בס"ד

בית הספר למדעי המחשב

שם הקורס: מבוא למדעי הנתונים

קוד הקורס: 2-7017810-1

תאריך בחינה: 26/01/2023, ד' שבט ה'תשפ"ג

סמסטר: א' מועד: א'

משך הבחינה: 2.5 **שעות**

שם המרצה: ד"ר אור חיים אנידג'ר

חומר עזר :**אין**, למעט שימוש במחשבון מדעי פשוט

פירוט הניקוד לכל שאלה:

שאלה	ניקוד מקס'	ניקוד בפועל
1	5	
2	5	
3	7	
4	8	
5	10	
6	15	
7	12	
8	13	
9	25	
סה"כ	100	

בס"ד

הוראות כלליות:

1. **הניסוח הוא בלשון רבים מטעמי נוחות ומתייחס לכולם!**
2. **במבחן ניתן לצבור עד 100 נקודות, והוא איננו מכיל שאלות בחירה; יש לענות על כל השאלות ללא יוצא מן הכלל.**
3. **יש להקפיד לענות תשובות מלאות, ומפורטות ככל הניתן. יחד עם זאת, תשובות לא מדויקות, או לא מלאות, או מלאות שלא לצורך - עלולות לגרור יריידת נקודות.**
4. **המבחן מחולק לשלושה חלקים:**
    a. **שאלות ידע כללי המבוססות על ההרצאות (50 נקודות).**
    b. **שאלות קוד והסתברות המבוססות על התרגולים (25 נקודות).**
    c. **שאלה המבוססת על תכנון פתרון לבעיית למידה (25 נקודות).**
5. **יש לציין במחברת הבחינה את מספר השאלה אשר עבורה ניתנת התשובה.**

בהצלחה! 😊

בס"ד

חלק א' - שאלות ידע כללי

שאלה 1 (5 נק'):
הסבירו מהו ההבדל בין קלסיפיקציה לבין רגרסיה? מדוע רגרסיה לוגיסטית יוצאת מהכלל?

שאלה 2 (5 נק'):
מהו קידוד one-hot ומתי נשתמש בו (מבחינת סוג המשתנה)? מה יהיה גודלו?

שאלה 3 (7 נק'):
עבור רגרסיה לוגיסטית, למדנו שמשוואת המודל היא מהצורה: $P(Y=1)=\frac{1}{1+e^{-(b_0+b_1X)}}$. ברמת הגרף הנוצר מעקומת הרגרסיה, איך משפיעים על ההצגה הגרפית הפרמטרים b_0, b_1?

שאלה 4 (8 נק'):
הגדירו והסבירו מה הן נקודות Support Vector? איזה עיקרון הן באות ליישם ברמת בניית מודל ה-SVM והמרחק שלו מהן?

שאלה 5 (10 נק'):
עבור המטריצה הבאה, חשבו את המטריקות הבאות: Accuracy, Precision, Recall (יש צורך במספר מדויק לכל מטריקה, וכמו כן להציגו ראשית באמצעות שבר מסוג $\frac{X}{Y}$).

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

שאלה 6 (15 נק'):
באיזו דרך ניתן לחשב K אופטימלי (כמות הקלאסטרים) עבור אלגוריתם k-means? מה מייצג הערך המאפשר לחשב את K? הסבירו כיצד היא באה לידי ביטוי בצורה גרפית (כולל תיאור גרפי), וכיצד תבחרו את K?

חלק ב' – שאלות קוד והסתברות

שאלה 7 (12 נק'):
נתון קוד. הניחו שקיימים המשתנים x_train, x_test, y_train, y_test.

1. $model = LogisticRegression()$
2. $model.fit(x\_test, y\_train)$

3. $y\_pred = model.prediction(x\_test)$
4. $Accuracy = model.score(y\_pred, y\_train)$

מצאו את שתי הטעויות הלוגיות בקוד, באילו שורות הן, וכתבו כיצד יש לתקן אותן, ומדוע? (תשובה ללא נימוק מלא לא תקבל ניקוד).

שאלה 8 (13 נק'):

בביה"ח "בינה מלא-חוטית" ישנן 4 בדיקות למחלת השפעת:

- ההסתברות שהרופא יבדוק מטופל בבדיקה A היא 50%, והיא נכונה ב- 95%.
- ההסתברות שהרופא יבדוק מטופל בבדיקה B היא 30% והיא נכונה ב- 70%.
- ההסתברות שהרופא יבדוק מטופל בבדיקה C היא 15% והיא נכונה ב- 60%.
- ההסתברות שהרופא יבדוק מטופל בבדיקה D היא 5% והיא נכונה ב-55%.

א. אדם נבחר באופן אקראי, הלך לרופא, ונבדק. מה ההסתברות שתוצאות הבדיקה שלו נכונה?

ב. נאמר שהבדיקה של אותו אדם הראתה תוצאה שגויה. מה ההסתברות שהוא נבדק בבדיקה C?

<u>חלק ג' – שאלת תכנון פתרון לבעיית למידה</u>

שאלה 9 (25 נק'):

בחברה המונה מס' רב של עובדים, מדי יום מתנהלות פגישות פנימיות בין עובדי החברה, ופגישות חיצוניות של עובדי החברה עם לקוחות, ספקים, וכו'. קיבלתם משימה לבנות מערכת זיהוי פנים עבור עובדי החברה (רשימה סגורה), לקוחותיה וספקיה (רשימה פתוחה ומתעדכנת).

לצורך כך, הניחו שקיימת עבורכם גישה למערכת זיהוי פנים אידאלית ומושלמת המותקנת בכניסה לחדרי הישיבות של החברה, כך שרק אדם אחד מופיע בכל תמונת פנים, והמערכת יודעת לאתר במדויק היכן נמצאות הפנים של האדם בתמונה, כך שהפנים המחולצות מהתמונה נשמרות בתת תמונה (מטריצה) בגודל קבוע (שורות ועמודות), ומקודדות בסופו של תהליך כל תמונת פנים - לווקטור מאפיינים קבוע באורך 128.

תארו ובנו מערכת אשר בהנתן פגישה מרובת-משתתפים, מחזירה את רשימת עובדי החברה, ורשימת הספקים\לקוחות אשר נכחו בפגישה על סמך יכולות ניתוח הפנים של המערכת. יש להתייחס כיצד תבצעו את ההפרדה בין עובדי החברה לבין לקוחותיה וספקיה בעת בניית המודל. בתשובה שלכם, הקפידו להתייחס למונחים מקצועיים כדוגמת Supervision, Unsupervision, Classification, Regression, סוגי המודלים וכיצד ייבנו, כיצד תתבצע ההפרדה בין עובדי החברה לבין לקוחותיה וספקיה, וכן אילו סוגים של Datasets תצטרכו, במידה ותצטרכו. בנוסף, יש להתייחס למטריקות בהן תבצעו שימוש לטובת הערכת הביצועים של המודלים שתבנו.