theta)$.

שלב א': בהינתן $\hat{\theta}_1$, נבחר $\mathbb{Q}(z)=\mathbb{P}(z|x,\hat{\theta}_1)$.

שלב ב': נכתוב $\log\left(\mathbb{P}(x|\theta)\right)=\sum\limits_z \mathbb{P}(z|x,\hat{\theta}_1)\log\left(\frac{\mathbb{P}(x,z|\theta)}{\mathbb{P}(z|x,\hat{\theta}_1)}\right)$ ואחרי ניקוי קבועים וגזירה לפי θ נקבל אומד חדש ל-θ, נסמנו ב-$\hat{\theta}_2$.

שלב ג': נחזור על שלב א' עם $\hat{\theta}_2$.

דוגמה 1.1. נבחר:

$$\mathbb{P}(x,z)=\left(\prod_i\prod_j \pi_j\binom{n}{k_i}q_j^{k_i}(1-q_j)^{n-k_i}\right)^{z_{ij}}$$

נכניס משתנים סמויים: $z_{ij}=1$ אם בניסוי ה-i בחרתי מטבע j ו-0 $z_{ij}=0$ אם בניסוי ה-i לא בחרתי במטבע j.

$$\log\left(\mathbb{P}(x,z)\right)=\sum_i\sum_j z_{ij}\log\left(\pi_j q_j^{k_i}(1-q_j)^{n-k_i}\right)$$

התעלמנו מ-$\binom{k_i}{n}$ כי הוא קבוע.

$$\mathbb{E}_z\left[\log\left(\mathbb{P}(x,z)\right)\right]=\sum_i\sum_j \gamma_{ij}\log\left(\pi_j q_j^{k_i}(1-q_j)^{n-k_i}\right)$$

נגזור לפי q_j:

$$\frac{\partial}{\partial q_j}\mathbb{E}_z\left[\log\left(\mathbb{P}(x,z)\right)\right]=\sum_i \gamma_{ij}\left(\frac{k_i}{q_j}-\frac{n-k_i}{1-q_j}\right)=0$$

נקבל

$$\sum_i \gamma_{ij}\frac{k_i}{q_j}-\sum_i \gamma_{ij}\frac{n-k_i}{1-q_j}=0$$

ובסופו של יום נקבל

$$q_j=\frac{\sum\limits_i \gamma_{ij}k_i}{\sum\limits_i \gamma_{ij}n}$$

וכן

$$\pi_j=\sum_i \frac{\gamma_{ij}}{n}$$

כאשר

$$\gamma_{ij}=\frac{\prod\limits_j \mathbb{P}(k_i|j)}{\sum\limits_{j'}\pi_{j'}\mathbb{P}(k_i|j')}$$

דוגמה 1.2. עבור תערובת גאוסיאנים (Gaussian Mixture Model-G.M.M.):

תזכורת: עבור חד מימד $\mathbb{P}(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ וברב מימד $\mathbb{P}(\bar{x}|\theta)=\frac{1}{\sqrt{(2\pi)^D\det(\Sigma)}}e^{-\frac{(\bar{x}-\bar{\mu})^T\Sigma^{-1}(\bar{x}-\bar{\mu})}{2}}$ (זה בהנחה שיש לי משתנה גאוסיאני אחד).

נחשב log לנראות:

$$\log\left(\mathbb{P}\left(\bar{x}_1,...,\bar{x}_n\right)\right)=\sum_{i=1}^n \frac{1}{\sqrt{(2\pi)^D\det\left(\Sigma\right)}}-\sum_{i=1}^n \frac{\left(\bar{x}_i-\bar{\mu}\right)^T\Sigma^{-1}\left(\bar{x}_i-\bar{\mu}\right)}{2}$$

נחשב את הגראדיאנט ונשווה ל-0:

$$\sum_{i=1}^n \Sigma^{-1}\left(\bar{x}_i-\bar{\mu}\right)=0$$

כלומר

$$\sum_{i=1}^n \left(\bar{x}_i-\bar{\mu}\right)=0$$

ולכן

$$\hat{\bar{\mu}} = \sum_{i=1}^{n} \frac{\bar{x}_i}{n}$$

ועם קצת אלגברה נגבל כי

$$\hat{\Sigma} = \sum_{i=1}^{n} \frac{\left(\bar{x}_i - \mu\right)\left(\bar{x}_i - \mu\right)^T}{n}$$

עכשיו החיים שלנו מסתבכים כי אנחנו רוצים לאמוד כמות מסויימת גדולה מ-1 של גאוסיאנים.

הבעיה:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} \log\left(\mathbb{P}(x_i|\theta_j)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{ij} \log\left(\pi_j c_j e^{-\frac{(\bar{x_i}-\bar{\mu}_j)^T \Sigma_j^{-1} (\bar{x_i}-\bar{\mu}_j)}{2}}\right)$$

כאשר

$$c_j = \frac{1}{\sqrt{(2\pi)^D \det\left(\Sigma_j\right)}}$$

לפי שלבי EM, נניח את $\pi_j, \bar{\mu}, \bar{\Sigma}_j$ וממנו נחשב את γ_{ij} ומזה נשפר את $\pi_j, \bar{\mu}, \Sigma_{ij}$ ונשפר את γ_{ij}....

$$\tilde{\gamma_{ij}} = \pi_j c_j e^{-\frac{(\bar{x_i}-\bar{\mu}_j)^T \Sigma_j^{-1} (\bar{x_i}-\bar{\mu}_j)}{2}}$$

ננרמל את $\tilde{\gamma_{ij}}$ בשביל $\sum_j \gamma_{ij} = 1$ כלומר

$$\gamma_{ij} = \frac{\tilde{\gamma_{ij}}}{\sum_j \gamma_{ij}}$$

נגזור וכן הלאה ונקבל

$$\hat{\bar{\mu}}_j = \sum_j \frac{\gamma_{ij}\bar{x}_i}{\sum_j \gamma_{ij}}$$

וגם

$$\hat{\bar{\Sigma}}_j = \sum_{i=1}^{n} \frac{\gamma_{ij}\left(\bar{x}_i - \hat{\bar{\mu}}\right)\left(\bar{x}_i - \hat{\bar{\mu}}\right)^T}{\sum_i \gamma_{ij}}$$

וכן

$$\hat{\pi}_j = \sum_{i=1}^{n} \frac{\gamma_{ij}}{n}$$

2 אלגוריתם K-Means.

אשכול קשיח:

באשכול רך ראינו ב-$G.M.M$, רק נותנים סיכוי לכל אשכול.

באשכול קשיח- נחליף את השיוך להיות חד ערכי.

הנחות:

א) נניח אשכול קשיח.

ב) נניח שהשונות בכל גאוסיאן שווה בכל מימד, ושווה בין הגאוסיאנים ואין שונות משותפת.

בהינתן ההנחות לעיל:

א) $\gamma_{ij} = \begin{cases} 0 \\ 1 \end{cases}$.

ב) $z_{ij} = 1$ עבר המרכז הקרוב ביותר ל-x_i ו-0 אחרת.

לכן $\bar{\mu}$ הוא ממוצע של נקודות באשכול.

אלגוריתם 2.1. אלגוריתם K-Means:

1. בחר k מרכזים באקראי.
2. שייך כל דגימה למרכז הקרוב אליה ביותר, וחשב מרכז מחדש.
3. אם חל שינוי, וכל האשכולות לא ריקים , חזור ל-(2).
4. אם התרוקן אשכול, פצל אשכול ל-2 אשכולים באקראי, וחזור ל-(2).

3 גרסה רכה של K-Means: אלגוריתם Fuzzy C Means.

אותן הנחות כמו K-Means אבל רך.

נגדיר את הטעות שלי להיות

$$LOSS = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\bar{x}_i - \bar{\mu}_j\|^2 \gamma_{ij}^m$$

כאשר

$$\sum_{j=1}^{n} \gamma_{ij} = 1, \gamma_{ij} \geq 0$$

אם $m \to 0$ אזי האשכול קשיח וכאשר $m \to \infty$ האשכול נהיה יותר ויותר קשיח.

נפתור באמצעות כופלי לגראנז':

$$L = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\bar{x}_i - \bar{\mu}_j\|^2 \gamma_{ij}^m - \sum_{i=1}^{n} \lambda_i \left(\sum_{i=1}^{n} \gamma_{ij} - 1\right)$$

נגזור ונקבל

$$\nabla \mu_j = \sum_{i=1}^{n} \|\bar{x}_i - \bar{\mu}_j\| \gamma_{ij}^m = 0 \Rightarrow \bar{\mu}_j = \frac{\sum_{i=1}^{n} \bar{x}_i \gamma_{ij}^m}{\sum_{i=1}^{n} \gamma_{ij}^m}$$

$$\frac{\partial L}{\partial \gamma_{ij}} = \frac{1}{2}\|\bar{x}_i - \bar{\mu}_j\|^2 m \gamma_{ij}^{m-1} - \lambda_i = 0 \Rightarrow \gamma_{ij} = \left(\frac{\lambda_i}{\frac{1}{2}\|\bar{x}_i - \bar{\mu}_j\|^2 m}\right)^{\frac{1}{m-1}}$$

$$\frac{\partial L}{\partial \lambda_i} = \sum_{j=1}^{n} \frac{\lambda_i^{\frac{1}{m-1}}}{\left(\frac{1}{2}\|\bar{x}_i - \bar{\mu}_j\|^2 m\right)^{\frac{1}{m-1}}} - 1 = 0 \Rightarrow \lambda_i = \left(\frac{1}{\sum_{j=1}^{n} \frac{1}{\left(\frac{1}{2}\|\bar{x}_i - \bar{\mu}_j\|^2 m\right)^{\frac{1}{m-1}}}}\right)^{m-1}$$

חלק IV

מטריקות.

כעת מה קורה אם יש לי רק מרחק בין עצמים d_{ij} ללא יכולת להגדיר הטלה ל-$\mathbb{R}^n$ של העצמים? יש לי 2 פתרונות לבעיה הזאת:

1. למצוא לכל i $\bar{x}_i$ ולהגדיר מטריקה $d(\bar{x}_i, \bar{x}_j) \approx d_{ij}$.
2. להגדיר גרף: התצפית ה-i תהפוך לקודקוד ה-i וכן $f(d_{ij})$ יהיה משקל הקשת דמיון ביניהם עבור f פונקציה מונוטונית יורדת. אם משתמשים בדמיון אז ניתן לזרוק קשתות עם ערך נמוך.

מטרה: אני רוצה לפרק את הגרף V ל2 אשכולות (או יותר) A, B כך ש-$A \cup B = V$ וכן $A \cap B = \emptyset$.

האמונה שלי אומרת לי שאני רוצה $|A| \sim |B|$ וכן $\min(\sum_{\substack{ij \\ i \in A \\ j \in B}} w_{ij})$ כלומר סכום משקלי הקשתות בין A ל-B יהיה מינימלי.

סימון: $z_i = \begin{cases} \frac{1}{2} & i \in A \\ -\frac{1}{2} & i \in B \end{cases}$ לכן לכל i, j באשכולות שונים $(z_i - z_j)^2 = 1$ וכן $(z_i - z_j)^2 = 0$ לכל i, j באותו אשכול. בנוסף לכל i מתקיים כי $z_i^2 = \frac{1}{4}$, לכן אני יכול עכשיו להגדיר פונקציית הפסד

$$LOSS = \sum_{i,j} w_{ij} (z_i - z_j)^2$$

טעות! אם אני אגדיר ככה את ההפסד אז אין בעיה לומר שפשוט כל הגרף שלי חלק מ-A וכן ש-B קבוצה ריקה וכביכול סיימתי אבל בעצם לא עשיתי פה כלום. לכן אני אגדיר את ההפסד שלי להיות

$$LOSS = \sum_{i,j} w_{ij} (z_i - z_j)^2 - \sum_{i} \underbrace{\left(\sum_{j} w_{ij}\right)}_{dij} (z_i - z_i)^2$$

כאשר יש לי אילוץ $\left|\sum_i z_i\right| < \varepsilon$.

דוגמה 3.1. יש לי מטריצת דרגות

$$W=\begin{pmatrix}0&2&0&5&0&3\\2&0&1&1&0&0\\0&1&0&2&1&0\\5&1&2&0&0&1\\0&0&1&0&0&4\\3&0&0&1&4&0\end{pmatrix}$$

כלומר מחוץ לאלכסון $d_{ij}=0$ וכן $d_{ii}=\sum\limits_j w_{ij}$.

מטריצת הדרגות שלי תהיה

$$D=\begin{pmatrix}10&&&&&\\&4&&&&\\&&4&&&\\&&&9&&\\&&&&5&\\&&&&&8\end{pmatrix}$$

כעת,

$$LOSS=\sum_{i,j}w_{ij}(z_i-z_j)^2-\sum_{i,j}d_{ij}(z_i-z_j)^2$$

לכן אני מקבל השראה להגדיר את <u>מטריצת הלפלסיאן</u> L_a להיות $L_a=W-D$.

במקרה שלנו נקבל

$$L_a=\begin{pmatrix}-10&2&0&5&0&3\\2&-4&1&1&0&0\\0&1&-4&2&1&0\\5&1&2&-9&0&1\\0&0&1&0&-5&4\\3&0&0&1&4&-8\end{pmatrix}$$

וכן

$$\begin{aligned}LOSS&=\sum_{i,j}L_{a_{ij}}(z_i-z_j)^2\\&=-2\bar{z}^TL_a\bar{z}\end{aligned}$$

כלומר אני רוצה לפתור

$$\begin{aligned}&\max z^TL_az\\&\text{s.t.} z_i\in\{-0.5,0.5\}\\&\left|\sum_i z_i\right|<\varepsilon\end{aligned}$$

כיוון ש-z_i בדיד זה ממש קשה לפתור, אז נרשה ל-z_i להיות רציף אבל עדיין נבקש $\sum\limits_i z_i^2=\frac{n}{4}$ כאשר כמובן n זה מספר הדגימות שלי.

נגדיר $\bar{v}=z\cdot\frac{2}{\sqrt{n}}$ ולכן $\|v\|^2=1$ כלומר v וקטור יחידה ולכן $\sum\limits_i v_i^2\le\underbrace{\frac{2}{\sqrt{n}}\varepsilon}_{:=\beta}$.

תכונות של מטריצת הלפלסיאן:

1. ע"ע ממשיים- כי היא סימטרית.
2. ו"ע $(1,...,1)^T$ כי סכום כל שורה הוא 0.

נפרק את v לפירוק לפי בסיס של וקטורים עצמיים של הלפלסיאן:

$$v=\sum_i\alpha_i\bar{u}_i$$

כאשר $\bar{u}_i$ ו"ע של L_a. לכן

$$L_a\cdot v=\sum_i\alpha_i\lambda_i\bar{u}_i$$

וכן

$$v^T L_a v = \sum_i \alpha_i^2 \lambda_i$$

כעת, הע"ע של L_a הם כולם אי חיוביים, כלומר אני ממקסם את זה

$$\max \sum_i \alpha_i^2 \lambda_i$$
$$\text{s.t.} \sum_i \alpha_i^2 = 1$$
$$\lambda_j \leq \lambda_i = 0$$

לכן אני ארצה $\alpha_1 = 1, \alpha_{i \neq 1} = 0$ אבל להזכירנו יש לי אילוץ $\sum_i v_i^2 \leq \beta < 1$, לכן הפתרון שלי יהיה $\alpha_1 = \beta, \alpha_2 = \sqrt{1-\beta^2}$ ואם חוזרים כל הדרך אחורה נקבל

$$z = \frac{\sqrt{n}}{2}\left(\beta \bar{u}_1 + \sqrt{1-\beta^2}\bar{u}_2\right)$$

לכן מהדוגמה הזאת בא לי הרעיון לאלגוריתם אשכול ספקטרלי:

4 אלגוריתם אשכול ספקטרלי:

אלגוריתם 4.1. צעדים לאשכול ספקטרלי:

1. חשב דרגה $d_{ij} = \sum_j w_{ij}$.
2. חשב $L_a = W - D$ כאשר $D = \begin{cases} d_{ij} & i = j \\ 0 & i \neq j \end{cases}$.
3. חשב ו"ע שני של L_a: $\bar{u}_2$.

$$\bar{z} = \frac{\sqrt{n}}{2}\left(\beta \bar{u}_1 + \sqrt{1-\beta^2}\bar{u}_2\right)$$

4. אם $z_i > 0$ שייך ל-A אחרת שייך ל-B.

5 אלגוריתם Prim.

נניח יש לי המון המון נקודות וגרף ביניהם שבו קשת מסמנת קרבה בין 2 קודקודים.

האלגוריתם מוצא לי עפ"מ.

אלגוריתם 5.1. אלגוריתם Prim.

1. בכל שלב חבר לעץ הקיים את הקשת הכי זולה מחוץ לעץ.
2. חזור ל-1.

התוצאה של אלגוריתם Prim הוא עץ, לכן אם אני אסנן את ה-k קשתות הכי יקרות שלי אני אקבל k אשכולות כמו שרציתי באשכול ספקטרלי.

6 אשכול היררכי.

אלגוריתם 6.1. אלגוריתם לאשכול היררכי:

1. שים כל דגימה באשכול משלה.
2. חבר שני אשכולות קרובים ביותר לאשכול חדש וחשב את המרחק שלו מאשר האשכולות.
3. כל עוד לא כולם מחוברים חזור ל-2.

שאלה מעולה: איך אני מגדיר מרחק בין אשכולות?

תשובה מעולה: זה תלוי.

כל אחד יכול להגדיר בעצמו ולראות מה הכי טוב לו.

דוגמה 6.2. דוגמאות למרחקים בין אשכולות: $d(A,(B,C))$

1. מרחק single link:

$$\min\left(d(A,B), d(A,C)\right)$$

2. מרחק full link:

$$\max\left(d(A,B), d(A,C)\right)$$

3. מרחק average link:
$$\frac{d(A,B)+d(A,C)}{2}$$
4. מרחק UPGMA:
$$\frac{|B|}{|B|+|C|}d(A,B)+\frac{|C|}{|B|+|C|}d(A,C)$$
5. מרחק energy distamce:
$$d(\{x_i\},\{y_j\})=2\sum_{i,j}d(x_i,y_j)-\sum_{i}d(x_i,x_i')-\sum_{j}d(y_j,y_j')$$

חלק V

אלגוריתמי Community detection.

השאלה שלנו: בהינתן רשת (גרף), איך מחלקים אותה לקהילות (אשכולות)?

תזכורת: יש לנו A_{ij}- דמיון בין i ל-j. $A_{ij}=0$ אם אין קשת בין i ל-j. בנוסף, $d_i=\sum_j A_{ij}$ הוא דרגת קודקוד.

אם הגרף כיווני דרגת כניסה היא $\sum_j A_{ji}$ ודרגת יציאה היא $\sum_j A_{ij}$.

7 אלגוריתם Newman Girvan.

הגדרה 7.1. המרכזיות של קשת/ קודקוד x הוא מספר המסלולים המינימלים שעוברים דרך הקשת/ קודקוד x.

נסמן על ידי Centrality(x).

אלגוריתם 7.2. אלגוריתם Newman Girvan:

1. חשב מרכזיות לכל קשת.
2. הוצא קשתות עם המרכזיות הכי גבוהה לפי סדר המרכזיות.
3. חזור על 2 עד שהגעת למספר הקשתות הרצוי.

8 אלגוריתם Louvain.

נגדיר את הרווח להיות
$$Q=\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_ik_j}{2m}\right)C(i,j)$$
כאשר:

k_i דרגה של קודקוד ה-i.

$\frac{1}{2}\sum_{i,j}A_{ij}=m=\frac{1}{2}\sum_i k_i$=מספר הקשתות.

$C(i,j)=\begin{cases}1 & \text{באותו אשכול } i,j\\ 0 & \text{אחרת}\end{cases}$.

שאלה חשובה היא כמה קשתות יש בקבוצה מול כמה ציפיתי.

אם $A_{ij}-\frac{k_ik_j}{2m}$ גבוה, רוצים $C(i,j)=1$, אחרת רוצים $C(i,j)=0$. זו בעיה כי לכל קודקוד אני צריך לקבוע באיזה אשכול הוא ויש מלא אשכולים וזה מאוד יקר. לכן הפתרון הוא להפוך את $C(i,j)$ לרציף תוך כדי שנייצר אילוצים ונוכל לגזור אותו.

אלגוריתם 8.1. אלגוריתם Louvain.

1. נסדר קודקודים בסדר אקראי.
2. נחבר כל קודקוד לאשכול הקרוב אליו, אם החיבור מעלה את Q (פורמלית, בהינתן 2 קודקודים שכנים i,j חשב את
$$\Delta(i,j)=Q(j \text{ עובר לאשכול } i)-Q(i,j \text{ במצב הנוכחי})$$
ואם $\Delta(i,j)>0$ תעביר את i לאשכול j.

3. אם אין איך להתקדם, חבר כל אשכול לקודקוד אחד וצור קשת פנימית עם פעמיים סכום הקשתות בכל אשכול $w_{k,k}=\sum_{\substack{i\in C_k\\ j\in C_k}} A_{ij}$ וקשת בין אשכולות כתור סכום הקשתות שביניהם $w_{k,l}=\sum_{\substack{i\in C_k\\ j\in C_l}} A_{ij}$.
4. חזור עד שאין שינוי.

9 אלגוריתם Leiden.

נגדיר את הרווח להיות $Q=\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_ik_j}{2m}\right)C(i,j)$.

הבדלים בין לובין ללידן:

1. מאפשר לפצל אשכולות.
2. מוסיף רעש.

סיכום:

אם יש לי גרף סביר שאני רוצה לחלק ל2 קבוצות- אשכול ספקטרלי.

גרף ענק- Leiden\ Louvain.

אם חשוב לי היררכיה של אשכולות- אשכול היררכי.

10 הערכת צפיפות- היסטוגרמה, parzen windows ו-KDE.

נניח יש לי סדרה של מספרים $1,2,3,1,4,5,1,8,2,7,1,4,3,2,7,5,8,10,9,7,4,3,8,10$.

דרך אחת להעריך כמה מופעים יש לי מכל מספר היא באמצעות ספירה של ממש (ונירמול):

מספר	10	9	8	7	6	5	4	3	2	1
צפיפות	$\frac{2}{24}$	$\frac{1}{24}$	$\frac{3}{24}$	$\frac{4}{24}$	$\frac{0}{24}$	$\frac{2}{24}$	$\frac{3}{24}$	$\frac{3}{24}$	$\frac{3}{24}$	$\frac{4}{24}$

לפעולה זו קוראים היסטוגרמה.

אני אקח אינטרוולים מרווחים שווה בשווה (כמה שבא לי) בין המקסימום ועוד קצת לבין המינימום פחות קצת ובין כל מרווח (באנגלית bin) אני אשים את כל המספרים שמתאימים לשם. כלומר במקרה הזה $\rho_i=\frac{n_i}{\Delta x}$. הבעיה היא שאם יש לי עוד מספר 545435453 אז יהיה לי היסטוגרמה דפוקה. לכן נגדיר $\rho=\frac{n_0}{\Delta x_i}$ כלומר כל מרווח (bin) להכיל n_0 ערכים. באופן טיפוסי $n_0=\sqrt{n}$.

כעת אם אני בשני מימדים אני יכול לעשות מקביל לזה ויצא לי $\rho_i=\frac{n_i}{\Delta x\Delta y}$.

אם יש לי k בינים בכל מימד ב-l-מימדים נקבל שיש לי k^l בינים וזה מאוד מאוד יקר. נניח שיש לי ∞ זיכרון אז ברוב התאים יהיו לי 0 איברים ובחלק מהתאים יהיה לי קצת מאוד מאוד איברים, כלומר ההיסטוגרמה שלי תהיה מאוד שטוחה.

פתרון אלטרנטיבי: במקום לעשות קוביות נעשה כדור $B(x,\varepsilon)=\{z:\|x-z\|\leq\varepsilon\}$. אני אסמן $V=$ נפח כדור ב-l-מימדים עם רדיוס ε ו-$n(x)=$ כמות הנקודות ברדיוס סביב x. לכן נגדיר $\rho(x)=\frac{n(x)}{V}$. הבעיה היא שיכול להיות שאני אדגום כדור ריק ואפספס דגימות. אני רוצה איכשהו להציל את התכונה של צפיפות בלי להיות תלוי במזל שלי באיפה אני דוגם.

פתרון: אם יש לי ε רדיוס בתוכו יש k נקודות סביב x אזי $\rho(x)=\frac{k}{V_\varepsilon}$ כאשר V_ε נפח כדור יחידה עם רדיוס ε. באופן טיפוסי נבחר $k=\sqrt{n}$.

נשאלת השאלה איך אפשר להפוך את זה לרציף? נסמן $\rho(\bar{x})=\sum_i\frac{\mathbb{I}(\|\bar{y}_i-\bar{x}\|\leq\varepsilon)}{V_\varepsilon}$. במקום אינדיקטור ניקח פונקציה יותר יפה שגם יהיה לי קל לגזור f ונקבל

$$\rho(\bar{x})=\sum_i\frac{f\left(\frac{\|\bar{y}_i-\bar{x}\|}{\varepsilon}\right)}{V_\varepsilon}$$

נבחר את $f(x)=\frac{1}{\left(\sqrt{2\pi\varepsilon}\right)^l}e^{-\frac{\|\bar{y}_i-\bar{x}\|^2}{2\varepsilon^2}}$. זה יוצא נורא יפה, לכן נגדיר $\text{KDE}(x)=\sum_i\frac{\mathbb{P}(y_i|x)}{V}$ סיכוי ל-y_i תחת הנחת להיות צפיפות נורמלית סביב x ברדיוס ε. לכאורה העלות של זה היא $O(|V|)$ אבל אפשרי גם ב-$O(\log|V|)$.

איך מזה אני בונה מזה אשכולות?

א) נקודות בצפיפות גבוהה שייכות לאשכול.

ב) נקודות בצפיפות נמוכה הן רעש

מבחינתנו אשכול חייב להיות מספיק צפוף בצורה רציפה.

11 אלגוריתם DBSCAN.

אלגוריתם 11.1. אלגוריתם DBSCAN.

א) בהינתן ε נגדיר צפיפות של נקודה כתור מסתר הנקודות האחרות שהו ברדיוס ε ממנה.

ב) נגדיר $\rho(x)=\frac{k}{V_\varepsilon}$ כאשר k הוא מספר הנקודות במרחק ε מ-x וכן V_ε נפח כדור l מימדי ברדיוס ε.

ג) נקודת גרעין היא נקודה עם צפיפות $\rho(x)\geq\rho_0$ וכן $k(x)\geq k_0$. באופן טיפוסי נבחר $k_0=5,\varepsilon=0.1$.

ד) נגדיר אשכול כתור אוסף של נקודות גרעין שמחוברות אחת לשניה ברדיוס ε וכל הנקודות שסביבן ברדיוס ε.

ה) כל הנקודות שלא חלק מהאשכול הן רעש.

חלק VI

הורדת מימדים.

נניח אני מסתכל על רשימת ציונים $\bar{x}$ מ-48 קורסים. אין 48 תכונות שקובעות את הציון, יש בערך 5 תכונות-$\bar{y}$. נאמר $\bar{x}\sim f(\bar{y})$ אבל

$$y\in\mathbb{R}^L,x\in\mathbb{R}^K,K>>L$$

אני לא מכיר את f ולא את y, אבל אני רוצה מודל שבהינתן $\bar{x}$ יביא לי $\bar{y}$. מה יכול לאפיין לי את y?

יש לנו מלא נתונים $x_1,...,x_n$ שמגיעים האיזשהו $x_i=f(y_i)$. המטרה שלנו :

א) מהו f?

ב) מהו y_i?

אנחנו רוצים להוריד את המימד של הנתונים שלנו תוך כדי שאנחנו שומרים על כמה שיותר אינפורמציה. איך נעשה את זה?

א) רוצים למזער טעות : $\tilde{x}_i=f(y_i)$ שיביא לנו $\|\tilde{x}_i-x_i\|^2$ מינימלית PCA.

ב) רוצים לשמר אינפורמציה ICA.

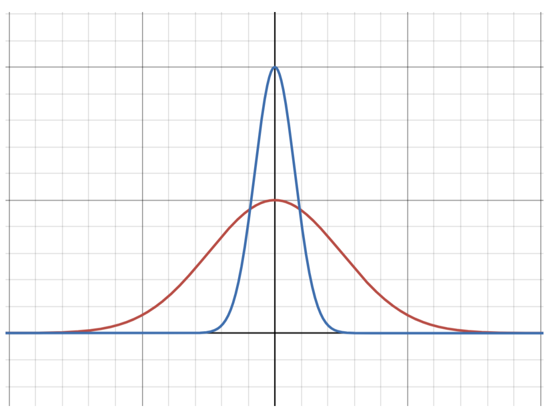
ג) מרחק $d(y_i,y_j)\sim d(x_i,x_j)$ M.D.S\ISOMAP.

ד) צפיפות : מסלול של הולך אקראי TSNE\UMAP.

12 חזרה על תורת האינפורמציה.

12.1 מדד לאי וודאות.

אם יש לי התפלגות ρ אזי ככל שהסטיית תקן שלי יותר גבוהה ככה יש לי אי וודאות יותר גבוהה.



לדוגמה, יש פה 2 התפלגויות, האדומה עם סטיית תקן יותר גבוהה ולכן יותר מפוזרת.

נפרמל את הרעיון הזה :

הגדרה 12.1. בהינתן התפלגות ρ נגדיר את **האנטרופיה** של ρ להיות

$$\mathbb{H}(\rho)=-\sum_i\mathbb{P}_i\log(\mathbb{P}_i)=\int_{\mathbb{R}}\mathbb{P}(x)\log(\mathbb{P}(x))dx$$

הגדרה 12.2. בהינתן שני משתנים מקריים X,Y נגדיר את **האינפרמציה המשותפת** שלהם להיות

$$\mathbb{H}(X,Y)=-\sum_{x,y}\mathbb{P}(x,y)\log\left(\mathbb{P}(x,y)\right)$$

הגדרה 12.3. בהינתן משתנה מקרי X ומשתנה מקרה Y אני רוצה לדעת כמה הידיעה של Y חידש לי על הידיעה על X. לכן נגדיר את **האינפורמציה המותנית** להיות

$$\begin{aligned} MI(X,Y) &= \mathbb{H}(X) - \mathbb{H}(X|Y) \\ &= \mathbb{H}(Y) + \mathbb{H}(X) - \mathbb{H}(X,Y) \\ &= MI(Y,X) \end{aligned}$$

13 בעיית PCA.

נניח שיש לי וקטור $\bar{x}_i = \sum_{j=1}^{m} \alpha_{i,j}\bar{U}_j$ כאשר $\bar{U}_i$ בסיס אורתונורמלי. אזי $\alpha_{ij} = \left\langle \bar{x}_i, \bar{U}_j \right\rangle = \bar{x}_i^T\bar{U}_j$ ואני רוצה להחליט איזה $\alpha_{i,j}$ אני שומר ואיזה אני זורק. יהיה לי $\tilde{x}_i = \sum_{j=1}^{k} \alpha_{i,j}\bar{U}_j$. נגדיר

$$\begin{aligned} L &= \sum_i \|x_i - \tilde{x}_i\|^2 \\ &= \sum_i \| \sum_{j=k+1}^{m} \alpha_{i,j}\bar{U}_j\|^2 \\ &\vdots \\ &= \sum_i \sum_{j=k+1}^{m} \alpha_{i,j}^2 \\ &= \sum_i \sum_{j=k+1}^{m} \left(\bar{x}_i^T\bar{U}_j\right)^T \left(\bar{x}_i^T\bar{U}_j\right) \\ &= \sum_{j=k+1}^{m} \bar{U}_j^T \underbrace{\left(\sum_i \bar{x}_i\bar{x}_i^T\right)}_{:=C} \bar{U}_j \\ &= \sum_{j=k+1}^{m} \bar{U}_j^T C\bar{U}_j \end{aligned}$$

נזכור כי U_j וקטורי יחידה $U_j^TU_j = 1$ ולכן נגדיר את הלגרנז'יאן להיות

$$La = \sum_{j=k+1}^{m} \bar{U}_j^T C\bar{U}_j - \sum_j \left(U_j^TU_j - 1\right)\lambda_j$$

כאשר λ_j מסמנים לי פה כופלי לגראנז'. נגזור:

$$\nabla_{U_i} La = 0 \Rightarrow 2\bar{\bar{C}}\bar{U}_j - 2\lambda_j\bar{U}_j = 0$$

נקבל

$$\bar{\bar{C}}\bar{U}_j = \lambda_j\bar{U}_j$$

כלומר $\bar{U}_j$ ו"ע של $\bar{\bar{C}}$. לכן $L = \sum_{j=k+1}^{m} \bar{U}_j^T C\bar{U}_j = \sum_{j=k+1}^{m} \lambda_j$ כאשר כאן λ_j מסמן לי ע"ע מתאים ל-$\bar{U}_j$. לכן נבחר λ_j להיות הקטנים ביותר.

אלגוריתם 13.1. אלגוריתם PCA.

1. העבר את ממוצע הנקודות ל-$\bar{0}$.
2. חשב את מטריצת ה-$C = \sum_j \left(\bar{x}_i - \bar{\mu}_i\right)\left(\bar{x}_i - \bar{\mu}_i\right)^T$.
3. חשב ו"ע של C.
4. חשב הטלה של k ו"ע עם ע"ע גבוהים ביותר $\alpha_{k_1}, ..., \alpha_{k_n}$.
5. היצוג של $\bar{x}_i$ הוא $\begin{pmatrix} \alpha_{i1} \\ \vdots \\ \alpha_{ik} \end{pmatrix}$.

זה הייצוג החדש.

14 אלגוריתם ICA.

נניח אני עכשיו בים ושומע מלא מלא גלים. מאחורי אני שומע אנשים שמדברים עלי ואני מעוניין בהינתן הגלי קול שלהם להבין מה הם אמרו עלי. הבעיה היא שאין לי רק את הגלי קול שלהם אלא זה מתערבב עם הגלי קול של הים. איך אני יכול להפריד בין הגלי קול של אלו שהיו מאחורי ובין הגלי קול של הים? יש לי $A\bar{y}=\bar{x}$ כאשר $\bar{y}$ זה מה שאני מחפש. מה הבעיה? כיוון ש-$\bar{x}$ שלי יהיה משתנה מקרי לא נורמלי, כשאני מכפיל את $\bar{y}$ (וקטור משתנים מקריים) במטריצה כלשהי B אני מקבל $B\bar{y}$ וקטור נורמלים. אבל אז איך זה יתכן ש-$\bar{x}$ וקטור נורמלי? הרי אני יודע שהוא קיים אז איך אני מגיע אליו?

הרעיון שלי הוא שבהינתן $\bar{x}_i$ נורמלים נחפש A כך ש-$y_i = A^{-1}x_i$ כמה שפחות נורמלי.

מה מאפיין התפלגות נורמלית משאר ההתפלגויות? להתפלגות נורמלית יש רק מומנטים 1 ו-2 שונים מ-0, כל שאר המומנטים שווים ל-0.

אלגוריתם 14.1. אלגוריתם ICA.

נגדיר $K=\mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^4\right]$. בהתפלגות נורמלית אנחנו רוצים שהוא יהיה כמה שיותר מינימלי.

בהינתן x_i ממימד n בחר מטריצה $C_{k\times n}$. חשב $y_i=Cx_i$ ובצע צעד בכיוון שממקסם

$$\sum_{\substack{\text{מימדים} \\ y\text{של}}}^{k} K\left(\begin{array}{c}\text{כל מימד} \\ y\text{של}\end{array}\right)$$

עוד דרך:

אלגוריתם 14.2. אלגוריתם ICA בשיטה אחרת:

בהינתן $\bar{x}_i$ נרצה להטיל אותו ל-y_i במימד 1 $y_i=\bar{U}^T\bar{x}_i$ כך ש-$MI(y_i,\bar{x}_i)$. פתרון:

א) בחר $\bar{U}$ אקראי.

ב) חשב $MI(y_i,\bar{x}_i)$.

ג) בצע צעד $\Delta U=\nabla_U MI(y_i,\bar{x}_i)$.

15 אלגוריתם C.M.D.S.

בהינתן d_{ij} בין כל הזוגות המטרה היא למצוא הטלה $\bar{x}_i$ כך ש-

$$\|x_i-x_j\|=d_{ij}$$

כלומר

$$(x_i-x_j)(x_i-x_j)^T=d_{ij}^2$$

כלומר

$$x_i^Tx_i-2x_i^Tx_j+x_j^Tx_j=d_{ij}^2$$

לכן אם נסמן $b_{ij}=x_i^Tx_h$ נקבל

$$d_{ij}^2=b_{ii}-2b_{ij}+b_{jj}$$

כעת בה"כ נניח $\sum\limits_i \bar{x}_i=0$. d_{ij}^2 ידוע אבל b_{ij} לא ידוע כי x_i לא ידועים. אבל:

$$b_{ij}=\left(\sum_i x_i^T\right)x_j=0$$

לכן

$$\sum_i\left(b_{ii}-2\cancel{b_{ij}}+b_{jj}\right)=\sum_i d_{ij}^2$$

נסמן $B=\bar{x}^T\bar{x}$ ונסמן $T=trace(B)$. לכן

$$T=\sum_j\sum_i\frac{d_{ij}^2}{2n}$$

ולכן

$$\sum_i T+nb_{jj}=\sum_j\sum_i d_{ij}^2$$

כלומר

$$b_{jj}=\frac{\sum\limits_i d_{ij}^2-T}{n}$$

ואז נקבל כבר

$$b_{ij}=\frac{b_{ii}+b_{jj}-d_{ij}^2}{2}$$

אלגוריתם 15.1. אלגוריתם C.M.D.S.

1. חשב $T = \sum_j \sum_i \frac{d_{ij}^2}{2n}$.
2. חשב $b_{jj} = \frac{\sum_i d_{ij}^2 - T}{n}$.
3. חשב $b_{ij} = \frac{b_{ii}+b_{jj}-d_{ij}^2}{2}$.
4. פרק $B = U^T \Sigma U$.
5. $X = \Sigma^{\frac{1}{2}} U$.

16 אלגוריתם ISOMAP.

טוב לדאטה על מניפה.

בהינתן גרף לא כיווני של מרחקים w_{ij} המרחק בין i, j נגדיר $K.N.N.$ בתור גרף כיווני שכל קודקוד מחובר ל-k השכנים הקרובים אליו.

אלגוריתם 16.1. אלגוריתם ISOMAP.

1. חשב גרף $K.N.N.$.
2. חשב מרחקים בגרף $K.N.N$ (דייקסטרה).
3. בצע $M.D.S$.

17 אלגוריתם LLE.

אני רוצה לשמר את המבנה הלינארי הלוקאלי של כל הנקודות.

אלגוריתם 17.1. אלגוריתם LLE.

1. בנה גרף $K.N.N.$.
2. קרב כל נקודה להיות צירוף לינארי של השכנים: x_i שכנים- $x_i \sim \Sigma_j w_{ij} x_j$ על ידי מציאת w_{ij} עבורם $\|x_i - \Sigma_j w_{ij} x_j\|^2$ מינימלי. שימו לב שזה נפתר בנפרד. עבור $\tilde{x}_i$ פותרים $\frac{1}{2}\Sigma_i \|\tilde{x}_i - \Sigma_j w_{ij} \tilde{x}_j\|^2$ מינימלי- פתרון גלובאלי. w_{ij} נתון- מחפשים את $\tilde{x}_i$. כדי לא להגיע לפתרון ה-0 דורשים $\tilde{x}_i$ על מעגל היחידה.

18 אלגוריתם EIGENMAPS.

שוב עובדים על גרף $K.N.N.$. היינו רוצים ש-d_{ij}^2 יהיה נמוך $\Longleftarrow \|\tilde{x}_i - \tilde{x}_j\|$ יהיה נמוך. נגדיר דימיון $r_{ij} = f(d_{ij})$ פונקציה יורדת מ-∞ ל-0. ההפסד יהיה

$$L = \sum_{i,j} r_{ij} \|\tilde{x}_i - \tilde{x}_j\|$$

ונחפש $\min(L)$. צריך אילוץ על $\tilde{x}_i$: נגדיר

$$D_I = \begin{pmatrix} \Sigma_j w_{1j} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_j w_{nj} \end{pmatrix}$$

ולכן האילוץ יהיה $\tilde{x}^T D_I \tilde{x} = I$ כאשר $\tilde{x}_i$ צמודים ל-D_I. לכן נקבל את הבעיה

$$\begin{aligned} &\min \sum_{i,j} r_{ij} \|\tilde{x}_i - \tilde{x}_j\| \\ &\text{s.t.} \tilde{x}^T D_I \tilde{x} = I \end{aligned}$$

נקבל שהפתרון הוא $L = W - D_I$.

19 הולכים מקריים, צוואר בקבוק לאינפורמציה ואלגוריתם T-SNE.

צוואר בקבוק: נעשה את הניסוי המשעשע הבא:

אליס ובוב שניהם רשומים לקורס כלשהו באוניברסיטה. רצה הגורל ויום אחד בוב לא יכול להגיע להרצאה ואליס ישבה וסבלה שעה וחצי לבדה בהרצאה. בסוף השיעור הגיע בוב אחרי שרץ בטורבו 5 בלוקים וביקש מהמרצה שיסכם לו את מה שקרה בשיעור, ובאורך פלא ב3 משפטים המרצה סיכם לבוב את מה שהלך בשיעור. נרגש ונפעם בוב הלך לאליס וסיפר לה את ה3 משפטים שהמרצה אמר לו והמומה אליס אמרה לו שזה סיכום מאוד מדוייק של כל מה שהלך בשיעור. מה המסקנה? היה אפשר לדחוס את השעה וחצי של סבל שחוותה אליס לכדי 3 משפטים וסתם היה בזבוז של זמן.

הילוך מקרי: נניח יש לי עכשיו גרף של נקודות במרחב כלשהו. נסמן $\frac{f(\|x_i-x_j\|)}{\sum\limits_{i\neq j} f(\|x_i-x_j\|)}=$ (הסיכוי שאקפוץ מ-i ל-j) $\mathbb{P}_{j|i}=\mathbb{P}$ כאשר $\sum\limits_j \mathbb{P}(j|i)=1$ וכן ניקח את f להיות $f\sim\mathcal{N}(0,\sigma)$. לכן

$$\mathbb{P}_{j|i}=\frac{e^{-\frac{\|x_i-x_j\|^2}{2\sigma_i^2}}}{\sum\limits_{i\neq j} e^{-\frac{\|x_i-x_j\|^2}{2\sigma_i^2}}}$$

מתקיים

$$\mathbb{P}_{j,i}=\frac{\mathbb{P}_{j|i}+\mathbb{P}_{i|j}}{2n}$$

ואכן מתקיים $\sum\limits_i\sum\limits_j \mathbb{P}_{i,j}=1$ אני רוצה $\tilde{x}_i$ בו הסיכוי לקפוץ מ-$\tilde{x}_i$ ל-$\tilde{x}_j$ דומה לסיכוי לקפוץ מ-i ל-j.

ברמה הרעיונית איך אלגוריתם $T-SNE$ יפעל?

1. חשב סיכויי קפיצה ל-k שכנים.
2. הפוך לסימטרי.
3. מצא $\tilde{x}_i$ עבורם סיכויי הקפיצה דומים לאלו של x_i אבל החלף את ההתפלגות הנורמלית בהתפלגות t:

$$g(\|\tilde{x}_i-\tilde{x}_j\|)=\frac{1}{1+\|\tilde{x}_i-\tilde{x}_j\|}$$

ללא שום פרמטר.

4. מצא $\tilde{x}_i$ שממזערים את $D_{KL}(\mathbb{P},\tilde{\mathbb{P}})$.

כאשר בחרתי את σ_i בחרתי אותו לקבל אנטרופיה של $\mathbb{P}_{j|i}$ כרצוני $-\sum\limits_j \mathbb{P}_{j|i}\log\left(\mathbb{P}_{j|i}\right)$.

כעת יש לי $\mathbb{P}_{j,i}=\frac{\mathbb{P}_{j|i}+\mathbb{P}_{i|j}}{2n}$. אני מחפש $\bar{y}_i$ עבורם הסיכוי לקפוץ מ-i ל-j הוא $\mathbb{Q}_{ij}$ עם התפלגות (ב). נגדיר $\mathbb{Q}_{ij}=\frac{\frac{1}{1+\|\bar{y}_i-\bar{y}_j\|}}{\sum\limits_{i,j}\frac{1}{1+\|\bar{y}_i-\bar{y}_j\|}}$ ואני רוצה להעריך את הדמיון בין ההתפלגות הנצפית שלי להתפלגות האמיתית, לכן נמדוד את הדמיון לפי $D_{KL}\left(\mathbb{P}_{ij},\mathbb{Q}_{ij}\right)$. איך מזה נמצא את $\bar{y}$? נעשה ירידה במורד הגראדיאנט.

20 אלגוריתם U-MAP.

זה $t-sne$ לקטני אמונה. נגדיר $\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{\|x_i-x_j\|^2}{2\sigma_i^2}}=\mathbb{P}_{ij}$ וכן $\mathbb{Q}_{ij}=\frac{1}{1+e^{-\frac{\|y_i-y_j\|^2+2\rho}{\sigma_y}}}$ המרחק בין $\mathbb{P}$ ל-$\mathbb{Q}$ יהיה

$$D\left(\mathbb{P},\mathbb{Q}\right)=\sum_{ij}\mathbb{P}_{ij}\ln\mathbb{Q}_{ij}+\sum_{ij}\left(1-\mathbb{P}_{ij}\right)\ln\left(1-\mathbb{Q}_{ij}\right)$$

21 מקודד אוטומטי.

יש לי תצפיות $\bar{x}_i$. יש לי מקודד $y_i=f(\bar{x}_i)$ וכן מפענח $\tilde{x}_i=g(y_i)$. אני רוצה $\tilde{x}_i\sim\bar{x}_i$. נגדיר את ההפסד להיות

$$L=\frac{1}{2}\sum_i D\left(g(f(\bar{x}_i)),\bar{x}_i\right)^2$$

ואני רוצה למזער את L. בדר"כ נבחר $D(x,y)=\|x-y\|^2$.

אני רוצה שהמקודד יהיה רציף מקומית, כלומר אם יש לי $x_i\to y_i\to\tilde{x}_i$ וכן $\tilde{x}_i\sim\bar{x}_i$ אזי $\bar{x}_i+\bar{\varepsilon}_1\sim\tilde{x}_i+\bar{\varepsilon}_2$.

כעת נניח אני רוצה ללמוד שני ערכים שונים לחלוטין מ-$f(x)$: $\mu_{\bar{x}}$ להיות ממוצע של x וכן $\sigma_{\bar{x}}$ להיות השונות של x. נגדיר $y(x)=\mu_x+\sigma_x\mathcal{N}(0,1)$ ואני רוצה למצוא $\hat{x}=g(y(\bar{x}))$ והפעם אני עובר מסביבה של $\bar{x}$ ל-y ל-$\tilde{x}$.

נגיד $\bar{x}$ במימד 1000, $\bar{y},\mu_x,\sigma_x$ ממימד 15. הבעיה היא שהמודל יקטין לי את σ עד לאיזה רמת דיוק שאני רוצה כדי לדייק לכן אני צריך להוסיף אילוץ $\bar{y}_i\sim\mathcal{N}(0,1)$ בכל מימד. לכן נעדכן את L להיות $L=\frac{1}{2}\sum\limits_i d\left(\tilde{x}_i,\bar{x}_i\right)^2+D_{KL}\left(\bar{y}_i,\mathcal{N}(0,1)\right)$. אם אני לא רוצה לחשוש מזה שחלק מהנתונים שלי נעלמים בזמן שאני מוצא את f ו-g אני יכול למצוא אותן כשכל פעם אני משתמש גם בחלק מהמידע שלי.

חלק VII

למידה מפוקחת עצמאית.

נניח יש לי ספר, ואת אותו ספר מתורגם לרוסית. איך אני יכול להבין שמדובר על אותו ספר?

נניח הספר בעברית הוא x_i והספר ברוסית הוא x_j. אני אייצר מלא מלא הפרעות על x_i ועם x_j:

$$x_i \to \{\tilde{x}_i\}$$
$$x_j \to \{\tilde{x}_j\}$$

כעת אני לוקח זוג $[x_1, x_2]$. ניצור הטלה שלהם $z_1, z_2 \in \mathbb{R}^n$ הטלות $\to x_1, x_2$. אני רוצה להגדיר הפסד שיביא לי ערך גבוה (מרחק נמוך) אם x_1, x_2 מאותה תמונה וערך נמוך (מרחק גבוה) אם x_1, x_2 לא מאותה תמונה. נגדיר SELF SUPERVIZED LOSS להיות

$$L = \frac{\sum_{ij} e^{-d\underbrace{(z_i, z_j)}_{\text{לא מאותה תמונה}}}}{\sum_{ij} e^{-d(z_i, z_j)}}$$

חלק VIII

זיהוי חריגים.

כיום יש 3 סוגים של חריגים:

1) Point Anomaly.
2) Contextual Anomaly.
3) Collective Anomaly.

יש כמה שיטות לזיהוי אנומליות:

- אשכול.
- צפיפות.
- מרחק.
- מודל סטטיסטי.
- Embedding.

מודל סטטיסטי לזיהוי אנומליות:

אנומליה = אירוע עם סיכוי נמוך.

בהינתן אוסף נקודות $\bar{x}_i$ ומודל שתלוי בפרמטרים סמויים $\bar{\theta}$, נחשב באמצעות EM את $\bar{\theta}$, אבל לא נכניס לחישוב ערכים עם סיכוי מתחת ל-ε (לבחירתי).

1. ניקח את הנקודות, נבנה מודל סטטיסטי.
2. נחשב סיכוי של נק' ונזהה אנומליות.
3. נוציא אנומליות מהנתונים.
4. נחזור ל-(1).

אשכול לזיהוי אנומליות:

נוכל לעשות $DBSCAN$ ואם יש לנו נקודה רחוקה מנקודת גרעין אזי היא חריגה.

נוכל גם לעשות $FUZZY-C-MEANS$ ונקודה ש"שייכת לכמה אשכולות" תהיה חריגה.

צפיפות לזיהוי אנומליות:

אפשר להפעיל KDE על הנתונים שלי ולהתבונן בערכי ה-KDE של הנקודות.

מרחק לזיהוי אנומליות:

נחשב מרחקים של כל הנקודות ונשאל מה המרחק של השכן ה-k.

בונוס: SVM לזיהוי אנומליות:

בהינתן מרכז $\bar{\mu}$ ורדיוס R נגדיר הפסד

$$L(x_i) = \begin{cases} 0 & \|x_i - \bar{\mu}\| < R \\ \|x_i - \bar{\mu}\| - R & else \end{cases}$$

לכן אני רוצה לפתור

$$\min_{\mu, R} \sum_i L(x_i) + f(R)$$

ואז אם $\|x_i - \bar{\mu}\| > R$ אזי x_i חריג. בדרך כלל: $f(R) = \alpha R^2$.

בחיים האמיתיים זה לא עובד.

נטיל למימד יותר גבוה: $x\to\phi(x_i)$ ונעשה

$$L(x_i)=\begin{cases}0 & \|\phi(x_i)-\bar{\mu}\|<R\\ \|\phi(x_i)-\bar{\mu}\|-R & else\end{cases}$$

$$\min_{\mu,R}\sum_i L(x_i)+f(R)$$

חלק IX

נתונים דינאמיים.

יש 2 סוגים של נתונים דינאמיים:

הראשון הוא בדידים בזמן: $O_1,O_2,O_3,...$ כאשר O_i הוא תצפית מזמן i.

הנחות שנניח לאורך הדרך:

א) מרקוביות: העתיד תלוי בהווה בלבד ולא בעבר:

$$\mathbb{P}\left(x_{t+1}|x_t,...,x_1\right)=\mathbb{P}\left(x_{t+1}|x_t\right)$$

ב) המציאות סמויה.

מקרה הכי פשוט:

א) הזמן בדיד.

ב) המצבים הסמויים סופיים ובדידים.

דינמיקה: $a_{ij}=\mathbb{P}\left(j|i\right)$

סיכוי התחלתי: $\mathbb{P}_i=\mathbb{P}\left(i\right)$.

סיכוי לתצפית $b_{ik}=\mathbb{P}\left(\text{תצפית }k\text{ ו מצב }i\right)$.

שאלות:

א) בהינתן מודל $(\mathbb{P},A,B)$ וסדרת תצפיות $o_1,...,o_T$., מה ההסתברות לסדרת התצפיות?

ב) בהינתן מודל ותצפיות כנ"ל, מה רצף האירועים הכי סביר (רצף המצבים הסמויים)?

ג) בהינתן אוסף של תצפיות- מה המודל הכי סביר?

22 מודלים מרקובים סמויים.

נגדיר:

סיכוי התחלתי לכל מצב: P.

סיכוי לעבור ממצב i למצב j: A.

סיכוי לראות תצפית k בהינתן מצב i: B.

דוגמה: $P=\begin{pmatrix}0.5\\0.3\\0.2\end{pmatrix}$, $A=\begin{pmatrix}0.8&0&0.2\\0&0.4&0.6\\0.1&0&0.9\end{pmatrix}$, $B=\begin{array}{cc}\begin{pmatrix}0.2&0.1&0.1&0.1&0.5\\0.7&0&0&0.3&0\\0&0.5&0.5&0&0\end{pmatrix} & \begin{matrix}1\\2\\3\end{matrix}\\ \begin{matrix}\text{ה}&\quad\text{ד}&\quad\text{ג}&\quad\text{ב}&\quad\text{א}\end{matrix} & \end{array}$.

אני רוצה לשאול מה הסיכוי של רצף תצפיות $\mathbb{P}\left(O_1,...,O_T\right)$. לפי משפט ההתפלגות השלמה

$$\mathbb{P}\left(O_1,...,O_T\right)=\sum_i\mathbb{P}\left(O_1,...,O_T,T\text{ מצב }i\text{ בזמן }\right)=\sum_i\alpha_i\left(T\right)$$

כאשר

$$\alpha_i\left(1\right)=\mathbb{P}_ib_{i,O_1}$$

ואז

$$\alpha_i(t)=\sum_j\mathbb{P}\left(O_1,...,O_{t-1},t-1\text{ מצב }j\text{ בזמן }\right)a_{ji}b_{i,O(T)}$$

בעת, בהינתן אוסף תצפיות אני רוצה לשאול מה הרצף הכי סביר. ביותר פורמלי: מה הרצף $S_1,...,S_t$ הכי סביר לתצפיות $O_1,...,O_t$. אני רוצה

$$\delta_i(t)=\mathbb{P}\left(S_1,...,S_t,O_1,...,O_t\right)$$

כתור התחלה:

$$\delta_i(1) = \mathbb{P}_i b_{i,O_1}$$

וכן

$$\delta_i(t+1) = \mathbb{P}\left(S_1,...,S_t,O_1,...,O_t\right) = \max\left(\mathbb{P}_i\left(S_1,...,S_{t-1},O_1,...,O_{t-1}\right)a_{ji}b_{i,o_t}\right)$$

בהינתן אוסף רצפי זמן של תצפיות, מה המודל הכי סביר? נעשה EM:

א) נניח מודל (P_0,A_0,B_0).

ב) נחשב רצף מצבים הכי סביר לכל רצף תצפיות.

ג) נחשב מודל חדש לפי רצף מצבים.

ד) כל עוד לא התכנסנו נחזור ל-(ב').

יש לי כמה שאלות כעת:

א) מה הסיכוי לרצף תצפיות? אם יש לנו

$$\alpha_i(T) = \mathbb{P}\left(O_1,...,O_T, i \text{ להיות במצב}\right)$$
$$\alpha_i(1) = \mathbb{P}_i b_{i,O_1}$$
$$\alpha_i(t+1) = \sum_j \alpha_j(t)a_{ji}b_{i,O_{t+1}}$$

אז

$$\mathbb{P}\left(O_1,...,O_T\right) = \sum_j \alpha_j(t)$$

ב) מה רצף המצבים הכי סביר בהינתן רצף תצפיות? זה מה שעשינו:

$$\delta_i(t) = \mathbb{P}\left(O_1,...,O_t, \text{במסלול הכי סביר } i \text{ להיות במצב}\right)$$

וכן

$$\delta_i(1) = \mathbb{P}_i b_{i,O_1}$$

אז

$$\delta_i(t+1) = \max\left(\delta_i(t)a_{ji}b_{i,o_t}\right)$$

נמצא את מקסימום $\delta_i(T)$ וחוזרים אחורה במסלול הכי סביר- כמו בתכנון דינאמי.

ג) בהינתן הרבה רצפי תצפיות, מה הוא המודל הכי סביר (לוקאלית)? נגדיר

$$\gamma_i(t) = \mathbb{P}\left(O_1,...,O_T, t \text{ בזמן } i \text{ להיות במצב}\right)$$
$$\beta_i(t) = \mathbb{P}\left(O_1,...,O_T | t \text{ בזמן } i \text{ להיות במצב}\right)$$

ואז

$$\beta_i(T) = 1$$
$$\beta_j(t+1) = \sum_i a_{ji}\beta_i(t)b_{i,O_t}$$

וכן גם

$$\alpha_i(t) = \sum_j a_{ji}\alpha_i(t-1)b_{i,O_t}$$

אז נקבל

$$\gamma_i(t) = \mathbb{P}\left(O_1,...,O_t | t \text{ בזמן } i \text{ להיות במצב}\right)\mathbb{P}\left(O_{t+1},...,O_T | t \text{ בזמן } i \text{ הייתי במצב}\right) = \alpha_i(t)\beta_i(t)$$

עכשיו אנחנו רוצים לדעת את הסיכוי של i בזמן t:

$$\mathbb{P}\left(i,\ t \text{ בזמן}\right) = \frac{\mathbb{P}\left(it \text{ בזמן}, O_1,...,O_T\right)}{\mathbb{P}\left(O_1,...,O_T\right)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{i'}\alpha_{i'}(t)\beta_{i'}(t)} \Rightarrow \sum_i \mathbb{P}\left(i,t \text{ בזמן}\right) = 1$$

מכאן שסיכוי ההתחלה החדש

$$_{new}\mathbb{P}_i(1) = \frac{\alpha_i(1)\beta_i(1)}{\sum_{i'}\alpha_{i'}(1)\beta_{i'}(1)}$$

ואז גם

$$_{new}a_{ij} = \frac{\sum\limits_{\text{זמנים}} \text{מעברים מ-}i\text{ ל-}j}{\sum\limits_{\text{זמנים}} \text{סיכוי להיות במצב } i} = \frac{\sum\limits_t \alpha_i(t) a_{ij} b_{i,O_{t+1}} \beta_j(t+1)}{\sum\limits_t \gamma_i(t)}$$

וכן

$$_{new}b_{ik} = \frac{\sum\limits_t \text{הייתי במצב } i \cdot \text{ראיתי תצפית } k}{\sum\limits_t \text{הייתי במצב } i} = \frac{\sum\limits_t \gamma_{it} \mathbb{I}_{O_k,t}}{\sum\limits_t \gamma_{it}}$$

וקיבלתי מודל חדש, כלומר זה שלב ה-EM. כמו בכל EM את המודל הזה אני שוב פעם אשפר עד שאני מגיע לקיצון מקומי.

23 טיפול במשתנים קטגוריאלים- אלגוריתם MCA.

יהי $x_i \in \{z_1, ..., z_k\}$. זה משתנה בדיד. איך אני עובר ממנו למשנה רציף?

אם זה בינארי עוברים לוקטור $0, 1$:

כלומר אם לא $x_i = 0$, אם כן $x_i = 1$ ואם לא ידוע עד שמים חציון/ ממוצע. אם יש יותר מ-2 אפשרויות נעשה one-hot: אם יש לי k קטגוריות אני אצור k עמודות וכל דגימה תקבל 1 בקטגוריה שלה.

לדוגמה: אם יש לי רשימה {אמא, אח, אמא, אבא, אבא} אז נקבל

אח	אבא	אמא
0	1	0
0	1	0
0	0	1
1	0	0
0	0	1

אם אני רוצה לעשות PCA עם one hot לא נקבל תוצאות טובות- לשם כך יש אלגוריתם MCA:

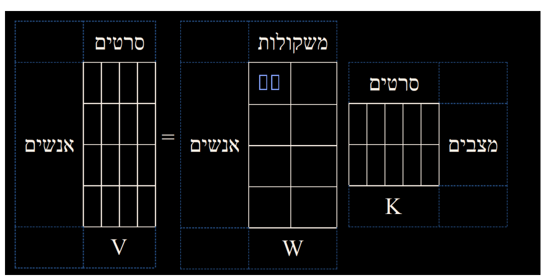
נמרכז את x: נוריד מכל עמודה את הממוצע שלה. אחרי זה נחשב $B = x^T x$, מטריצה בגודל מספר העמודות $\times$ מספר העמודות. נטיל על הו"ע של B כמו ב-PCA ואם יש לי שורה של לא ידוע נחליף בשורת אפסים.

אם יש לי דאטא לא קטגוריאלי משולב עם כן קטגוריאלי, נפצל אותם, נעשה לכל אחד את ההורדת מימדים שלו ואז נאחד אותם ועל התוצאה נעשה עוד הורדת מימדים.

אחרי MCA: קיבלתי וקטור רציף, אפשר לשלב עם משתנים רציפים אחרים אם רוצים להמשיך להוריד מימד.

24 בונוס.

כעת יש לי שאלה אחרת: בהינתן נדב ורשימת הסרטים שהוא ראה, אני רוצה לחזות איזה סרטים הוא רוצה לראות. כלומר אם יש לי מטריצה של ראה סרט לא ראה סרט אני רוצה



אני רוצה $V \sim WH =: \tilde{V}$. אני רוצה $\tilde{V}$עם ערכים שהם לא שליליים. איך אני עושה את זה? הכי פשוט: נכריח את H, W להיות רק עם ערכים חיוביים. אני לא יכול לעשות $W^{t+1} = W^t +$? כי? יכול להרוס לי את הכל ולהכניס מינוסים. לכן אני אעשה (משתנה חיובי) $\cdot$ $W_{ij}^{t+1} = W_{ij}^t$. איך אני עושה את זה? בעזרת מורד הגרדיאנט עם טוויסט:

בהינתן $F(h)$ אני אגדיר $G(h, h')$ כך ש-$G(h, h') \geq F(h)$ וכן $G(h, h) = F(h)$. כדי למזער כל פונקציה $F(h)$:

א) נגדיר $G(h, h')$.

ב) נגדיר $h^0 = h$ עבור h כרצוני וכן $h^{t+1} = \arg\min G(h, h^t)$. כעת, $F(h^{t+1}) = G(h^{t+1}, h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t)$. נדלג על המון המון שלבים ואז נקבל את העדכון

$$_{new}H_{kj} = \frac{H_{kj} \left[W^T V\right]_{kj}}{\left[W^T W H\right]_{kj}}$$

$$_{new}W_{ik} = \frac{H_{ik} \left[V H^T\right]_{ik}}{\left[W H H^T\right]_{ik}}$$

וסיימנו את הקורס.