למידה סטטיסטית וניתוח נתונים - תרגיל 2 (PCA) נק׳: 10
דדליין להגשה: יום שני, 17.3.23 ב-00:00

1. **ביצוע PCA hands-on**: בשאלה שלפנינו נלמד להעריך את יכולותיו של אלגוריתם זה בתיאור מידע באופן ״חסכוני״ ו״חכם״, וכיצד הוא עשוי להועיל לכיווץ מידע. בקובץ התרגיל מצורפת תמונת jpg עליה נעבוד.

א. ראשית, השתמשו בספריה PIL על מנת לפתוח את הקובץ בפייתון. מהם מימדי התמונה, ומה כל מימד מייצג?

ב. סכמו את התמונה על מימדי הצבע ובצעו נורמליזציה (הסרת ממוצע, חילוק בסטיית התקן). האם הנורמליזציה משנה את טיב התמונה?

ג. בצעו פירוק SVD לתמונה המנורמלת בעזרת הספריה. מהו מס׳ הערכים הסינגולריים עבור התמונה שלנו, ולמה? הציגו plot של הערכים הסינגולריים, ועל-גביו plot מקווקו של כל הערכים הסינגולריים שערכם גדול מ0.01% של הערך הסינגולרי הגדול ביותר שקיים, ורשמו כמה כאלה ישנם.

ד. כעת נרצה לנסות לשחזר את התמונה עם מספר שונה של Principal Components. ראשית, שמרו את גודל התמונה המנורמלת והמקורית (בבייטים) כמשתנים כלשהם. כעת הניחו חמש כמויות שונות של PCs בעזרתן נרצה לבצע את שחזור התמונה: 12, 24, 48, 96 והמספר שחישבתם בסעיף ג׳. שחזרו את התמונה בעזרת שלושת הקומפוננטות של הSVD שיצרתם בסעיף ג׳ עם מספרי ה-PCs הנ״ל (בעזרת מעט אלגברה לינארית חשובה!), והציגו את חמשת השחזורים בפלוט. האם שמים לב להבדל? לבסוף, חשבו את גודל השחזור האחרון (עם מס׳ הPCs מסעיף ג׳), והשוו עם גדלי התמונות שחישבתם בתחילת הסעיף. האם הכיווץ חסך בהרבה זכרון, ביחס להרעה באיכות התמונה?

2. בשאלה זו נעסוק בחקר תולדות חיבור התנ״ך וכיצד ניתן לזהות אסכולות סופרים שונות. בתרגיל מצורפים מס׳ קבצים המתייחסים לחלוקה היפותטית של ספר שמות לשני מקורות שחוברו בידי קבוצות שונות של סופרים: האסכולה הכהנית והלא-כהנית (דהיינו, כל אסכולה אחרת). בשאלה זו נבחן את טיב ההפרדה הלשונית בין שתי קבוצות אלו, וננסה לאפיין את מידת השוני ביניהן.
הקובץ exodus_pickled המצורף לתרגיל הינו מטריצה (ספרסית) המייצגת קידוד מתמטי של הטקסט ספר שמות: **מרחב השורות** מייצג את הפסוקים (כל שורה הינה וקטור המייצג פסוק מסוים), ו**מרחב העמודות** מייצג את מרחב המשתנים המסבירים שלנו, במקרה זה - כל ערך בוקטור מייצג שלשת מילים עוקבות (מילים שמופיעול אחת אחרי השניה ברצף) הקיימת בספר, ולכן אורך כל וקטור הוא כמס׳ כל שלשות המילים הייחודיות בספר שמות. הערך המספרי של כל ערך בוקטור מייצג (פחות או יותר, לא ניכנס לזה) אם השלשה הספציפית הזו קיימת בפסוק הספציפי הזה או לא. למשל: העמודה ה-1087 מייצגת את השלשה ״אמה, ו, חצי״, ולכן בוקטור של פסוק בו השלשה קיימת - הערך המספרי יהיה חיובי, ובכל פסוק בו איננה - הערך של העמודה ה-1087 יהיה 0 (לא ניכנס למה הציינון לא בינארי, אבל זה הרעיון). הקובץ המצורף features.txt מציין לאיזה שלשת מילים כל עמודה משוייכת. כמו כן, הקובץ המצורף labels.txt מכיל שיוך היפותטי של כל פסוק לאחת משתי קבוצות: 0, או 1 (דהיינו - כוהני או לא-כוהני).

א. טענו את הקבצים. את exodus_pickled ניתן לפתוח בעזרת הספריה pickle שמקודדת נתוני פייתון בצורה יעילה. את קבצי הטקסט ניתן לפתוח בעזרת numpy.loadtxt (במקרה של features, יש לציין שה-delimiter הוא '\t' ו-dtype הוא 'str'). מהו המימד במטריצה exodus_pickled המתייחס לוקטורים features, labels? הסבירו.

ב. האם אנו יכולים לעשות plot של הפסוקים במתכונתם הנוכחית ולבדוק האם השיוך לקבוצות מבטא הבדל אמיתי ביניהן? מדוע?

ג. כעת נרצה לבצע הורדת מימדת לנתונים שלנו. בצעו פירוק SVD למטריצה exodus_pickled. בחנו את המימדים של שלושת הרכיבים. לפי המימדים בלבד, קבעו איזה רכיב מבטא את וקטורי החשיבות של כל שלשת מילים לכל PC.

ד. **עבור ה-PC הראשון**, מצאו את המקומות בוקטור החשיבות (loadings) של שלשות המילים את עשרת המיקומים בעלי הערכים (המוחלטים!) הגבוהים ביותר. הדפיסו את עשר שלשות המילים במיקומים הללו. מה משותף לרובן, ומה מרמזת לנו החשיבות של עשר שלשות מילים אלו ל-PC הראשון?

ה. **אינטרפרטציה של התוצאות**: השתמשו במנוע החיפוש בתנ"ך בכדי לנסות לזהות את מקור שלשות המילים שמצאתם בסעיף ד' (למשל - גוש רציף של טקסט) בספר שמות. האם המקור (באופן כללי) זהה? אם כן, מהי משמעות החשיבות של שלשות מילים אלו ל-PC הזה? מה ניתן להסיק מכך על הטקסט?

ו. **הטלה על PCs**: כעת נרצה להטיל את הנתונים שלנו על שני ה-PCs הראשונים ולהציגם באופן גרפי, באופן הבא:

- ראשית, ציינו כיצד מטילים כל פסוק (שורה) על שני הPCs הראשונים? מה המימד של מה שמתקבל בסוף, עבור כל פסק? (רמז: הטלה משמעו מכפלה סקלרית. מהו הרכיב משלושת הרכיבים של הSVD עימו צריך לחשב מכפלה סקלרית זו? תחשבו על המימדים של הרכיבים, ועל המימד של הפסוקים).
- בצעו את ההטלה של כל הפסוקים על שני הPCs הראשונים. הציגו plot דו-מימדי של הנתונים המוטלים, כאשר ציר ה-x הוא ציר ההטלה על ה-PC הראשון, וציר ה-y על הPC השני. כעת יש להשתמש בוקטור ה-labels שצורף לתרגיל. הציגו את כל הפסוקים המשויכים לקבוצה 0 בצבע אחד, ואלו המשויכים לקבוצה-2 בצבע אחר, כאשר כל פסוק הוא נקודה ב-plot הדו-מימדי.
- האם, לדעתכם, ההפרדה ההיפוטתית לקבוצה 0 או 1 בעלת מתאם חיובי עם שני צירי השונות הראשונים? מהי המשמעות של התוצאה הגרפית של ה-PCA ביחס לנסיון לזהות את ההפרדה בין הקבוצות 0 ו-1 ע"י ספירה של מופעים של שלשות מילים בפסוקים?

ז. **סכמו את השאלה והממצאים**:
* מהי השאלה המחקרית
* מהם הנתונים עימם אנו עובדים וכיצד כל אחד מהם רלוונטי לשאלה המחקרית (features, labels, dataset..)
* מהן התרומות של פירוק SVD שעזרו לבחון את השאלה המדעית, ואיזו אינפורמציה חדשה סיפקנו בעזרתו?
* האם אתם מכירים כלים סטטיסטיים נוספים שהייתם מעריכים שעשויים לתרום תרומה נוספת לחקר הענין?