פתרון מועד X – מבחן בלמידת מכונה 2021-22 חורף, סמסטר א
3 שעות

1. [5] הצדקת Cross-Entropy: באילו תנאים מזעור ה- CE הינו זהה למיקסום ה- likelihood? סמן את כל התשובות הנכונות
   a. הרעש הבלתי ניתן להפחתה מתפלג בצורה אחידה
   b. הדגימות ב- D הן בלתי תלויות סטטיסטית אחת בשניה
   c. הסתברויות הפריור (Prior) של כל ההיפותזות שוות
   d. כל 2 מאפיינים אינם תלויים סטטיסטית זה בזה
   e. המאפיינים אינם קורלטיביים בינם לבין עצמם

2. [5] ביצעו רגרסיה שהשתמשה Thin Plate Splines. ניסו את הרגרסיה בדרגות גמישות 1,2,3,4,5 ובכל דרגה שניסו, בנו מודל חיזוי וחישבו את ה loss על קבוצת וולידציה. בדרגת גמישות 4 שגיאת הוולידציה הגיע לערכה הנמוך ביותר ואולם בפער מהשגיאה על קבוצת האימון שהיתה קרובה ל 0. רשום באילו טכניקות כדאי להשתמש כדי לשפר את שגיאת הוולידציה? (רשום כל מה שנכון):
   a. forward Feature Selection
   b. הוספת מאפיינים מחושבים
   c. רגולריזצית לאסו
   d. ביצוע Cross validation על בסיס BootStrapping
   e. הוספת דאטה מתוגת חדשה לקבוצת הוולידציה
   f. שימוש ב Bagging על בסיס Bootstrapping
   g. הוספת דאטה מתויגת לקבוצת האימון הקימת

3. [5] מבצעים רגולריזצית L1 ברגרסיה לוגיסטית. רשום את כל המשפטים המתקיימים, כאשר מגדילים בהדרגה את העונש.
   a. הביאס אינו קטן בהכרח
   b. אם 2 מאפינים הם קורלטיבים, המשקולות של שניהם יתקרבו ל 0
   c. כל המשקולות יקטנו
   d. ככל שנגדיל את העונש , חלק מהמשקולות יעלמו
   e. אם 2 מאפינים הם קורלטיבים, אחד מהם יעלם בסיכוי גבוה

4. [5] אם ברצונך לדרג קלסיפיירים ע"י מתן מספר אחד לכל קלסיפייר המציין את איכותו המשקללת בין סוגי השגיאות השונים. באילו מדדים ניתן להשתמש לשם כך? רשום כל מה שנכון:
   a. F1
   b. Precision
   c. Recall
   d. AUC
   e. Accuracy
   f. Balanced accuracy

5. [6] סמן כל מה שנכון לגבי SVM וההיפר-פארמטר C (ה- Capacity עליו למדנו בכיתה).

a. הגדלת C מעלה את הגמישות
b. הגדלת C מגדילה את מספר וקטורי התמך Support Vectors
c. הגדלת C מקטינה את שגיאת ה Variance
d. הגדלת C מגדילה את מספר וקטורי התמך Support Vectors
e. C=5.5 עלול לאפשר מישור הפרדה המסווג עד 5 דוגמאות בצורה שגויה

6. [10] PCA : נתון dataset המתאר את נתוני הפשיעה לכל מדינה בארה"ב (4 מימדים- לכל מדינה). רוצים לצמצם ל 2 מימדים ולהציג ויזואלית על ידי שימוש ב-PCA.

לאחר הצמצום מקבלים 2 טרנספורמציות לינאריות אחת לרכיב הראשי הראשון והשנייה לרכיב הראשי השני:

	PC1	PC2
Murder	0.5359	-0.4182
Assault	0.5832	-0.1880
Urban pop	0.2782	0.8728
Rape	0.5434	0.1673

a. מה ניתן לומר על כל אחד מהרכיבים? רשום כל מה שנכון
   1) הרכיב הראשון משקלל יותר את אחוז האוכלוסייה העירונית והרכיב השני משקלל יותר את הפשיעה.
   2) הרכיב הראשון משקלל יותר את הפשיעה והשני משקלל יותר את אחוז האוכלוסייה העירונית.
   3) הרכיב הראשון משקלל יותר את סטיית התקן של כל מאפיין והרכיב השני משקלל יותר את הממוצע.
   4) הרכיב הראשון משקלל את המאפיינים לאחר נירמול סטנדרטי והרכיב השני לפני נירמול.

b. נתונה הדוגמא:

$$X = (\text{Murder}= 5,\ \text{Assault}= 209, \text{Urban Pop}= 15, \text{Rape}= 20)$$

הערה: הניחו כי הדוגמא הזו כבר נורמלה, והשתמשו בטרנפורמציות הנתונות לעיל.

חשבו את הוקטור Z (ממימד 2) המתקבל לאחר הכפלה בכל אחד מהרכיבים הנתונים בשאלה:

$$Z = (139.6093,\ -24.945)$$

7. [20] על סמך מדגם של 100 נבדקים בסיכון למחלה, מצאו בקופת חולים כי המחלה מופיעה ב 40% מהמדגם. חישבו גם כי לבדיקת סקר ישנה ספציפיות (Specificity) 0.3 ורגישות (Recall) 0.9., העלות של הבדיקה 1 ש"ח, העלות של ביצוע טיפול באבחנה חיובית הינה 100 ש"ח ואילו עלות חולה שלא טופל (ז.א. אי זיהוי המחלה) הינה 600 ש"ח

a. מלאו את מטריצת העמימות שעל פיה חישבו בקופה את המספרים לעיל.

	Predicted 0	Predicted 1
Actual 0		
Actual 1		

	Predicted 0	Predicted 1
Actual 0	18	42
Actual 1	4	36

b. באיזה אלפא תשתמש כדי לחשב F-measure משוקלל (על פי הנוסחא):
תזכורת. α הוא מספר בין 0 ל 1, והיחס בין α ל $1-\alpha$ הוא יחס העלות של FP לעלות של FN

$$F-Measure = \frac{1}{\alpha \cdot \frac{1}{Precision} + (1-\alpha) \cdot \frac{1}{Recall}}$$

$$alpha = \frac{FP\ price}{FP\ price + FN\ price}$$

$$\frac{100}{100+600} = \frac{100}{700} = 0.143$$

c. מהו ערכו ה F-measure המשוקלל?

1/(0.143*1/(36/78)+0.857*(1/0.9)=1/(0.143*2.167+0.857*1.11)=1/(1.261)=0.793

d. מה העלות לקופה כאשר מבצעים את בדיקת הסקר ל- 1000 נבדקים מתוך האוכלוסיה שבסיכון?

$$\left(100 \cdot \frac{FP}{Total} + 600 \cdot \frac{FN}{Total} + 1\right) \cdot 1000$$

$$(100 \cdot \frac{Actual\ Negative}{Total} \cdot \frac{FP}{Actual\ Negative} + 600 \cdot \frac{Actual\ Positive}{Total} \cdot \frac{FN}{Actual\ Positive}) \cdot 1000$$

$$(100 \cdot \frac{Actual\ Negative}{Total} \cdot (1-Specificity) + 600 \cdot \frac{Actual\ Positive}{Total} \cdot (1-Recall)) \cdot 1000$$

$$(100 \cdot 0.6 \cdot 0.7 + 600 \cdot 0.4 \cdot 0.1 + 1) \cdot 1000 = (42 + 24 + 1) \cdot 1000 = 67{,}000$$

e. מהי העלות לקופה אם לא תבצע כלל את בדיקת הסקר ל- 1000 הנבדקים הללו?
<u>האם כדאי לקופה לבצע את בדיקת הסקר?</u>

$$\left(600 \cdot \frac{Actual\ Positive}{Total}\right) \cdot 1000$$
$$(600 \cdot 0.4) \cdot 1000 = 240{,}000$$

כדאי לבצע את הסקר

8. [24] מעוניינים לבנות Naive Bayes Classifier (NBC) שמסווג ל 3 קטגוריות A,B,H .
B,A הם קטגוריות של מחלות ו H היא הקטגוריה ל"בריא". הקלט הוא דו מימדי (x1,x2) הכולל 2 בדיקות בינאריות. בדיקה יכולה להיות חיובית או שלילית.
ה recall (רגישות) של בדיקה חיובית x1 עבור חולה במחלה A היא 0.9 ועבור חולה במחלה B היא 0.8 .
בדיקה x1 שלילית תסווג בריאים עם קצב שגיאת FP (false positive rate=1-specificity) של 0.1
ה recall (רגישות) של בדיקה x2 עבור חולה במחלה A היא 0.9 ועבור חולה במחלה B היא 0.9. בדיקה x2 שלילית תסווג בריאים בספציפיות 0.9

ידוע כי מחלה A מופיעה באוכלוסיה הכללית ב 1% מיקרים וכך גם מחלה B.
ידוע גם כי 1.5% מהאוכלוסיה חולים ב A או ב B.

a. מהם (שיערוך) הסתברויות ה priors של כל הקטגוריות?
P(A)=0.01; p(B)=0.01; p(H)= 0.985

b. מהן (שיערוך) ההסתברויות המותנות (ללא החלקות) שאותן יש לדעת כדי לסווג ע"פ NBC את הדוגמא (+,-)?
הנחיה (ורמז) לגבי צורת הרישום: ההסתברות המותנית לבדיקה x_i חיובית בהינתן קטגוריה C, צריכה להירשם בצורה: $p(x_i=+|C)=$

$p(x_1=-|A)=0.1;\ p(x_1=-|B)=0.2;\ p(x_1=-|H)=0.9;$
$p(x_2=+|A)=0.9;\ p(x_2=+|B)=0.9;\ p(x_2=+|H)=0.1;$

c. איזו קטגוריה תיבחר ע"פ MAP לדוגמא (+,-) ע"פ השיערוכים שחישבת? הראה את החישובים אותם יש לעשות כדי לבחור קטגוריה
(כדי להימנע משגיאות נגררות, רשום את נוסחת החישוב בה השתמשת, לפני הצבת השיערוכים מהסעיפים הקודמים)
argmax{0.01*0.1*0.9; 0.01*0.2*0.9; 0.985*0.9**0.1}=argmax{0.0009; 0.0018; 0.089}
קטגוריה H

d. איזו קטגוריה תיבחר ע"פ ML?
(כדי להימנע משגיאות נגררות, רשום את נוסחת החישוב בה השתמשת, לפני הצבת השיערוכים מהסעיפים הקודמים)

argmax{0.1*0.9; 0.2*0.9; 0.9**0.1}=argmax{0.09; 0.18; 0.09}
קטגוריה B

e. מהי ה joint probability המשוערכת על פי NBC של p((-,+),A)
(כדי להימנע משגיאות נגררות, רשום את נוסחת החישוב בה השתמשת, לפני הצבת השיערוכים מהסעיפים הקודמים)
0.0009

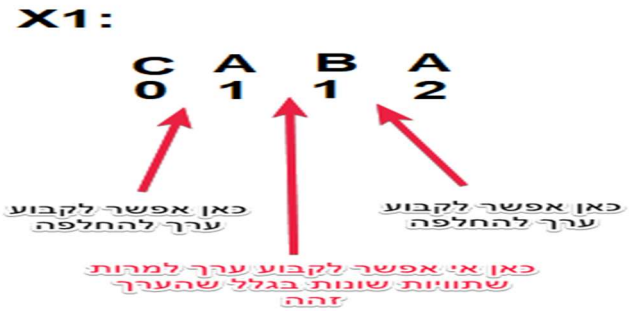
f. מכיוון שאת הרגישות והספציפיות העריכו רק על סמך 90 דוגמאות (המדגם) החליקו את שיערוכי ההסתברויות בעזרת m-estimate ע"י תוספת של 10 דוגמאות וירטואליות.
ההנחה היא ששכיחות החיוביים בבדיקה x1, $p(x_1=+)$, באוכלוסייה הכללית היא 0.002

מה יהיה השיערוך של p(x1=-|A) לאחר החלקה?

(90*0.1+10*0.002)/(90+10)=9.02/100=0.0902

9.]20[עצים: נתונה קבוצת האימון של מרחב קלט דו מימדי (x_1,x_2) (2 המאפיינים הם נומרים) והמתויגת לקלסיפיקציה של 3 קטגוריות {A,B,C}
D={((2,2),A), ((1,2),B), ((0,4),C), ((1,4),A)}

a. על פי השיטה לפיצול מאפין נומרי שלמדנו בכיתה, אילו שאלות מהצורה x<? הינן מועמדות לתפקיד השאלה המפצלת שבשורש? רשום את קבוצות התיוגים (labels) של הדוגמאות שנבחרות לכל ילד בעבור כל שאלה אפשרית לפיצול בשורש העץ.
(רמז בקבוצת האימון D שבשאלה, יש רק 3 שאלות מועמדות לתפקיד הצומת שצריך לבדקן)



$x_1<0.5$ וילדים {C},{A,A,B}
$x_2<3$ וילדים (A,B},{A,C}
$x_1<1.5$ וילדים {A},{A,B,C}

b. מהי האנטרופיה לפני הפיצול?:
-2/4log(2/4)-2*(1/4log(1/4))=0.5*1+0.5*2=1.5

c. איזו שאלה תיבחר לשורש, מהי האנטרופיה המשוקללת של הילדים של פיצול זה? רשום קודם את תכולת התיוגים (לייבלים) של הילדים של שורש כזה.

השורש שיבחר $x1<0.5$

$$\text{Wentropy } (x_{1<0.5}) = 0.25(0)+0.75(-(2/3)\log(2/3)-(1/3)\log(1/3))=0.75(0.92)=0.69$$

ושל השאלה $x2<3$: (לא נידרשים לחשב)

$$\text{Wentropy } (x_2) = 0.5(1)+0.5(1)=1$$

d. מהו ה Information Gain של בחירת $x1<1$ בשורש?

$1.5-0.69=0.81$

בהצלחה