תרגול # 12 ב' - 22/6/20

★ תרגול # 11 היה תרגול הכנה לקראת המבחן.

נדבר על: הורדת מימד - מוטיבציה

- PCA: 2 פרושים שונים (משקולים):
  - את המרחב האפיני הקרוב ביותר.
  - את המרחב שממקסם את השונות.

מוטיבציה:

לדאטה ממימד גבוה יש קשה מלים בפנינו 2 סוגים של אתגרים:

1. חישובי - כל שמס הפיצ'רים עלה, עולה סיבוכיות מחלקת ההיפותזות, ונצטרך יותר דגימות כדי ללמוד, ויקח לנו יותר זמן לעשות זאת (כי רוב האלגוריתמים שרגילים הם פולינומיאליים בזמן ובמ- m).
2. data-exploration: נרצה לחקור את הדאטה, אך זה קשה לחקור כמות גדולה של פיצ'רים: להבין את ההתפלגות שלהם, מה היחסים ביניהם, איך הם משפיעים על המציאה... כמו כן -

data-visualization: מימדים גבוהים לא ניתן לצייר את הדאטה.

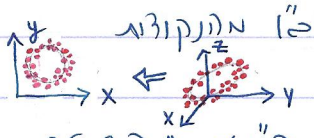
⇐ נרצה לעשות הליך הורדת מימד: $\varphi: \mathbb{R}^d \to \mathbb{R}^k$, $(k << d)$ $\{x_i\}_{i=1}^m \mapsto \{\varphi(x_i)\}_{i=1}^m$.

נקבע טרנספורמציה φ על הדאטה שלנו שתביא עבורנו שלומים (לינאריים או לא לינאריים) ל הפיצ'רים במרחב חדש, שבו נקבע למידה.

★ אלא אם הדאטה בשמת יושב ב-$\mathbb{R}^k$, תהליך הורדת מימד כרוך בשימוש מידע. בתהליך זה נרצה לשרוט מחלק שאנו אוהבים משמרים.

דוגמאות:

- יש מעגל (טבעת) ב-$\mathbb{R}^3$ שמקביל למישור xy. אז כ"ו מהנקודות הן אז נק' על המעגל + רעש גאוסי מסביב לכל הכיוונים.



מכיוון שאותנו נעניין שהנק' יושבות על מעגל, וזה החלה ה"מעניין" בדאטה (העיקרי), יכול להיות שנרצה להפחית מימד למימד 2, וזה יקל עלינו את העבודה.

- נניח שאנו חוקרים מחלה מסוימת, וחוקרים עבורה רבים גנטי. רוצים לסווג אנשים, שחולים במחלה וכאלו שלא, ולומר האם יש אצל אדם מוטציות גנטיות שגורמות למחלה.

גנום הוא כבר ארוך מאד האלפבית $\Sigma = \{A,T,C,G\}$ באורך 3.2×10^9. נשווה את הגנום של כמה אנשים, נמצא את ההבדלים ונשאל אם עצמנו האם להבדלים האלו שמצאנו יש קשר למחלה.

האם יש לנו באמת 3.2×10^9 פיצ'רים ?!

<u>דוגמא 1:</u> נבטמבא לפיצ'רים שבהם יש לנו הבדלים בין אנשים שונים.

⇐ הבדלתנו להבטמבא ל- 300×10^6 פיצ'רים.

<u>דוגמא 2:</u> נאדיר שנים של אדם מסוים להיות ה-"reference", כל יתר השנים נשווה ל-"reference" הזה.

⇐ הבדלתנו להבטמבא ל- 5×10^6 פיצ'רים.

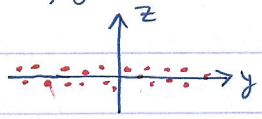
★ הדוגמאות באו להראות שניתן לנבל יתר קודם על הבעיות כדי להקטין את מימדיהן.

★ זה לא Feature Selection. מוותרים על חלק מהמימדים.

כשאנחנו מקבלים הורדת מימד אנחנו מאבדים נפח מהבאוטה. תלוי שהורדת המימד נשמר את החלק שאנחנו רוצים. <u>איך הבעיות ניתן לשמר?</u>

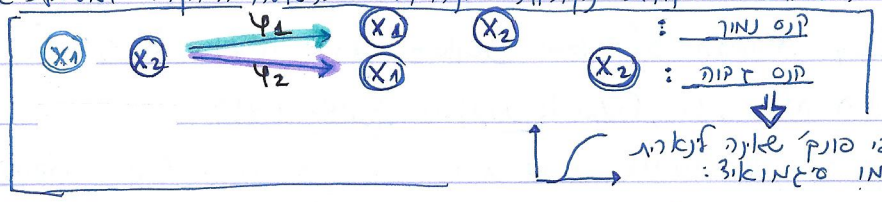
1. לשמור על השונות בבאוטה, כדי שיהיה להמשיך לחקור את הבעיה שאותה אנו מעניינים ללמוד.

<u>דוג':</u> בדוגמת המעגל, אם נטיל את הבאוטה על מישור zy נקבל נקודות מרוכזות סביב ציר ה-y, בהיפוך שנקבל לפי הנתון הבאוסני שהוספנו:



אם הבאוטה לא מקבל למישור, אלא נמצא בזווית של $\theta \in (0°, 90°)$ אז ההטלה היתה מתבצעת על קומבינציה כלשהי של 3 הצירים, כך שלא ניתן את המרחב שנקבל בסוף יפיק לנו תמונה כמו ההטלה הקודמת על xy.

2. לשמר מרחקים קרובים. נציג את <u>אלגוריתם טיסנים</u> שמבצע הורדת מימד לא לינארית: נתון שאופן שונה (לא לינארי) הטלות של נקודות שמצויות קרוב זו לזו בבאוטה המקורי אך רחוקות זו מזו במימד המופחת לעומת הטלות שלוקחות נקודות שקרובות בבאוטה המקורי ומטילות אותן קרוב זו לזו.



3. לשמר צורה פנימית של הבאוטה. נתבונן:

<u>המרחב האמבאנטי</u> - המרחב שבו הבאוטה שלנו מוטבע בצורתו המקורית.

למשל בדוגמא של המעגל המרחב האמבאנטי הוא $\mathbb{R}^3$.

<u>המרחב האינטרינסי</u> - זה המרחב עליו יושב הבאוטה. בדוגמת המעגל: $\mathbb{R}^2$.

PCA

(1) PCA הוא פירוש של הורדת המימד לתת מרחב אפיני כלשהו, הקרוב ביותר לנקודות הדאטה.

מתפשים מטריצות $w^*, u^* = \underset{w,u}{\operatorname{argmin}} \sum_{i=1}^{m} \|x_i - uwx_i\|_2^2$

(2) פירוש נוסף של PCA הוא הורדת מימד שמשמרת שונות:

מתפשים: $M^* = \underset{M \text{ orthogonal}}{\operatorname{argmax}} \operatorname{Var}(XM)$, כלומר הטלה למרחב ממימד נמוך יותר k שתפיצור של הדאטה בתת המרחב יהיה מקסימלי.

השונות שאותה נרצה לשמר היא שונות ה- sample שמסומנת ע"י:

$\text{Sample Variance} = \frac{1}{m}\sum_{i=1}^{m}\|x_i - \bar{x}\|_2^2$ ל- $\bar{x}$ ממוצע $\{x_i\}_{i=1}^{m}$.

★ יש עוד 2 פירושים ל- PCA שנמשר לא נדבר.

נדבר כאן מהפירושים הכלולים ולכסוף נראה שהם שקולים: ☺

1) תת-מרחב אפיני:

כפי שהראינו בהרצאה, יש לנו פתרון סגור לבעיה זו.

כלל נעבור על ההוכחה מהשיעור ונתמקד בחלקים מסוימים בה:

<u>בפתרון ה- PCA</u>: נעבור על הדאטה ונחשב אותו (נפחית את $\bar{x}$), ונגדיר את המטריצה: $A = \sum_{i=1}^{m} x_i x_i^T$. נחשב את הע"ע והו"ע של A ונבחר מהם את w, u: עמודות u יהיו אז הוקטורים העיקריים (לפי גודל) של A, ו- $w = u^T$.

<u>למה זה כך?</u> כשנרצה להקרין את x נתפש מטריצה $d \times d$ מדרגה k שמעבירה את x לתת המרחב $S = \operatorname{Im}(uw)$, כלומר מתפשים העתקה $\varphi: X \to S$, אך נרצה שהעתקה זו תמזער את סכום המרחקים, כלומר את $\|x_i - uwx_i\|^2$ (★).

מסתבר שגוף המטריצות המקיימת זאת הן המטריצות שמייצגות את ההטלה האורתוגונלית על תת המרחב.

<u>הטלה אורתוגונלית</u> - יהי $V \subset \mathbb{R}^d$ ממימד k, ויהי בסיס אורתונורמלי ל- V: $\{v_1, \dots, v_k\}$. אז: $P = \sum_{i=1}^{k} v_i v_i^T$ מטריצת ההטלה האורתוגונלית.

<u>תכונות מטריצת ההטלה:</u>

1. P סימטרית.
2. לכל $x \in \mathbb{R}^d$, $u \in V$: $\|x - Px\|_2 \le \|x - u\|_2$ כאשר $Px \in \mathbb{R}^d$ היא ההטלה של x לתת המרחב.

⇐ הנק' הקרובה ביותר ל- $x \in \mathbb{R}^d$ על תת-המרחב היא ההטלה שלו על תת המרחב: Px, ולכן כדי למזער את (★) נצטרך למצוא מטריצת הטלה אורתוגונלית.

כלומר בעת האופטימיזציה השו מצמצר $\|x - uu^T x\|^2$

(כאן $w = u^T$ כי אנחנו בוחרים מטריצת הטלה אורתוגונלית שהיא מהצורה $\sum_{i=1}^{k} u_i u_i^T$).

לאחר פתוח, זה שקול למקסם: $\|x\|^2 - \mathrm{trace}\Big[u^T \underbrace{\Big(\sum_{i=1}^{m} x_i x_i^T\Big)}_{:=A} u\Big]$.

נזכיר:

נתון ש-A סימטרית ו-PSD (positive semidefinite): מטריצה $\mathbb{R}^{k\times k}$ סימטרית שמקיימת לכל $x \in \mathbb{R}^k$: $x^T A x \geq 0$.

נוכיח ש-A היא PSD, יהי $v \in V$. אז:

$$v^T A v = v^T\Big(\sum_{i=1}^{k} x_i x_i^T\Big) v = \sum_{i=1}^{k} v^T x_i x_i^T v =$$

$$= \sum_{i=1}^{k} \langle v, x_i\rangle \langle x_i, v\rangle \underset{\text{סימטרית מ"פ}}{=} \sum_{i=1}^{T} \langle v, x_i\rangle^2 \geq 0$$

מטוון ש-A סימטרית ו-PSD קיים לה פירוק SVD: $A = VDV^T$

ל-V אורתוגונלית ו-D אלכסונית שעל האלכסון שלה מופיעים הע"ע של A (שהם ריבועי הערכים הסינגולריים של A), ו-V מופיעים הו"ע המתאימים לע"ע שמופיעים ב-D.

נחזור לבעיה אותו אנחנו מנסים למקסם:

$$\|x\|^2 - \mathrm{trace}[u^T A u] \underset{SVD}{=} \mathrm{trace}[u^T V D V^T u] \underset{B^T := u^T V}{=}$$

$$= \mathrm{trace}[B^T D B] \underset{\text{הזזת trace}}{=} \sum_{i=1}^{d} [B^T D B]_{i,i} = \sum_{i=1}^{d} D_{i,i} \langle B_i, B_i\rangle \underset{\text{הזזת מ"פ}}{=}$$

$$= \sum_{i=1}^{d} D_{i,i} \underbrace{\sum_{j=1}^{m} B_{ij}^2}_{\leq 1} \leq \sum_{i=1}^{d} D_{i,i}$$

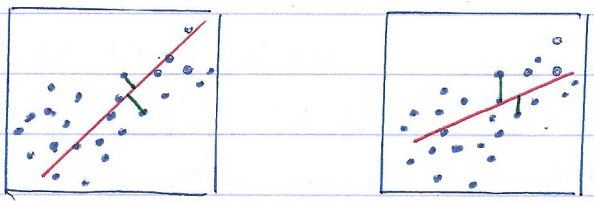
הספר UML מראים שלכל i: ≤ 1

כלומר קיבלנו ש: $\mathrm{trace}[u^T A u] \leq \sum_{i=1}^{d} D_{i,i}$

וזה הערך המקסימלי שאפשר לקבל. מהו ה-argmax?

u שמוגדרות הם אז הו"ע הגדולים כימר ב-A, כמו כן מתקיים:

$$\mathrm{trace}[u^T A u] = \sum_{i=1}^{k} D_{i,i}$$



PCA OLS

אינטרפרטציה גאומטרית:

יש אמן דאטה ו-2 פרמטרים המוגדרים לפי:

ב-2 הבעיות חיפשנו את מרחב קרוב ביותר לנקודות שלנו.

<u>איך הפרדנות יכולות שונים?</u> ההסבר נעשה בהנחות של כל מהמודלים -

- OLS: הנחנו שהתפלגות $\{y_i\}$ ושגיאת הקירוב על את מרחב של $\{x_i\}$ ואז רעש גאוסי שנמצא <u>רק בכיוון של ה-y</u>. כלומר:

$$y = \beta_0 + x^T\beta_1 + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2)$$

לכן כל נקודה יכולה לזוז <u>רק</u> בכיוון ה-y. לכן את המרחב צריך לקחת בחשבון רק את ההטלות של קואורד' ה-y שם הנקודה.

- PCA: הלמידה אינה מפוקחת, אין y, רוצים את ההטלה האורתוגונלית על את המרחב. מפני שלא יודעים נופל <u>בשווה בשווה הכיוונים</u>, כל נק' יכולה לזוז <u>לכל</u> הכיוונים באופן שווה ולכן נרצה לעשות הטלה אורתוגונלית.

הנאמרה נתון: $(x,y) \sim N(\mu, \sigma^2 I_d)$.

2) מקסום השונות:

כאן נחפש את את המרחב הקרוב ביותר לכל הנק' מימוש x, שממקסם את פיזור הנקודות על את המרחב: $M^* = \underset{M \text{ orthogonal}}{\arg\max}\ Var(XM)$

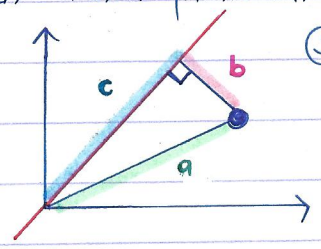
כמו בדוגמא מקודם: הטלת $\mathbb{R}^3$ על xy ממקסמת את השונות לעומת הטלה על yz.

<u>תהליך ה-PCA</u>: בראשית נמרכז את הנאמרה (נחסיר $\bar{x}$).

כעת נתחיל לבדוק תת-מרחב שונים עליהם נטיל את הנקודות ונבדוק בפיזור שלהם (נמדוד את השונות). לבסוף נבחר את את המרחב שעליו פיזור הנקודות לאחר ההטלה הוא הגבוה ביותר.

שקילות הפרושים:

נתון נקודה אחת בנאמרה, ונחפש לה <u>קרובה 3 אנלוגים</u>: מרחק הנקודה מהראשית, מרחק הנק' <u>מתת-מרחב</u> (שאת שיבו ננסה להעריך), ואת מרחק ההטלה האורתוגונלית של הנק' על את-המרחב מהראשית, קיבלנו משולש ☺



ממשפט פיתאגורס: $a^2 + b^2 = c^2$

כשנקטין את b נצטרך להגדיל את c, ולהפך.

אם $b = \sum \|x_i - UWx_i\|_2^2$

פונק' המטרה בפרוש האופני, $c = Var(XM) - 1$ פונק' המטרה בפרוש <u>שונות למקסם</u> את השונות.

<u>שונות למינימום</u> ולכן 2 הבעיות שקולות. ☺