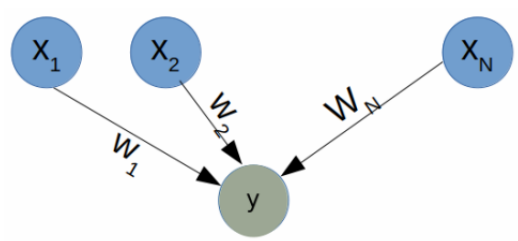
חישוביות וקוגניציה - תרגיל 6

להגשה עד: 14/12/21

<u>שימו לב</u>: שאלה 1 היא שאלה אנליטית ושאלה 2 היא שאלת תכנות

שאלה 1

נתון נוירון לינארי המקבל קלט N מימדי $\mathbf{x}$ (כמתואר בציור) . המשקולות המחברון בין הקלט לנוירון מיוצגים ע`י הוקטור $\mathbf{w}$ כלומר $y = \mathbf{w}^\top \mathbf{x}$. תוחלת הקלט היא אפס, כלומר $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, ומטריצת הקורלציה של הקלט מסומנת ע`י $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] = \mathbf{C}$. מטרת הרשת היא למקסם את השונות של הפלט, כפונקציה של $\mathbf{C}, \mathbf{w}$, תחת אילוץ על $\mathbf{w}$.



1. כתבו ביטוי לשונות הפלט (כלומר $\mathrm{Var}[y]$) כפונקציה של $\mathbf{C}$ ו־ w
2. מצאו את הווקטור $\mathbf{w}$ האופטימלי שממקסם את השונות שמצאתם בסעיף 1, בהינתן ש־ $||w||^2 = 1$. הדרכה אפשרית:

(א) הניחו כי $\{\vec{b_i}\}_{i=1}^n$ הוא הבסיס <u>האורתונורמלי</u> שבו C היא מטריצה אלכסונית, כלומר, הבסיס שנוצר על ידי הווקטורים העצמיים של C (זכרו כי C היא מטריצה לכסינה

וסימטרית, כלומר, קיים בסיס כזה). רשמו ביטוי כללי ל־w ול־Cw באמצעות רכיבי הבסיס הנ"ל.

(ב) רשמו את הביטוי לשונות שקיבלתם בסעיף 1 באמצעות (א)

(ג) השתמשו באילוץ על הגודל של w כדי להסיק מיהם המקדמים האופטימליים של w בבסיס $\{\vec{b}_i\}_{i=1}^n$.

3. נניח שהקלט הוא דו־מימדי ומתפלג באופן הבא: $x_1 \sim \mathcal{N}(0,1)$ ו $x_2 \sim \mathcal{N}(0,4)$, ו x_1, x_2 הם בלתי תלויים

(א) מהי מטריצת הקורלציה של הקלט?

(ב) מצאו את w האופטימלי במקרה זה תחת האילוץ של סעיף 2

(ג) ציירו במישור באופן סכמטי איך ייראה מדגם אופייני של נקודות קלט, וכן ציירו את הכיוון של w

4. נניח כי הקלט הוא חד־מימדי. מהו הערך של w שממקסם את השונות בפלט, אם אין אילוץ על ערכו של w?

5. השוו את התוצאה שקיבלתם בסעיף 2 לתוצאה שראיתם בכיתה עבור הפתרון האופטימלי להורדת מימד בעזרת נוירון לינארי (PCA). התייחסו לפונקציית המטרה (מטרת האופטימיזציה) ולתוצאה שהתקבלה.

שאלה 2

בתרגיל הבא תשתמשו בשיטת PCA כדי לנתח נתונים שהתקבלו משאלון קצר אותו תעבירו בעצמכם. לשם כך מצורף שאלון של 6 שאלות שכל אחת מהן עוסקת בהיבט אחר של שביעות רצון מהחיים.

חלק א' - העברת השאלון והעלאת התוצאות

1. כתבו את השאלון על נייר, או צרו עותק של ה־google form שמופיע באתר.

2. העבירו את השאלון ל־10 אנשים קרובים והכניסו את התוצאות שלהם לטבלה המצורפת/ הורידו את התוצאות מהטופס של גוגל לתוך קובץ אקסל. **שימו לב** שאינכם יכולים להשתמש בנתוניו של מי שכבר מילא את השאלון עבור סטודנטית אחרת.

3. את הטבלה יש **להעלות בנפרד לתיקיית ההגשה quizes** כקובץ אקסל ששמו בפורמט הבא, **כאשר הספרות הן ארבע הספרות האחרונות בתעודת הזהות שלכם**: quiz_1234.xlsx

חלק ב' - הרצת PCA

קטע הקוד המצורף בקובץ ex6_PCA קורא את הטבלה והופך אותה למטריצה [לשם כך, עליכם להכניס במקום המתאים את שם הקובץ שלכם, ואם לא שמרתם אותו באותה התיקייה בה רץ הקוד, אז גם את הנתיב המתאים]. תוכלו להשתמש בו על מנת לייצר מטריצה שבה כל שורה מייצגת משיב וכל עמודה מייצגת שאלה.

1. רשמו את שם הקובץ שלכם בקובץ הקוד, ואם צריך, אז גם את הנתיב (אפשר לראות בקוד דוגמה).
2. נרמלו את העמודות של המטריצה, כדי לייצר מטריצה שבה הממוצע של כל שאלה על כל הדוגמאות יהיה 0 (אם אתן משתמשות ב-np.mean וודאו שאתן משתמשות בפרמטר $axis = 0$).
3. חשבו את מטריצת השונות המשותפת של הדוגמאות.
4. השתמשו בפונקצייה $np.linalg.eig$ כדי למצוא בסיס של ווקטורים עצמיים של המטריצה מהסעיף הקודם, וערכים עצמיים המתאימים להם. שימו לב שהפונקציה מחזירה שני אובייקטים, רשימה של ערכים עצמיים, ומטריצה שבה העמודה ה-i היא ווקטור עצמי המתאים לערך העצמי ה-i.
https://numpy.org/doc/stable/reference/generated/numpy.linalg.eig.html

חלק ג' - הצגת הנתונים והסקת מסקנות

1. הוציאו גרף בו הציר האופקי הוא מספר הרכיבים וציר ה-y הוא אחוז השונות המוסברת באמצעות אותם רכיבים.

- כמה ווקטורים נדרשו לכם על מנת לבטא 50% מהשונות, 75%, 95%? מה המשמעות של כך בנוגע לשאלון שהעברתן?

2. רשמו מה ההרכב של הוקטור העצמי השייך לרכיב העצמי הגדול ביותר, כלומר, מהו הצירוף הליניארי של הצירים המקוריים (כל ציר הוא שאלה), ומהו מבטא.

- האם תוכלו להסביר מדוע אלה הרכיבים שיצאו?

- אילו שאלות מהוות חלק משמעותי בצירים החדשים? מדוע, לדעתכן?

3. הציגו את הדאטא שלכן בתצוגה דו־מימדית בה ציר אחד הוא הרכיב החזק ביותר וציר שני הוא הרכיב החלש ביותר.

- הסבירו כיצד ניתן לראות בגרף את ההבדל בין הרכיבים. (תוכלו לנסות לקבע טווח זהה עבור שני הצירים ולראות מה קורה)

- מה המשמעות של הרכיב הכי חלש?

4. האם ניתן להסיק משהו על המבנה של השאלון שהעברתן באמצעות תוצאות הניתוח שלכן? (שימו לב שלכל אחת מכן התשובה על השאלה הזו יכולה להיות שונה בגלל תוצאות שונות)

- אם הייתן צריכות לייצר שאלון חדש שבוחן את אותו הנושא ובו שאלה אחת (ישנה או חדשה), מה היא הייתה?

הערות:

1. על מנת לייצר את המטריצה C וודאו שאתם מבצעים כפל מטריצות בצורה הנכונה - אם העמודות הן שאלות, אז השחלוף (טרנספוז) צריך להופיע באיבר הראשון שבכפל.

אם יש שש שאלות, התוצר הסופי צריך להיות מטריצה 6X6. אם כפלתם לא נכון, המטריצה הזו תהיה גדולה מדי, וגם הערכים העצמיים שתחשבו עלולים לצאת מרוכבים.

2. הערכים העצמיים שהפונקציה מחזירה לא יוצאים בהכרח מסודרים לפי הגודל, בשביל הנוחות שלכם כדאי למיין. (דרך אחת היא למיין ראשית את האינדקסים ואז לבנות מטריצה חדשה לפי האינדקסים הללו)

3. את אחוז השונות המוסברת ניתן לחשב בתור החלק היחסי של הערך העצמי מתוך סכום הערכים העצמיים, כפי שמופיע במסמך המצורף.