בס"ד

בית הספר למדעי המחשב

שם הקורס: מבוא למדעי הנתונים

קוד הקורס: 2-7017810-1

תאריך בחינה: 22/02/2023, א' אדר ה'תשפ"ג

סמסטר: א' מועד: ב'

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 את ההבדל בין למידה פרמטרית לבין למידה לא פרמטרית.

שאלה 2 (5 נק'):
מבין המודלים הבאים – KNN, SVM, עץ החלטה (Decision Tree), רגרסיה לינארית ורגרסיה לוגיסטית, אילו מודלים מאופיינים כבעלי למידה פרמטרית, ואילו כבעלי למידה לא פרמטרית?

שאלה 3 (7 נק'):
מודל SVM איננו מותאם מלכתחילה לבעיות multi-class (ריבוי מחלקות), אלא לבעיות סיווג בינאריות. תארו והסבירו שתי שיטות לבניית מודל SVM עבור בעיות קלסיפיקציה multi-class?

שאלה 4 (8 נק'):
א. מנו 2 חסרונות של אלגוריתם k-Nearest Neighbors.
ב. מהו פיתרון אפשרי לכל אחד מן החסרונות שציינתם, וכיצד הוא אכן פותר את החסרון?

שאלה 5 (10 נק'):
נניח שברשותכם סט מבחן (test-set) המורכב מ-150 דגימות, כך ש-100 מהדגימות הן בעלות label של "כלב" ואילו 50 מהדגימות של הן בעלות label של "חתול". לצורך כך, בניתם מודל קלסיפיקציה כלשהו, שמטרתו בהינתן תמודה מסט המבחן, היא לסווג "כלב" או "חתול". בפועל, המודל סיווג 90% מהכלבים כ-"כלב" ו-40 אחוז מהחתולים כ-"חתול".

בנו את ה-confusion matrix עבור הנתונים הנ"ל (יש להציג חישוב מלא לכל ערך במטריצה).

שאלה 6 (15 נק'):
למדנו 3 שיטות עבור חישוב מרחק בין שני קלאסטרים. מהן? הסבירו אותן בפירוט.

חלק ב' – שאלות קוד והסתברות

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

1. $y_{train} = [0.1, 0.2, 0.3]$
2. $y_{test} = [0.4, 0.5, 0.6]$
3. $model = LogisticRegression(\texttt{random\_state=123})$
4. $model.fit(x\_train, y\_train)$
5. $y\_pred = model.prediction(x\_test)$
6. $Accuracy = model.score(y\_pred, y\_test)$

א. מה משמעות איתחול הפרמטר random_state=123 בשורה 3?
ב. מצאו את הבעיה בקוד, באילו שורות הן, וכתבו כיצד יש לתקן אותה.

שאלה 8 (13 נק'):

בכד ישנם 4 כדורים כחולים, 2 כדורים אדומים ו-3 כדורים צהובים.

הוצאנו כדור באופן אקראי מבלי להחזיר אותו לכד.

מה ההסתברות ששלפנו בפעם הראשונה כדור צהוב, בהינתן שבפעם השנייה לא שלפנו כדור צהוב?

חלק ג' – שאלת תכנון פתרון לבעיית למידה

שאלה 9 (25 נק'):

בבורסה של מדינת חלם עלה רעיון לבנות אלגוריתם אשר יהיה מסוגל לנבא בתאריך נתון את ערכה של מנייה נתונה, בשבוע שלאחריו. ישנו מאגר נתונים של 10 שנים אחורה עבור המניות השונות בבורסה, אשר ניתן לייצר ממנו מידע עבור ערך המניות בכל תאריך, הימים בשבוע, אירועים חריגים שקרו במדינת חלם (למשל, גילוי חיידק כלשהו במוצרים של חברה כלשהי). בנוסף, בכל חודש, ישנו צורך להתמודד עם מניות חדשות של חברות חדשות אשר הצטרפו לבורסה באותו חודש.

א. תארו ובנו מערכת אשר בהנתן מערכת בורסאית המכילה מידע כמובא לעיל, תנבא את ערכה של מנייה נתונה בתאריך מסוים, עבור שבוע לאחר מכן. בתשובה שלכם, הקפידו להתייחס למונחים מקצועיים כדוגמת Supervision, Unsupervision, Classification, Regression, סוגי המודלים וכיצד ייבנו, כיצד תייצרו את סט המאפיינים הדרוש עבור חיזוי של שבוע מראש, וכן אילו סוגים של Datasets תצטרכו, במידה ותצטרכו. בנוסף, יש להתייחס למטריקות בהן תבצעו שימוש לטובת הערכת הביצועים של המודלים שתבנו.

ב. בניתם מערכת מוצלחת, אך לרוע המזל חלק מהשדות בעמודות מסוימות, בשורות מסוימות – נמחקו כחלק מתקיפת סייבר (ערך null). כיצד תשלימו ערכים אלה? (תשובה מהצורה "לכתוב 0 במקום null" לא תתקבל).

ג. האם השלמת ערכים אלה תבוצע באמצעות חישוב על העמודות (Features) של המטריצה, או על השורות שלה? (הדוגמאות)