למידת מכונה

מבחן סופי

מועד א' סמסטר ב

תשפ"א

מרצה: פרופסור אילן שמשוני

מתרגל: גילי כהן ואריה סטוליארסקי

משך הבחינה שעתיים וחצי.
ניתן להשתמש בכל חומר עזר.
יש לענות על כל **שבע** השאלות **במחברת המצורפת.**
שים לב כי בידך 5 דפים כולל דף זה.
הבחינה מיועדת לגברים ונשים כאחד ומנוסחת בלשון זכר מטעמי נוחות בלבד.
נא לכתוב בצורה מסודרת וברורה במחברת.

בהצלחה!

שאלה מס'	ניקוד מקסימאלי	ניקוד
1. בניית מודל	25	
2. מטריצת בלבול	9	
3. תיאוריה	10	
4. SOM ו t-SNE	16	
5. SVM	16	
6. אלגוריתמים אדפטיבים	10	
7. עצי החלטה	24	

1. תהליך בניית מודל (25 נק')

נתון ה-dataset הבא, באמצעותו נרצה לבנות מערכת רוקחות אוטומטית, המנפיקה תרופות לפי תיק רפואי:

ID	Age	Sex	Blood pressure	Cholesterol	Na_to_K	Drug
670652128	23	F	HIGH	HIGH	25.355	Nexemide
492859899	47	M	LOW	HIGH	13.093	Doxicon
169374018	47	M	LOW	HIGH	10.114	Doxicon
814085227	28	F	NORMAL	HIGH	7.798	Vitavol
874498623	61	F	LOW	HIGH	18.043	Nexemide
248391664	22	F	NORMAL	HIGH	8.607	Vitavol
748533993	49	F	NORMAL	HIGH	---	Nexemide
029759877	41	M	LOW	HIGH	11.037	Doxicon
186361419	60	M	NORMAL	HIGH	15.171	Nexemide
677928907	43	M	LOW	NORMAL	19.368	Nexemide
454853495	47	F	LOW	HIGH	---	Doxicon
958556060	34	F	HIGH	NORMAL	19.199	Nexemide
031123464	43	M	LOW	HIGH	15.376	Nexemide
550332065	74	F	LOW	HIGH	20.942	Nexemide
290757012	50	F	NORMAL	HIGH	---	Vitavol
299687163	16	F	HIGH	NORMAL	15.516	Nexemide

הסבר על הפיצ'רים והלייבל:

- ID: ערך ייחודי, תעודת הזהות של המטופל.
- Age: גילו של המטופל.
- Sex: מגדרו של המטופל.
- Blood pressure: לחץ הדם של המטופל, מתחלק ל-3 קטגוריות: LOW, NORMAL, HIGH.
- Cholesterol: רמת הכולסטרול בדם. מתחלקת ל-2 קטגוריות: NORMAL ו-HIGH.
- Na_to_K: היחס בין רמת הנתרן לאשלגן בדם.
- Drug: תרופה המתאימה למטופל, תהיה אחת מבין ה-3 הבאות: Nexemide, Doxicon, Vitavol. זהו ה-label.

ה-dataset הנ"ל מכיל כ-10,000 תצפיות.

ענו על השאלות הבאות, המתייחסות לתהליך בניית מודל מתאים.

א. אילו תהליכים הייתם מבצעים על הנתונים בשלב ה-preprocessing, פרטו על לפחות 3 תהליכים, הסבירו עליהם בקצרה וכיצד הייתם מיישמים אותם עבור ה-dataset הזה. התייחסו לתהליכי transformation, imputation ו-correlation.
ב. התבקשתם להפעיל אלגוריתם SVM על ה-dataset. הציגו תהליך preprocessing אחד נוסף שהייתם מבצעים ומדוע.
ג. כעת התבקשתם להפעיל גם KNN, האם יהיו שינויים נוספים שצריך לבצע? הסבירו מדוע.
ד. במהלך אימון המודל, קיבלתם אחוזי דיוק של 0.97, אולם לאחר הרצת האלגוריתם על סט המבחן התקבלו תוצאות דיוק של 0.48 בלבד. ציינו 2 סיבות לתוצאות אלו והסבירו כיצד הייתם מתמודדים איתן.
ה. לאחר תהליך בירוקרטי רב, הוחלט לבנות מודל אשר יעבוד על תיק רפואי נרחב יותר, אשר מכיל כ-5,500 מאפיינים רפואיים שונים. מהן הבעיות העיקריות העלולות להתרחש משינוי זה וכיצד תתמודדו איתן?

2. Confusion Matrix(9 נקודות)

נתונה confusion matrix הבאה:

		Actual		
		Class 0	Class 1	Class 2
Predicted	Class 0	10	4	3
	Class 1	6	2	5
	Class 2	3	7	7

א. חשבו את מדד ה-specificity עבור class 0.
ב. חשבו את מדד ה-f1 score עבור class 2.
ג. הסבירו את המושגים "false positive" ו –"false negative".

3. תאוריה (10 נק')

מדוע אנו זקוקים לסט ולידציה וסט מבחן במהלך בניית מודל?

4. t-SNE ו SOM (16 נק')

הקלט לאלגוריתמים t-SNE ו SOM הוא n נקודות X_i ממימד גבוה.

א. מה הפלט של אלגוריתם t-SNE ואלגוריתם SOM ?
ב. בהינתן נקודת קלט X_i היכן בתמונת הפלט הדו-מימדית יופיע המייצג שלה?
ג. מה ההבדל בין אלגוריתמים אלה ל SVD מבחינה חישובית ומבחינת מטרה?
ד. האם נוהגים להשתמש בתוצאות של SVD לצורכי ויזואליזציה? אם כן איך עושים זאת?

5. SVM (16 נק')

נניח שניתן לתת כפרמטר לפונקצית ה SVM מערך באורך n בשם weights שנותן משקל לכל נקודה.

א) תנו שני שימושים לאופציה כזאת.
ב) איך תעדכנו את פונקצית המחיר של SVM על מנת שהאלגוריתם ישתמש במערך הזה?
ג) נניח שניתן לתת כפרמטר מערך בשם class_weight שנותן משקל לכל מחלקה.
תנו שימוש לאופציה הזאת.
ד) איך תעדכנו את פונקצית המחיר של SVM על מנת שהאלגוריתם ישתמש במערך הזה?

פונקצית המחיר:

$$\tfrac{1}{2} \langle \mathbf{w}^{T}, \mathbf{w} \rangle + C\Sigma\xi_i$$

Such that:

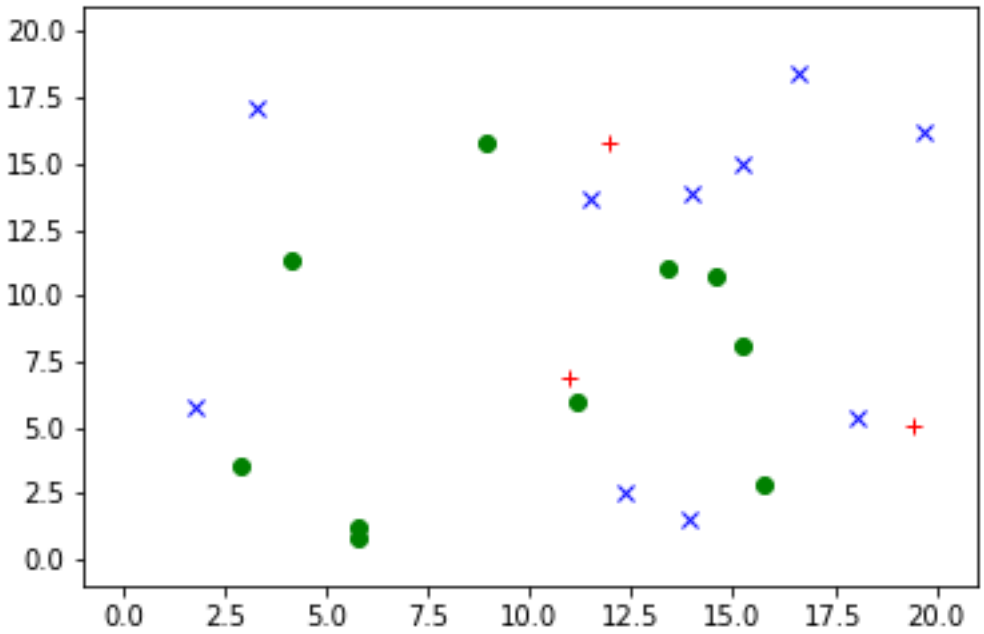
$$y_i \left(\langle \mathbf{w}^{T}, \mathbf{x_i} \rangle + b\right) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \text{ for all } i$$

6. אלגוריתמים אדפטיבים (10 נק')

א) למדנו על הרחבה של אלגוריתם mean shift שהוא אדפטיבי ומותאם לנתונים ממימד גבוה.
איך הוא הותאם להיות אדפטיבי לצפיפות הנתונים ואיך הוא פותח כדי לטפל בנתונים ממימדים גבוהים.
ב) יש טענה שגם אלגוריתם t-SNE הוא אדפטיבי לצפיפות הנתונים המשתנה בנתונים. איך זה נעשה?

7. עצי החלטה (24 נקודות)

נתונים הנתונים הבאים. חלק שייכים למחלקה X השאר למחלקה O. יש גם שלוש נקודות שמסומנות ב + שהן נקודות לסיווג.



א) בנו עץ החלטה בעומק 3 (עד 7 צמתים בעץ). נסו שהפתרון יהיה קרוב לעץ שהיה מחושב על ידי אלגוריתם (לא חייבים לנסות את כל האפשרויות).

ב) לכל צומת חשבו את ערכי הג'יני ואת אחוז הנקודות בשתי המחלקות.

ג) סווגו את שלושת הנקודות לסיווג. לכל נקודה חשבו את המחלקה ואת ההסתברות להיות באותה מחלקה.