

מבוא לבינה מלאכותית – תרגול 8

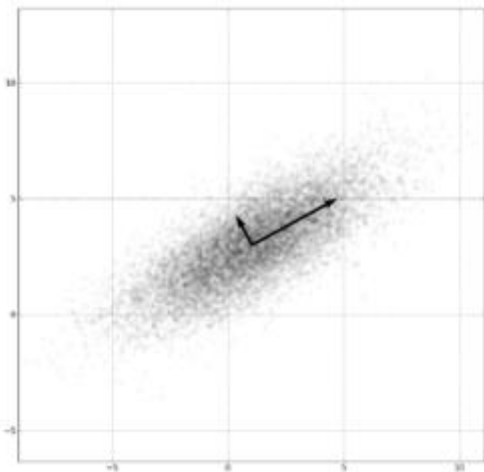
נושאים:

- PCA
- ICA
- אם יהיה זמן – MDS

רקע:

ככל שיש יותר פרמטרים להעריך במודל, כך המשימה לדייק בהערכה הזו דורשת יותר – חישוב מסובך או ארוך יותר, יותר דאטא ועוד. זו בעיה שלפעמים לא ניתן לפתור (במקרים רבים אין דאטא נוסף שאפשר להכניס). במקום זאת, ישנם אלגוריתמים שמטרתם היא להקטין את מספר המשתנים והפרמטרים, במטרה מצד אחד להקל על התכנסות והערכה טובה ומצד שני לא לאבד מידע ככל האפשר.

:Principal Component Analysis (PCA)



זוהי שיטה שמניחה על הדאטא את הדבר הבא:

ישנה הצגה של הדאטא ע"י וקטורים הלקוחים מהתפלגות נורמלית (אולי רב ממדית), עם תוחלת 0 ומטריצת Covariance אלכסונית. במקרה כזה, פיצ'רים בלתי מתואמים הם בהכרח בלתי תלויים.

נסתכל על 3 גישות להסתכלות על מה שמתבצע באלגוריתם PCA:

1. שונות מקסימלית – ככל שיש יותר שונות בדאטא, כך נוכל להבחין בדברים, להפריד לאשכולות וכד'. מהסתכלות הפוכה, כאשר אין שונות, הדאטא מרוכז במקום אחד ולא מפוזר, כך שלא ניתן להבחין כמו שצריך במה שרוצים.
אלגוריתם PCA ניתן לניסוח כבעיה שממקסמת את השונות הנותרת לאחר הטרנספורמציה.
2. שגיאת השחזור מינימלית – נניח שאנחנו מתארים את המרחב בתור צירוף ליניארי של מספר קבוע של וקטורים. וקטור הפיצ'רים שאחרי הטרנספורמציה יהיה הקרוב ביותר לווקטור המקורי, מבין כל הווקטורים שניתן לכתוב כצירוף של מספר וקטורים קבוע זה.
3. הסרת קורלציות – חלק מהפיצ'רים יכולים להיות מתואמים אלה עם אלה, ולכן במרחב המקורי ישנם פיצ'רים מיותרים או מעורבבים אלה עם אלה. המטרה המקורית של PCA

(היסטורית) הייתה לבצע טרנספורמציה למרחב הפיצ'רים כך שהפיצ'רים החדשים לא יהיו מתואמים (ובמקרה הגאומטרי – גם בלתי תלויים) אלה עם אלה.

כל הבעיות האלה נפתרות בעזרת אלגוריתם אחד, PCA:

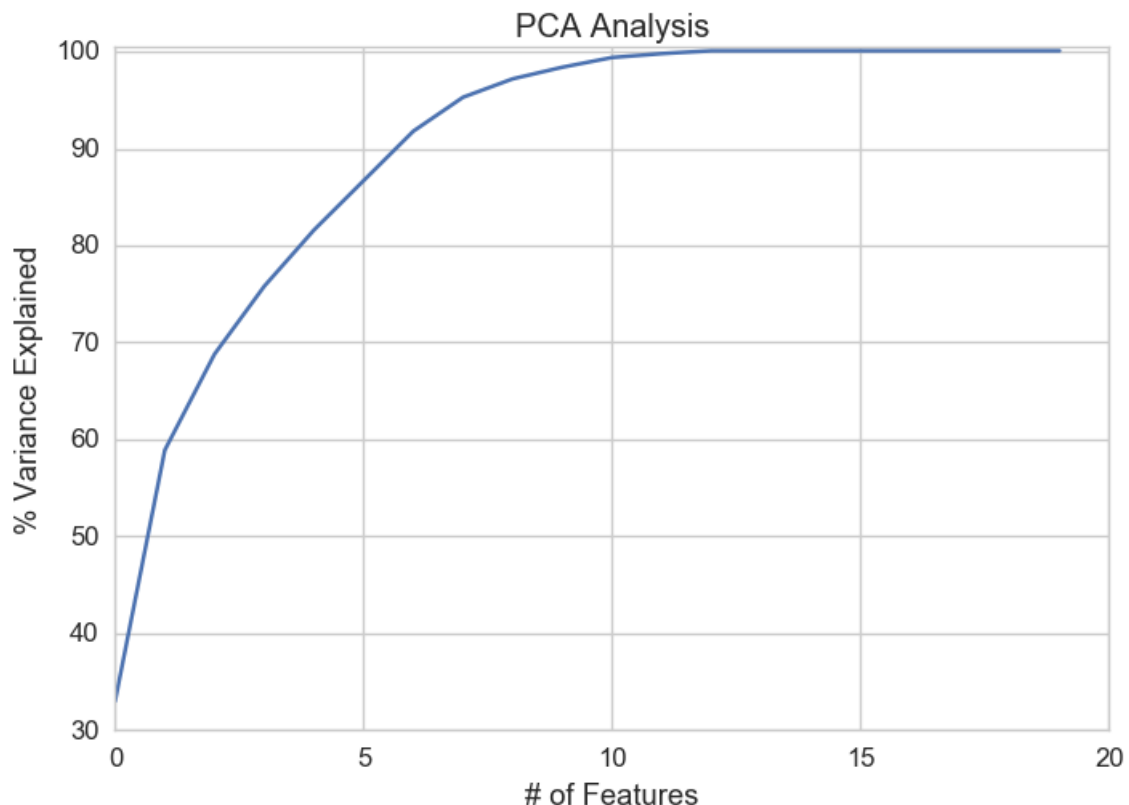
אלגוריתם:

בהינתן אוסף נקודות $\{x_i\} \subseteq X$ כאשר X מרחב הפיצ'רים, מתבצעים:

1. "הלבנה" של הדאטא – חיסור הממוצע מכל נקודה, כדי שהממוצע של כל הפיצ'רים יהיה 0.
2. בניית מטריצת Covariance: $C = \sum_i \vec{x}_i \cdot \vec{x}_i^T$. שימו לב כי $\vec{x}_i \cdot \vec{x}_i^T$ היא מטריצה ($\vec{x}_i$ וקטור עמודה) חיובית, וכך גם C .
3. חישוב ערכים עצמיים גדולים ביותר ווקטורים עצמיים מתאימים להם (לגבי בחירת כמות הווקטורים, ראה בהמשך).
4. הפיצ'רים החדשים יהיו ההטלה של הפיצ'רים הישנים על מרחב הווקטורים העצמיים.

מה מקבלים ואיך בוחרים את הממד החדש:

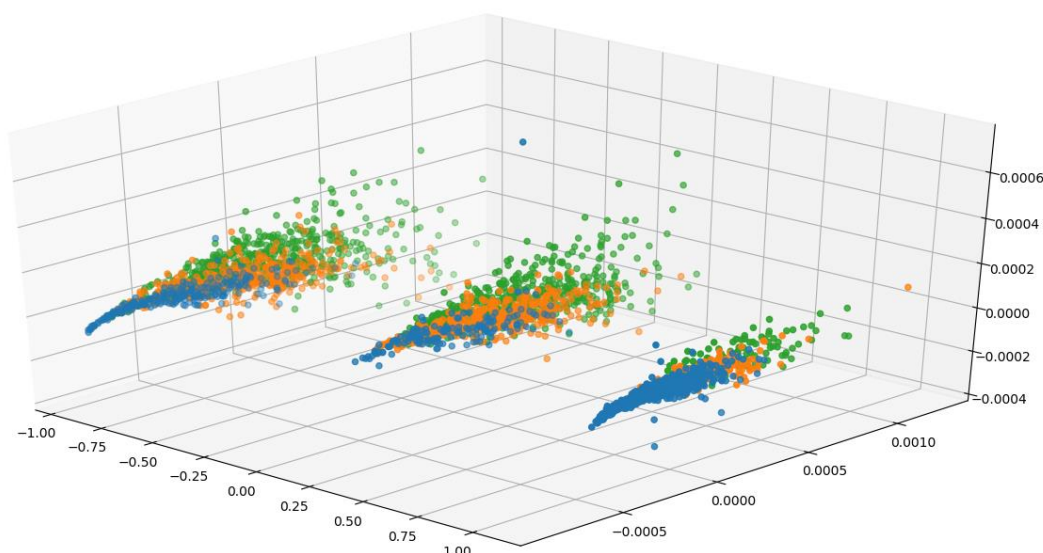
גרף טיפוסי של השונות המוסברת כתלות בערך העצמי הקטן ביותר שניקח:



מכאן רואים שמספיק להעתיק את הדאטא למרחב בממד לא גבוה מאוד כדי לקבל שונות מספיקה. אין אפילו כלל אצבע לגבי כמה ממדים לבחור, אבל יותר ממדים יסבירו יותר מהשונות הכוללת ומצד שני בעיות שנרצה לפתור תהיינה פשוטות יותר ככל שהממד יהיה נמוך יותר.

דוגמה:

דאטאסט של צדפות, עם 8 פיצ'רים שונים (כן, זה מהקורס בלמידת מכונה. את הפיצ'ר של מין הצדפה החלפתי ב-0, 1, 2). לאחר נרמול מסוים ואז לאחר PCA שהוריד את הממד ל-3, הדאטא נראה כך:



הווקטורים העצמיים (מעוגלים ל-5 ספרות אחרי הנקודה):

	0	1	2	3	4	5	6	7
0	1.0	-4e-05	-4e-05	-4e-05	-0.0001	-0.0001	-0.0001	-9e-05
1	0.00021	0.17332	0.18291	0.19901	0.47812	0.48418	0.47483	0.45651
2	0.0	0.04952	0.08189	0.36061	-0.03983	-0.58589	-0.22048	0.68362

לא כתבתי כאן אבל הווקטור העצמי הראשון מסביר כ-99% מהשונות (הסיבה היא כי לא נרמלתי את הרכיב הזה והוא נשאר בגדלים שהיה, בשונה מפיצ'רים אחרים שקטנו משמעותית). נשים לב: ככל שהרכיב גדול יותר, כך הפיצ'ר משמעותי יותר לייצוג החדש.

הערה – באלגוריתמים שמיישמים PCA, לעתים משתמשים בפירוק לערכים סינגולריים (SVD) של מטריצת הדגימות במקום במטריצת ה-Covariance. בשיטה זו ניתן לחשב בדרך אחרת את הווקטורים העצמיים העיקריים. לרוב שיטה זו מועדפת על פני לכסון.

הערה – קיימת גרסת PCA עם kernel, לא ניכנס למצב זה מכיוון שהוא דומה ברעיונו ל-kernel SVM.

Independent Component Analysis (ICA)

זוהי שיטה אחרת להורדת ממדים, בהנחות שונות על המודל:
במרחב החדש הדאטא יהיה מיוצג באמצעות וקטורים שלקוחים מהתפלגות לא נורמלית, עם תוחלת 0 וב"ת אלה באלה.

שתי הסתכלויות על הפירוק (מן הסתם אם נכנסים לפרטים יש בהן הבדלים קטנים, אולם הרעיון באופן אידיאלי [ולא ניתן למימוש כאלגוריתם] נותן תוצאות זהות):

1. אם כותבים $\vec{x} = A\vec{y} + \vec{\epsilon}$ כאשר $\vec{x}$ הדאטא המקורי, $\vec{y}$ ייצוג הדאטא במרחב החדש, A ההעתקה ו- $\vec{\epsilon}$ רעש, שחזור שגוי של $\vec{y}$ מ- $\vec{x}$ (שיוביל לערבוב של הרכיבים הנכונים והלא-נורמליים) ייתן בסיסי גבוה וקטורים שלקוחים מהתפלגות נורמלית (לפי משפט הגבול המרכזי).
לכן, אלגוריתם ICA מנסה לשחזר את $\vec{y}$ בצורה כזו שההתפלגות תהיה רחוקה ככל האפשר מהתפלגות נורמלית.
2. בעזרת תורת האינפורמציה, ההתפלגות במערכת בעלת מומנט שני (כלומר עם שונות סופית) שעבורה האנטרופיה מקסימלית היא התפלגות נורמלית.
לכן, ICA מנסה למזער את האנטרופיה.

דוגמה למצב בו הדאטא מאוד לא נורמלי ולכן המרחב המועתק באמצעות PCA לא מוצלח כמו המרחב המועתק באמצעות ICA:



:Multi-Dimensional Scaling (MDS)

השיטות שמופיעות כאן נועדו לפתור את הבעיה הבאה:

נתון דאטא בצורת מטריצת מרחקים בין הנקודות, אבל לא ייצוג שלהן כאוסף וקטורים במרחב. המטרה היא לייצג אותם ככאלה, כאשר שומרים על המרחקים ככל האפשר או בהתאם למה שנדרוש. בכל מקרה, אלה שיטות שדורשות זמן חישוב די ארוך.

:Classical MDS (cMDS)

השיטה ידועה גם כ-Principal Coordinate Analysis (PCoA).

קלאסי – מחפשים ייצוג במרחב של הדאטא, כך שההפרשים בין המרחקים של נקודות במרחב למרחקים במטריצת המרחקים יהיו הקטנים ביותר האפשריים.

דרך הפתרון כוללת ייצוג של מטריצה שאמורה להיות מטריצת $X^T X$ כאשר X המרחב המיועד, ואז SVD למטריצה: $B = UDV$ כאשר U ו- V מטריצות אוניטריות (במקרה הממשי, $UU^T = I$ ו- D מטריצה אלכסונית).

$$X = D^{\frac{1}{2}} U^T, \text{ בשלב הבא,}$$

הערה – ביצוע שלבים אלה בצורה פשוטה ייתנו ל- X מימד פיצ'רים ששווה לכמות הנקודות בדאטא, שזה הרבה. כדי למנוע זאת, משתמשים רק בערכים העצמיים הגדולים ביותר של D . אינטואיטיבית, זה אומר להיפטר מפיצ'רים שהערכים שלהם קרובים מאוד ל-0.

הבעיה באלגוריתם הזה היא שהוא לוקח זמן – SVD עולה $O(n^3)$ כאשר n כמות הנקודות.

:Metric MDS (mMDS)

הכללה מסוימת של cMDS, כך שהמרחקים בין הנקודות יהיו לא בהכרח לפי מטריקה אוקלידית (כלומר הייצוג החדש יתאים למטריצת מרחקים מועתקת למרחב חדש, או העתקה של הייצוג החדש מתאימה למטריצת המרחקים המקורית).

מה שעושים הוא כותבים פונקציית loss (מכונה Stress) שאותה ממזערים על מרחב הנקודות, באמצעות שיטות כמו גרדיאנט.

:Non-Metric MDS (nMDS)

כאן דורשים תנאי חלש יותר מקודם –

אם $d_{ij} < d_{kl}$ במטריצה המקורית, אז $d(x_i, x_j) \leq d(x_k, x_l)$ במרחב החדש לכמה שיותר נקודות.
דרך פתרון:

1. בחירת נקודות אקראיות מהתפלגות נורמלית בממד שנבחר מראש, עם תוחלת 0 ומטריצת Covariance אקראית.

2. לכל נקודה בדאטא תותאם נקודה מאוסף הנקודות הנ"ל.

3. לכל שתי נקודות, מחליפים ביניהן אם מספר הזוגות שעבורם התנאי מופר יקטן.

