מבוא למערכות לומדות - הרצאה 10 - למידת ייצוג (הורדת מימד)

11 ביוני 2015

יהא נתון מדגם $S=\{x_1,\ldots,x_m\}\subset\mathbb{R}^n$. שיטות הורדת מימד מחפשות תת מרחב V ממימד $k\ll n$ והעתקה $\Psi:\mathbb{R}^n\to V$ המהווה ייצוג טוב של המדגם. בהינתן ייצוג כנ"ל, אנו יכולים לייצג כל דוגמא x ע"י k מספרים בלבד (המקדמים $\Psi(x)$ בהצגתו כצירוף לינארי לפי איזשהו בסיס של V). ייצוג קומפקטי כזה מאפשר לעבוד עם מחלקות עשירות יותר, ועשוי לשפר משמעותית את זמן האימון, ואת זמן הריצה של ההיפותזה הנלמדת.

אנו נלמד שתי שיטות להורדת מימד:

- הראשונה נקראית PCA ומוצאת ת"מ $V\subset\mathbb{R}^n$ שהוא הקרוב ביותר (במובן מסויים) לנקודות המדגם, ומגדירה את $\Psi:\mathbb{R}^n\to V$ להיות ההטלה האורתוגונלית על V.
- השיטה השנייה שנלמד נקראית "הטלות מקרית" ובה אנו פשוט לוקחים העתקה לינארית אקראית $\Psi:\mathbb{R}^n\to\mathbb{R}^k$. למרבה הפלא, השיטה הנ"ל מובילה לעיתים לתוצאות טובות, ועל כן מהווה חלופה מהירה ל-PCA.

בסוף השיעור נאמר כמה מילים על שיטות הנקראות "למידת מילון" הדומות להורדת מימד.

1 ניתוח גורמים ראשיים (PCA)

יהא נתון מדגם $S=\{x_1,\ldots,x_m\}\subset\mathbb{R}^n$. נסמן ב-

$$P_V:\mathbb{R}^n\to V$$

את **ההטלה האורתוגונלית על** V. ניזכר בכמה עובדות בסיסיות מאלגברה לינארית:

טענה 1.1 1. לכל $x\in\mathbb{R}^n$, $P_V(x)$ היא הנקודה ב-V הקרובה ביותר (לפי המרחק האוקלידי) ל-x כלומר, מתקיים

$$d(x,P_V(x))=\min_{v\in V}d(x,v)=:d(x,V)$$

2. P_V הינה העתקה לינארית

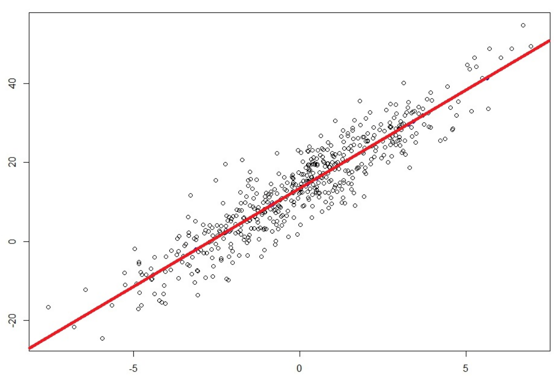
3. אם $U \in M_{n\times k}$ היא מטריצה שעמודותיה מהוות בסיס א"נ של V אזי

$$P_V(x) = UU^T x$$

אלגוריתם ה־PCA מחפש בסיס למרחב V ממימד k הממזער את את סכום (או, באופן שקול, ממוצע) ריבועי המרחקים של נקודות המדגם מ־V. כלומר מרחב הממזער את

$$L^{\mathrm{PCA}}(V) := \sum_{i=1}^{m} d^2(x_i, V) = \sum_{i=1}^{m} d^2(x_i, P_V(x_i))$$

לדוגמא, בתמונה הבאה, הקו האדום הוא המרחב ממימד 1 הממזער את סכום ריבועי המרחקים של הנקודות ממנו.



למרבה הפלא, ניתן למצוא ביעילות תת מרחב V הממזער את ממוצע ריבועי המרחקים. הדרך בה עושים זאת מסתמכת על אלגברה לינארית, או באופן יותר קונקרטי, התורה של לכסון אורתוגונלי של מטריצות סימטריות. נסמן

$$A = \sum_{i=1}^{m} x_i x_i^T \in M_{n\times n}$$

המטריצה A הינה מטריצה סימטרית מוגדרת חיובית למחצה[1] (תרגיל). מאלגברה לינארית, אנו יודעים שקיים בסיס אורתונורמלי של ו"ע

$$u_1, \ldots, u_n \in \mathbb{R}^n$$

[1]כלומר, מתקיים $A = A^T$ ובנוסף $x^T Ax \geq 0$ לכל $x \in \mathbb{R}^n$.

עם ע"ע אי־שליליים

$$\lambda_1 \geq \ldots \geq \lambda_n \geq 0$$

מתברר ש־

משפט 1.2 (PCA) המרחב הנפרש ע"י $u_1, \ldots, u_k$ ממזער את L^{PCA} על פני כל המרחבים ממימד $\geq k$.

ההוכחה מתבססת על אלגברה לינארית, ובאופן יותר קונקרטי, המשפט הספקטרלי, לפיו למטריצה סימטרית קיים לכסון אורתוגונלי. אנו נאמר כמה מילים על ההוכחה, אך לא נוכיח את המשפט באופן מלא (ראו תרגיל רשות, וכמו כן, פרק 23 בספר של שי ושי). נסכם:

PCA

קלט: מדגם $S = \{x_1, \ldots, x_m\} \subset \mathbb{R}^n$, מימד $k \geq 1$

1. הגדר $A = \sum_{i=1}^m x_i x_i^T$
2. מצא ל־A בסיס א"נ של ו"ע $u_1, \ldots, u_n \in \mathbb{R}^n$ עם ע"ע $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$
3. החזר את $u_1, \ldots, u_k$

1.1 הערות

מציאת בסיס א"נ

על מנת להפעיל את האלגוריתם, עלינו למצוא ל־A בסיס א"נ. נעיר שקיימים אלגוריתמים יעילים סטנדרטיים העושים זאת.

מרכוז

הרבה פעמים, מועיל לעשות מרכוז של המדגם לפני שמפעילים PCA. כלומר, להחליף את $x_1, \ldots, x_m$ ב־$(x_1 - \mu), \ldots, (x_m - \mu)$ כאשר $\mu = \frac{1}{m}\sum_{i=1}^m x_i$. הפעולה הנ"ל ממרכזת את המדגם, במובן שוקטור ה־0 הופך להיות המרכז (הממוצע) של הנקודות.

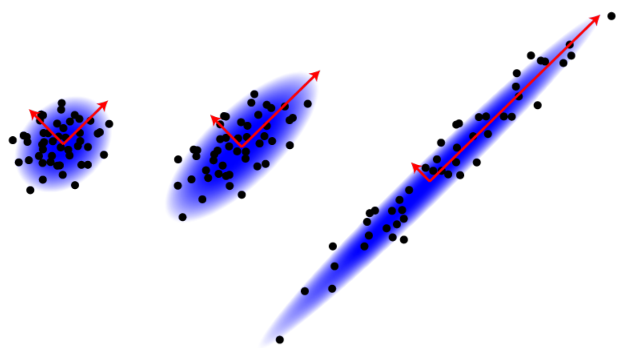
המטריצה A ופרשנות גיאומטרית של הגורמים הראשיים

הוקטורים $u_1, \ldots, u_n$ נקראים **הגורמים הראשיים (Principal Components)** של המדגם. הם מאופיינים באופן הבא

- הוקטור u_1 מגדיר את המרחב החד מימדי (כלומר, קו) עם התכונה הבאה: אם מטילים את נקודות המדגם עליו, סכום ריבועי הנורמות של הנקודות המוטלות הוא הגדול ביותר (על פני כל המרחבים ממימד 1).

- עבור $i > 1$, u_i מגדיר את המרחב החד מימדי עם התכונה הבאה: אם מטילים את נקודות המדגם עליו, סכום ריבועי הנורמות של הנקודות המוטלות הוא הגדול ביותר, על פני כל המרחבים ממימד 1 הניצבים ל־$\text{span}\{u_1, \ldots, u_{\{i-1\}}\}$.

בתמונה הבאה, מופיעים (עבור שלושה מדגמים שונים) המרחבים החד מימדיים המוגדרים ע"י שני הגורמים הראשיים הראשונים (אורך החצים פרופורציונאלי לסכום ריבועי הנורמות של הנקודות, אחרי שמטילים אותן על הקו):



האפיון הנ"ל נובע מהעובדה הבאה, שהוכחתה מושארת כתרגיל.

למה 1.3 תהא $A \in M_{n \times n}$ מטריצה סימטרית. יהא $u_1, \ldots, u_n$ בסיס א"נ של ו"ע של A עם ע"ע $\lambda_1 \geq \ldots \geq \lambda_n$ אזי

1. u_1 ממקסם את $u^T A u$ על פני כל הוקטורים מנורמה 1.
2. לכל $i > 1$, u_i ממקסם את $u^T A u$ על פני כל הוקטורים מנורמה 1 הניצבים ל־$\text{span}\{u_1, \ldots, u_{\{i-1\}}\}$.

בהקשר שלנו, $A = \sum_{i=1}^m x_i x_i^T$. לכן לכל וקטור יחידה $u \in \mathbb{R}^n$ מתקיים

$$u^T A u = \sum_{i=1}^m u^T x_i x_i^T u = \sum_{i=1}^m \langle u, x_i \rangle^2$$

הביטוי הימני הינו סכום ריבועי ההטלות של הדוגמאות על $\text{span}\{u\}$. לכן, האפיון של הגורמים הראשיים נובע מהלמה הנ"ל.

שקילות של PCA למקסום סכום הנורמות של ההטלות, ומילה על ההוכחה

יהא $V \subset \mathbb{R}^n$ ת"מ ממימד k מתקיים

$$L^{\text{PCA}}(V) \quad = \quad \sum_{i=1}^m \|x_i - P_V(x_i)\|^2$$

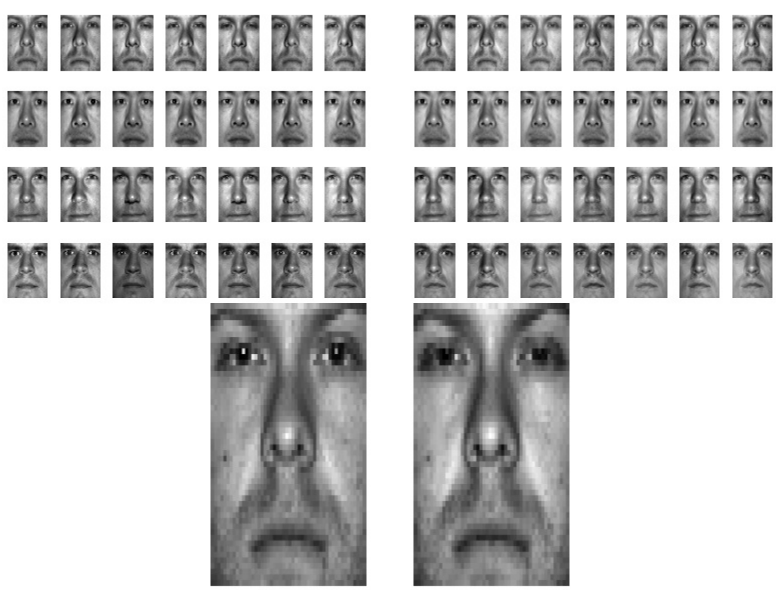
$$
\begin{aligned}
&= \sum_{i=1}^{m} \|x_i\|^2 - 2\langle x_i, P_V(x_i)\rangle + \|P_V(x_i)\|^2 \\
&= \sum_{i=1}^{m} \|x_i\|^2 - 2\langle P_V(x_i), P_V(x_i)\rangle + \|P_V(x_i)\|^2 \\
&= \sum_{i=1}^{m} \|x_i\|^2 - \|P_V(x_i)\|^2
\end{aligned}
$$

(כאן, השוויון השלישי נובע מכך שעבור הטלות $P : \mathbb{R}^n \to \mathbb{R}^n$, מתקיים $\forall x,\ \langle x, P(x)\rangle = \langle P(x), P(x)\rangle$). נשים לב שהביטוי $\sum_{i=1}^{m} \|x_i\|^2$ אינו תלוי ב־V. לכן, מציאת V הממזער את $L^{\mathrm{PCA}}(V)$ שקולה למציאת V הממקסם את $\sum_{i=1}^{m} \|P_V(x_i)\|^2$. כלומר, מציאת מרחב V הממקסם את סכום הריבועים של ההטלות של נקודות המדגם עליו.

מהפסקה הקודמת, נובע שעבור $k = 1$, הבחירה האופטימלית היא $V = \mathrm{span}\{u_1\}$. על מנת להוכיח את משפט 1.2 מראים שעבור k כלשהו, הבחירה האופטימלית היא $V = \mathrm{span}\{u_1, \ldots, u_k\}$.

דוגמא

בתמונה הבאה מודגמת הטלה, באמצעות PCA, של אוסף של תמונות שחור־לבן בגודל 50×50 (שכל אחת מיוצגת ע"י וקטור ב־$\mathbb{R}^{2500}$). בצד שמאל, מוצגות התמונות המקוריות, ובצד ימין התמונות שהתקבלו אחרי הטלה למימד 10. כפי שאפשר לראות, הייצוג החדש של התמונות הינו די טוב - לא איבדנו הרבה כאשר הורדנו דרסטית את המימד.



2 הטלות מקריות

יהא נתון מדגם $S = \{x^1, \ldots, x^m\} \subset \mathbb{R}^n$. דרך נאיבית למצוא ייצוג של המדגם ע"י מ"ו וקטורי ממימד $k \ll n$ היא לבחור מטריצה אקראית

$$W \in M_{k \times n}$$

ולקחת בתור ייצוג את ההעתקה (הלינארית) $\Psi : \mathbb{R}^n \to \mathbb{R}^k$

$$\Psi(x) = Wx$$

השיטה הנ"ל נראית מאד נאיבית - בבחירת הייצוג אנו בכלל לא מסתכלים על המדגם! למרות זאת, אנו נראה שבהסתברות גבוהה אנו מקבלים ייצוג של המדגם המקורי, **ללא עיוות גדול** של המרחקים. כלומר, עבור כל i, j מתקיים

$$d(x^i, x^j) \approx d(\Psi(x^i), \Psi(x^j)) \tag{1}$$

לכן, לעיתים, הטלות מקריות יכולות להוות תחליף מהיר ופשוט ל-PCA

ניגש לתיאור יותר מפורט. ראשית, הדרך בה נבחר את W היא ע"י כך שהרכיבים $\{W_{ij}\}_{i\in[k], j\in[n]}$ יהיו משתנים מקריים בלתי תלויים המתפלגים נורמלית עם תוחלת 0 ושונות $\frac{1}{k}$. נזכיר ש-

- משתנה מקרי נורמלי עם תוחלת μ ושונות σ^2 הוא מ"מ עם פונקציית צפיפות

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- אם $X_1, \ldots, X_n$ הם מ"מ נורמליים בלתי תלויים עם תוחלות $\mu_1, \ldots, \mu_n$ ושונויות $\sigma_1^2, \ldots, \sigma_n^2$ אז לכל

$$\alpha_1, \ldots, \alpha_n \in \mathbb{R}$$

המשתנה המקרי

$$\alpha_1 X_1 + \ldots + \alpha_n X_n$$

הוא מ"מ נורמלי עם תוחלת $\alpha_1\mu_1 + \ldots + \alpha_n\mu_n$ ושונות $\alpha_1^2\sigma_1^2 + \ldots + \alpha_n^2\sigma_n^2$.

ניגש כעת להראות שמתקיים (1) בהסתברות גבוהה (על פני בחירת W). נראה זאת קודם עבור זוג בודד, ואח"כ נשתמש בחסם האיחוד על מנת לעבור לכל הזוגות.

למה 2.1 לכל זוג דוגמאות x^i, x^j

$$\Pr_W\left(\left|\frac{d^2(\Psi(x^i), \Psi(x^j))}{d^2(x^i, x^j)} - 1\right| > \epsilon\right) \le 2\exp\left(-\frac{k\epsilon^2}{6}\right)$$

אנו נסתמך על הלמה הבאה (ראו הוכחה בפרק B.7 בספר), המהווה מעין אנלוג לחסם הופדינג, עבור ריבועים של מ"מ נורמליים

למה 2.2 יהיו $Z_1, \ldots, Z_k$ מ"מ נורמליים ב"ת ש"ה עם תוחלת 0. נסמן

$$Z = \sum_{i=1}^{k} Z_i^2, \;\; \sigma^2 = E[Z]$$

אזי,

$$\Pr\left(\left|\frac{Z}{E[Z]} - 1\right| > \epsilon\right) \leq 2\exp\left(-\frac{k\epsilon^2}{6}\right)$$

הוכחה: נסמן $x = x^i - x^j$. מתקיים $\|x\| = d(x^i, x^j)$. כמו כן, $d(\Psi(x^i), \Psi(x^j)) = \|Wx^i - Wx^j\| = \|Wx\|$. לכן, די להראות שמתקיים

$$\Pr\left(\left|\frac{\|\Psi(x)\|^2}{\|x\|^2} - 1\right| > \epsilon\right) \leq 2\exp\left(-\frac{k\epsilon^2}{6}\right) \tag{2}$$

נכתוב

$$\Psi(x) = \begin{pmatrix} \Psi_1(x) \\ \vdots \\ \Psi_k(x) \end{pmatrix} = \begin{pmatrix} W_{11}x_1 + \ldots + W_{1n}x_n \\ \vdots \\ W_{k1}x_1 + \ldots + W_{kn}x_n \end{pmatrix}$$

מכיוון ש־x קבוע, הקואורדינטות של $\Psi(x)$ הן ב"ת. כמוכן, כל אחת מהן היא צירוף לינארי של מ"מ נורמליים ב"ת, ולכן כל קואורדינטה היא משתנה מקרי נורמלי עם תוחלת (בה"כ נביט על הקואורדינטה הראשונה)

$$x_1 E[W_{11}] + \ldots + x_n E[W_{1n}] = 0$$

ושונות

$$\mathrm{var}\left(\sum_{i=1}^{n} x_i W_{1i}\right) = \sum_{i=1}^{n} \mathrm{var}(x_i W_{1i}) = \sum_{i=1}^{n} x_i^2 \mathrm{var}(W_{1i}) = \sum_{i=1}^{n} x_i^2 \frac{1}{k} = \frac{\|x\|^2}{k}$$

מכאן, $\Psi(x) = \begin{pmatrix} \Psi_1(x) \\ \vdots \\ \Psi_k(x) \end{pmatrix}$ הוא וקטור מקרי עם קואורדינטות ב"ת שכל אחת מתפלגת נורמלית עם תוחלת 0 ושונות $\frac{\|x\|^2}{k}$. בפרט,

$$E[\|\Psi(x)\|^2] = \sum_{i=1}^{k} E\|\Psi_i(x)\|^2 = \sum_{i=1}^{k} \frac{\|x\|^2}{k} = \|x\|^2$$

מלמה 2.2 מתקיים (2) ∎

כעת, מחסם האיחוד, נובע שכאשר $k \gg \log(m)$ מתקיים (1):

משפט 2.3 (ג'ונסון - לינדנשטראוס) נסמן $\epsilon = \sqrt{\frac{6\log\left(\frac{m^2}{\delta}\right)}{k}}$. אזי, בהסתברות $\leq 1-\delta$ על פני הבחירה של W מתקיים

$$\forall i,j \in [m], \quad \left|\frac{d^2(\Psi(x^i),\Psi(x^j))}{d^2(x^i,x^j)} - 1\right| \leq \epsilon$$

3 הצצה ללמידת מילון

יהא נתון מדגם $S = \{x_1,\ldots,x_m\} \subset \mathbb{R}^n$. בלמידת מילון מחפשים $K \ll m$ נקודות (הנקראות **מילים**) $v_1,\ldots,v_K \in \mathbb{R}^n$ המייצגות טוב את המדגם במובן שרוב הדוגמאות הן, בקירוב, צירוף לינארי של מספר קטן, $k \ll K$ של מילים. כלומר, רוב הדוגמאות מקיימות

$$x_j \approx \sum_{i=1}^{K} \alpha_i^{(j)} v_i$$

עבור וקטור $(\alpha_1^{(j)},\ldots,\alpha_K^{(j)}) \in \mathbb{R}^K$ עם $\geq k$ קואורדינטות שאינן 0.

בדומה ל־k-means, רוב האלגוריתמים הלומדים מילון מבוססים על Alternate Minimization: הם מחזיקים מילון

$$v_1,\ldots,v_K$$

וכמו כן, מקדמים

$$(\alpha_1^{(j)},\ldots,\alpha_K^{(j)})$$

לכל דוגמא. האלגוריתמים מתחילים עם בחירה אקראית של המילון. ואז, בכל שלב, מבצעים:

- **אופטימיזציה על המקדמים** - בהינתן המילון הנכחי, לכל דוגמא x_j, מעדכנים את $(\alpha_1^{(j)},\ldots,\alpha_K^{(j)})$ כך שלמשל, ימזער את

$$\left\|x_j - \sum_{i=1}^{K} \alpha_i^{(j)} v_i\right\|^2$$

על פני כל הוקטורים $(\alpha_1,\ldots,\alpha_K) \in \mathbb{R}^K$ עם $\geq k$ קואורדינטות שאינן 0.

- **אופטימיזציה על המילון** - בהינתן המקדמים הנוכחיים, מעדכנים את המילון כך שלמשל, ימזער את

$$\sum_{j=1}^{m}\left\|x_j - \sum_{i=1}^{K} \alpha_i^{(j)} v_i\right\|^2$$

נעיר שבדומה ל־k-means, הבעיה של מציאת המילון האופטימלי הינה בעיה קשה. למעשה, הקושי שלה חמור אפילו יותר. ב־k-means, בהינתן המרכזים, קל היה לחשב את הייצוג של וקטור חדש $x \in \mathbb{R}^n$ (הייצוג הוא פשוט המרכז הקרוב ביותר). בלמידת מילון, אפילו בהינתן המילון, קשה לחשב את הייצוג של וקטור $x \in \mathbb{R}^n$ ביחס למילון. כלומר, קשה למצוא מקדמים $(\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K$ עם $k \geq$ קואורדינטות שאינן 0 הממזערים את $\left\| x - \sum_{i=1}^K \alpha_i v_i \right\|$. לעובדה הנ"ל יש שתי השלכות:

- לאחר שלמדנו את המילון, בהינתן דוגמא חדשה $x \in \mathbb{R}^n$, הייצוג שלה יחושב **באופן יוריסטי**. למשל, ע"י האלגוריתם החמדני המתחיל עם $(\alpha_1, \ldots, \alpha_K) = (0, \ldots, 0)$ ואז מבצע k עדכונים, שבכל אחד מהם משנים את הערך של קואורדינטה בודדת, באופן אופטימאלי מבין כל השינויים הללו.
- במהלך למידת המילון, בשלב של אופטימיזציה על המקדמים, עלינו לחשב את הייצוג של הדוגמאות לפי המילון הנכחי. החישוב הנ"ל אף הוא יחושב ע"י יוריסטיקה, בד"כ אותה יוריסטיקה בה נשתמש בהמשך על מנת לחשב את הייצוג של דוגמאות חדשות

לסיום, נעיר שבניגוד לשלב של אופטימיזציה על המקדמים בהינתן המילון, ניתן לבצע ביעילות את השלב של אופטימיזציה על המילון בהינתן המקדמים, שכן, מדובר בבעיה קמורה.