

למידת מכונה

מבחן סופי

מועד ב' סמסטר ב

תשפ"א

מרצה: פרופסור אילן שמשוני

מתרגל: גילי כהן ואריה סטוליאנסקי

משך הבחינה שלוש שעות.

ניתן להשתמש בכל חומר עזר.

יש לענות על כל **שבע** השאלות **במחברת המצורפת**.

שים לב כי בידך 4 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. תיאוריה	10	
2. Confusion Matrix	15	
3. תהליך בניית מודל	25	
4. Mean Shift	20	
5. Viola and Jones	15	
6. Divisive Clustering	15	
7. Neural Nets	20	

שאלה 1 – תיאוריה (10 נק'):

מה דומה ומה שונה בין GridSearchCV ל-RandomizedSearchCV ומתי נרצה להשתמש בכל אחד מהם?

שאלה 2 – Confusion Matrix (15 נק'):

לאור מספר הזיופים הרב המתרחש במדינה בשנים האחרונות, חברה לביטוח מעוניינת לבנות מערכת לזיהוי זיופים לתביעות ביטוח. לשם כך נבנו עבור החברה מספר מודלי למידת מכונה.

נתונות מטריצות הבלבול הבאות השייכות למודלים 'א' ו'ב' בהתאמה:

		Actual	
		תביעה אמיתית	תביעה מזויפת
Predicted	תביעה אמיתית	70	10
	תביעה מזויפת	130	190

		Actual	
		תביעה אמיתית	תביעה מזויפת
Predicted	תביעה אמיתית	170	110
	תביעה מזויפת	30	90

- איזה מה-labels יותר חשוב לחברת הביטוח בהתאם לסיפור. יש לציין מדוע.
- בהתאם לתשובתכם בסעיף א' - מהו הממד שנשתמש בו ע"מ להשוות את המודלים השונים?
- לפי הממד שבחרתם, איזה מבין המודלים מתאים יותר לביצוע המשימה – א' או ב'? יש לצרף חישובים.

שאלה 3 – תהליך בניית מודל (25 נק):

נתון ה-dataset הבא:

feature 1	feature 2	feature 3	feature 4	feature 5	feature 6	feature 7	label
17	2	0.91	--	Type A	0.33	0.34	ABC
506	8	0.03	--	Type A	0.68	0.68	DEF
333	6	0.32	--	Type A	0.97	0.94	XYZ
105	3	0.77	--	--	0.1	0.12	ABC
77	2	0.05	0.022	Type B	0.45	0.46	DEF
278	5	0.44	--	Type A	0.04	0.04	ABC
390	8	0.97	--	Type A	0.87	0.89	XYZ
--	--	--	--	--	0.3	0.28	ABC

א. הסבירו את המושגים הבאים (2-3 שורות לכל מושג) וצינו כיצד הייתם ממשים אותם ב-dataset הנ"ל:

- Data imputation
- Feature encoding
- Discretization

ב. עבור מודל מסוג neural nets, תארו פעולה טרנספורמציה נוספת שיש לבצע בשלב ה-preprocessing.

ג. נתונה מטריצת הקורלציה הבאה:

	feature 1	feature 2	feature 3	feature 4	feature 5	feature 6	feature 7	label
feature 1	1	0.32	0.02	0.04	-0.12	-0.56	0.17	0.05
feature 2	0.32	1	-0.88	-0.06	0.03	0.17	-0.19	0.33
feature 3	0.02	-0.88	1	0.11	0.04	0.22	-0.15	0.12
feature 4	0.04	-0.06	0.11	1	-0.03	0.09	0.07	0.01
feature 5	-0.12	0.03	0.04	-0.03	1	0.4	-0.02	0.1
feature 6	-0.56	0.17	0.22	0.09	0.4	1	0.94	0.75
feature 7	0.17	-0.19	-0.15	0.07	-0.02	0.94	1	0.8
label	0.05	0.33	0.12	0.01	0.1	0.75	0.8	1

- בין אילו פיצ'רים קיימת קורלציה גבוהה? ציינו 2 זוגות לפיצ'רים כאלו.
- האם הייתם מאחדים/מורידים פיצ'רים אלו? נמקו (אם מאחדים – פרטו כיצד ובאיזה אופן יתבצע האיחוד, אם מורידים – פרטו איזה פיצ'ר הייתם מורידים ומדוע).

ד. ביצעתם פעולת clustering על ה-data. ציינו 2 שימושים שניתן לעשות באמצעות הפלט שהתקבל כתוצאה מפעולה זו. התייחסו לשלב ניתוח הנתונים הראשוני ולשלב בניית המודל. ציינו 6 מודלים שמבצעים פעולה זו.

ה. הסבירו מהם outliers. כיצד הם משפיעים על המודל? תארו 2 שיטות למציאתם.

שאלה 4 – Mean Shift (20 נק'):

(א) הפרמטר העיקרי של האלגוריתם הוא h (bandwidth). איך משתנות תוצאות האלגוריתם כאשר h גדל או קטן.

(ב) נתונות 10 הנקודות הבאות:

1, 1.5, 2, 2.1, 2.15, 2.2, 2.25, 2.3, 2.45, 3.0

הריצו את אלגוריתם ה-Mean Shift עבור $h=0.25$ ו- $h=0.3$, מהן התוצאות שמתקבלות? האם הדוגמא הזאת תומכת בטענה שהעליתם בסעיף א?

(ג) תארו את אלגוריתם Edison לסגמנטציה של תמונות. איזה פרמטרים המשתמש נותן ואיך זה נכנס לנוסחא של ה-Mean Shift.

שאלה 5 – Viola and Jones (15 נק'):

(א) באלגוריתם של Viola ו-Jones משתמשים ב integral image. הסבירו איך מחשבים אותו בהינתן תמונה.

(ב) תנו דוגמא מפורטת לשימוש בו.

(ג) אם ה- integral image לא היה חלק מהאלגוריתם, כיצד זה היה משפיע עליו?

שאלה 6 – Divisive Clustering (15 נק'):

(א) מה הקלט והפלט לאלגוריתם Divisive Clustering.

(ב) האם כל נקודת קלט צריכה להיות מיוצגת על ידי וקטור של מספרים? הסבירו.

(ג) הסבירו את השלבים של האלגוריתם.

(ד) כמה פלטים נוצרים באלגוריתם ואיך מהם בוחרים את הפלט הסופי?

(ה) אם מטריתכם ליצור 3 קלסטרים, כיצד תשתמשו באלגוריתם כדי להשיג את מטריתכם?

שאלה 7 – Neural Nets (20 נק'):

(א) תארו ארכיטקטורה של רשת נוירונים שמקבלת קלט X ומסווגת אותו לאחת מ- 10 מחלקות.

(ב) כיצד תיבחר התוצאה של האלגוריתם?

(ג) באיזה loss function תשתמשו כדי לאמן את הרשת. תנו את הנוסחה שלה.

(ד) תארו ארכיטקטורה של רשת נוירונים שמקבלת קלט X ומחזירה ערך מספרי (למשל מתכונות של בית הערכה למחירו).

(ה) כיצד תיבחר התוצאה של האלגוריתם?

(ו) באיזה loss function תשתמשו כדי לאמן את הרשת. תנו את הנוסחה שלה.