שאלון Y

1. [10] רגרסיה לוגיסטית: רוצים לסווג תמונות בעזרת רגרסיה לוגיסטית רבת קטגוריות.
בתמונות יש שלושה סוגים של חיות (כלב, חתול ועכבר) וכל חיה נמצאת במצב רגשי אחד מתוך שלושה (שמחה, עצובה או מפוחדת). בנוסף לתמונות, מצורף לדאטה קובץ CSV שמכיל את התיוגים של התמונות (ה- Labels):

```
[13] dataset
```

Picture Id	Animal	Emotion
1	Dog	Happy
2	Cat	Sad
3	Mouse	Scared
4	Mouse	Happy
5	Dog	Sad
6	Dog	Happy
7	Cat	Happy
8	Dog	Scared
9	Cat	Scared
10	Mouse	Happy

אנו רוצים לסווג כל תמונה ל- 2 סוגי סיווגים; איזו חיה מופיעה בתמונה, ומה הרגש שהיא מביעה.

a. אם נשתמש בייצוג של OneHotEncoder (OHE) על הלייבלים (Labels):

```
[14] enc = OneHotEncoder(categories=[['Dog', 'Cat', 'Mouse'], ['Happy', 'Sad', 'Scared']])
     enc.fit_transform(dataset).toarray()
```

מה יהיה המימד של וקטור הלייבלים t?

6

b. כיצד יראה וקטור הלייבלים t עבור תמונה מסוימת של כלב שמח?

[1, 0, 0, 1, 0, 0]

c. בודקים את החיזוי על התמונה מהסעיף הקודם ומקבלים כי כל החיזויים הם 1/3 (כלומר וקטור החיזויים y שיוצא מהקלסיפיקציה הוא $[\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3}]$). חשבו את ה MCCE עבור תמונה זו [השתמשו ב- ln (לוג עם בסיס טבעי - e)]

$$MCCE(y,t) = -\sum_{i}^{k} t_i \ln(y_i) + (1-t_i)\ln(1-y_i)$$

$$-\left(ln\left(\frac{1}{3}\right) + ln\left(1-\frac{1}{3}\right) + ln\left(1-\frac{1}{3}\right) + ln\left(\frac{1}{3}\right) + ln\left(1-\frac{1}{3}\right) + ln\left(1-\frac{1}{3}\right)\right) = 3.819$$

d. אם נשתמש בייצוג של Dummy Encoding, מה יהיה מימד וקטור הלייבלים t?

4

2. [8] הצדקת CE: כאשר מצדיקים את מיזעור ה- CE ממקסמים את $P(h_w|D)$ ומשתמשים במשפט Bayes:

$$P(h_w|D) = \frac{P(D|h_w)P(h_w)}{P(D)}$$

a. מדוע איננו חיבים לחשב את $P(D)$?
   i. $P(D)$ אינו תלוי בהיפותזה, לכן הוא קבוע לכל ההיפותזות
   ii. $P(D)$ תמיד שווה ל- 1
   iii. $P(D)$ לא ידוע, לכן ניתן להתעלם ממנו
   iv. $P(D)$ תלוי ב- $P(h_w|D)$, לכן כשנחשב את $P(h_w|D)$ נדע את $P(D)$

b. מה הבסיס להשמטת $P(h_w)$?
   i. אנו יכולים לחשב ידנית את ה- Prior לכל היפותזה, לכן אנו יכולים להוסיף זאת לאחר המיקסום
   ii. אין בסיס תיאורטי להשמטת $P(h_w)$
   iii. איננו יודעים מה ה- Prior לכל היפותזה, לכן ב- Maximum Likelihood מניחים שההסתברות לכל ההיפותזות זהה
   iv. לא משמיטים את $P(h_w)$, מחשבים אותו ומכפילים ב- $P(D|h_w)$ כדי למצוא את $P(h_w|D)$

c. לאיזו הנחת אי-תלות אנו זקוקים כדי לחשב את $P(D|h_w)$?
   i. $P(D)$ ו- $P(h_w|D)$ אינן תלויות סטטיסטית זו בזו, לכן כשנחשב את $P(h_w|D)$ נדע את $P(D)$
   ii. הדוגמאות ב- D אינן תלויות סטטיסטית זו בזו, לכן ההסתברות שווה למכפלת ההסתברויות של כל דוגמא לחוד
   iii. אנו לא זקוקים להנחת אי-תלות כדי לחשב את $P(D|h_w)$
   iv. ההיפותזות אינן תלויות סטטיסטית זו בזו, לכן ניתן להניח ש- $P(h_w)$ שונה בין ההיפותזות

3. [6] שגיאות: בקלסיפיקציה בעזרת רגרסיה לוגיסטית רואים שהאימון אינו מצליח להוריד את שגיאת האימון.

a. מה יכולה להיות הסיבה לכך?
   i. שגיאת ביאס (Bias) גדולה מידי
   ii. שגיאת ווריאנס (Variance) גדולה מדי
   iii. שגיאה בלתי ניתנת להפחתה קטנה מידי
   iv. שגיאת ה- Cross Entropy גדולה מידי

b. כיצד ניתן לעזור לרגרסיה לוגיסטית להפחית את השגיאה הזו? בחרו בכל התשובות המתאימות
   i. להוריד פיצ'רים
   ii. להוריד דאטה
   iii. להגדיל את קבוע הרגולריזציה
   iv. להוסיף דאטה
   v. לייצר פיצ'רים נוספים מהפיצ'רים הנתונים

4. [5] KPI: קלסיפייר אומן כדי לשערך את ההסתברות P(+|x)
כאשר משתמשים בסף (Threshold) של 0.5 מקבלים Precision ו- Recall מסוימים. מה יקרה ל- FP ומה יקרה ל- FN אם נקטין את הסף?
FP יגדל, FN יקטן

5. [5] KPI: בקלסיפיקציה בינארית, בהנחה שקבוצת הטסט (Test Set) מאוזנת (50% מהדאטה חיוביים ו- 50% שליליים), מה יהיו ה- F1 score וה- Balanced-Accuracy עבור קלסיפייר בינארי "טיפש" שחוזה תמיד P(+|x) = 0 על כל x? הניחו ש- $\frac{0}{0}$ שווה ל- 1.

$$Recall = \frac{TP}{TP+FN} = \frac{0\%}{50\%} = 0$$

$$Precision = \frac{TP}{TP+FP} = \frac{0\%}{0\%} = 1$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{50\%}{50\%} = 1$$

$$\text{Balanced Accuracy} = \frac{Specificity + Recall}{2} = \frac{1+0}{2} = 0.5$$

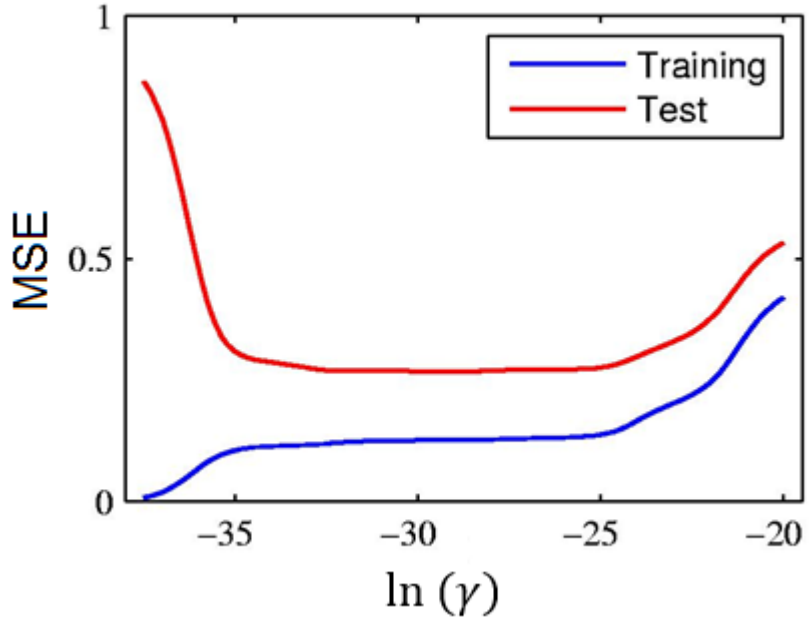
$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precesion} + \text{Recall}} = \frac{2 \cdot 1 \cdot 0}{1+0} = 0$$

6. [6] CV: רוצים לשערך את השגיאה של קלסיפייר מסוים, בעזרת Dataset מתוייג של 100 דוגמאות.
   a. החליטו להשתמש ב- Leave 2 Out Cross Validation. לכמה אימונים נזדקק (מספר מדויק)?
   $$\binom{100}{2} = 4950$$
   b. החליטו להשתמש ב- 10 Fold Cross Validation. לכמה אימונים נזדקק (מספר מדויק)?
   10

7. [5] נתונה קבוצת אימון של 1000 דוגמאות מתויגות. חישבו ומצאו שבשעה ניתן לבצע בערך 60 אימונים על קבוצת אימון של 630 דוגמאות כולל חישוב מדד F1 על קבוצת וולידציה של כ 370 דוגמאות. מעוניינים לנצל משאבי חישוב הניתנים ל- 10 שעות כדי לבצע שיערוך מדויק ככל האפשר של מדד F1 על קבוצת הטסט. באיזו טכניקה נשתמש?
   i. נשתמש בשיטת ה- bootstrapping כדי להגריל 1000 קבוצות אימון בגודל של 630
   ii. נבצע Leave 370 out CV
   iii. נשתמש בשיטת ה- bootstrapping כדי להגריל 600 קבוצות אימון בגודל של 1000
   iv. נבצע 3-foldCV
   v. נבצע LOOCV

8. [5] רגולריזציה: נתונה רגרסיה לוגיסטית המשתמשת רגולריזציית לאסו (L1). מבצעים CV כדי למצוא קבוע רגולריזציה (γ) בטווח $(10^{-8}, 1)$. מה יקרה לשגיאה ככל שנגדיל את קבוע הרגולריזציה?
   i. שגיאת הוואריאנס תגדל וחלק ממקדמי המאפיינים יתאפסו
   ii. שגיאת הוואריאנס תקטן וחלק ממקדמי המאפיינים יתאפסו
   iii. שגיאת הוואריאנס תגדל וחלק ממקדמי המאפיינים יתקרבו ל- 0
   iv. שגיאת הוואריאנס תקטן וחלק ממקדמי המאפיינים יתקרבו ל- 0

9. [5] רגולריזציה: עבור רגרסיה מודדים שגיאה על קבוצת הטסט ועל קבוצת האימון. ומשתמשים בקבוע רגולריזציה (γ). לפי הגרף הנתון, מהו טווח העונשים (γ) שעבורו למודלים הנלמדים יהיה High-Bias? [זכרו כי בגלל שערכי קבוע הרגולריזציה (γ) קטנים, משתמשים בסקלה לוגריתמית (לפי בסיס טבעי - e)]



עבור $-25 < \ln(\gamma) < -20$ יש High-Bias

כלומר הטווח הוא $1.38 \cdot 10^{-11} < \gamma < 2.06 \cdot 10^{-9}$

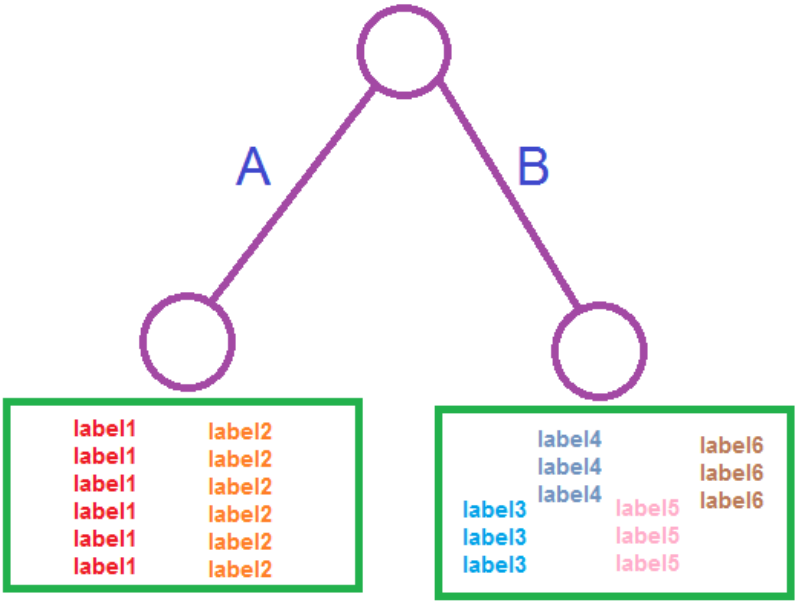
10. [4] SVM: בנינו מודל SVM עם קרנל גאוסיני לקבוצת אימון D וקיבלנו שבנקודה בה ה-Margin מתמקסם, אחד ה- Slack Variables הוא 1.5. מה ניתן להסיק?

i. לפחות דוגמת אימון אחת מסווגת ע"י המודל בצורה שגויה
ii. C>0 והדוגמאות ב- D אינן ניתנות להפרדה לינארית
iii. לפחות 2 דוגמאות אימון נמצאות בתוך ה- Margin
iv. C>0 ולפחות דוגמת אימון אחת נמצאת בתוך ה- Margin

11. [5] SVM: מה יקרה ככל שנגדיל את ה Capacity (C) באימון SVM? בחרו בכל התשובות המתאימות

i. ה- Margin יגדל
ii. ה- Margin יקטן
iii. שגיאת ה- Variance תגדל
iv. שגיאת ה- Variance תקטן
v. נקבל יותר דוגמאות אימון שמודל ה- SVM סיווג באופן שגוי
vi. נקבל יותר נקודות תמך (Support)

12. [9] עצי החלטה: צומת בעץ מפצל קבוצת דוגמאות ל- 2 קבוצות של דוגמאות: A ו- B.
קבוצה A מכילה 2 ערכי מטרה (Label) שונים ו- 6 דוגמאות לכל לייבל. קבוצה B מכילה 4 ערכי לייבל שונים ו- 3 דוגמאות לכל לייבל. (שימו לב כי כמות הדוגמאות זהה בכל אחת מהקבוצות). [בשאלה זו השתמשו ב- ln (לוג עם בסיס טבעי - e)]



a. איזו קבוצה היא פחות הומוגנית (ע"פ מדד האנטרופיה)?
   i. A
   ii. B
   iii. האנטרופיה זהה ב- 2 הקבוצות ולכן ההומוגניות זהה
   iv. לא ניתן לחשב את האנטרופיה לאחר הפיצול מכיוון שלא ידועה האנטרופיה לפני הפיצול

b. חשב את מדד האנטרופיה לכל אחת מהקבוצות

$$H = -\sum_{i=1}^{N} p_i \cdot ln(p_i)$$

$$H_A = -\left(\frac{1}{2}\cdot \ln\left(\frac{1}{2}\right) + \frac{1}{2}\cdot \ln\left(\frac{1}{2}\right)\right) = 0.693$$

$$H_B = -\left(\frac{1}{4}\cdot \ln\left(\frac{1}{4}\right) + \frac{1}{4}\cdot \ln\left(\frac{1}{4}\right) + \frac{1}{4}\cdot \ln\left(\frac{1}{4}\right) + \frac{1}{4}\cdot \ln\left(\frac{1}{4}\right)\right) = 1.386$$

c. חשב את האנטרופיה המשוקללת של A, B

$$H_{Children} = \sum_{child \in Children} p_{child} \cdot H_{child}$$

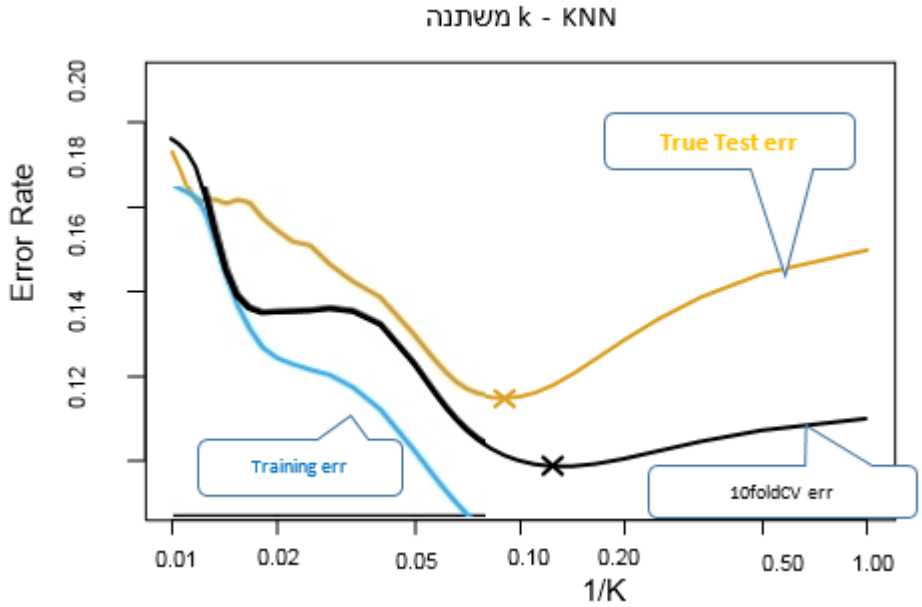
$$\frac{12}{24}\cdot 0.693 + \frac{12}{24}\cdot 1.386 = \frac{0.693 + 1.386}{2} = 1.0395$$

d. חשב את ה- Information Gain (IG)

$$IG = H_{Parent} - H_{Children}$$

$$-2\cdot \frac{6}{24} ln\left(\frac{6}{24}\right) - 4\cdot \frac{3}{24} ln\left(\frac{3}{24}\right) - 1.0395 = 0.693$$

13. [7] KNN: מעוניינים לבחור K עבור אלגוריתם KNN. מבצעים 10-fold CV על ה- Train ובוחרים K עבור אלגוריתם ה- KNN. לאחר מכן מריצים חיזוי של אלגוריתם ה- KNN עם ה- K הנבחר על ה- Test. מקבלים את הגרף הבא:



a. איזה K ייבחר?
   i. k=20
   ii. k=12
   iii. k=8
   iv. k=1

b. מה יכולה להיות הסיבה להבדל בין הגרף השחור לצהוב?
   i. ה- CV מעריך שביצועי המודל טובים יותר מהביצועים האמיתיים (על ה- Test), מכיוון שמודלי ה- CV כוונו לוולידציה (שהייתה שונה במקצת מהטסט)
   ii. ה- CV מעריך שביצועי המודל טובים יותר מהביצועים האמיתיים (על ה- Test), מכיוון שמודלי ה- CV נבנו רק על סמך 90% מקבוצת האימון
   iii. ה- CV מעריך שביצועי המודל גרועים יותר מהביצועים האמיתיים (על ה- Test), מכיוון שמודלי ה- CV נבנו רק על סמך 90% מקבוצת האימון
   iv. ה- CV מעריך שביצועי המודל גרועים יותר מהביצועים האמיתיים (על ה- Test), מכיוון שמודלי ה- CV כוונו לוולידציה (שהייתה שונה במקצת מהטסט)

14. [5] K-Means: נתון קלאסטר שנוצר בעזרת K-Means.
בקלאסטר יש 3 דוגמאות אימון [בצורה $((x_1, x_2), t)$]:
$C_1 = \{((1,1),1), \quad ((-1,-1),0), \quad ((2,2),0)\}$
השתמשו בנוסחת ה- WCV (בנוסחה יש נורמה 2 בריבוע):

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} ||x_i - x_{i'}||_2^2$$

נורמה 2:

$$\|V\|_2 = \sqrt{\sum_{j=1}^{n} v_j^2}$$

חשבו את ציון ה- WCV של הקלאסטר

$$WCV(C_1) = \frac{1}{3}(2(1 - -1)^2 + 2(-1 - 2)^2 + 2(2 - 1)^2) = 9.33$$

15. [5] ICD: נתונים 2 קלאסטרים שנוצרו בעזרת קליסטור היררכי.
בכל קלאסטר יש 3 דוגמאות אימון [בצורה $((x_1, x_2), t)$]:
$C_1 = \{((1,1),1), \quad ((-1,-1),0), \quad ((2,2),0)\}$
$C_2 = \{((1,2),1), \quad ((-2,-2),0), \quad ((3,3),0)\}$
בהנחה שפונקציית ה- Dissimilarity היא נורמה 2:

$$\|V\|_2 = \sqrt{\sum_{j=1}^{n} v_j^2}$$

חשבו את מדד ה- ICD (אי-הדמיון) בין הקלאסטרים, בשיטת Minimal
מדד ה- ICD בשיטת Minimal בין הקלאסטרים הללו הוא 1 מכיוון שהמרחק המינימלי בין 2 נקודות בין הקלאסטרים הוא בין הנקודות $((2,2),0)$ בקלאסטר C_1 ו- $((1,2),1)$ בקלאסטר C_2.
המרחק ביניהן הוא 1.

16. [10] מעוניינים לבצע קלסיפיקציה בינארית בעזרת מודל באסיאני נאיבי (NBC). קבוצת האימון מכילה 2000 דוגמאות חיוביות (+) ו- 1000 דוגמאות שליליות (-). מרחב הקלט הוא דו ממדי. למאפיין x_2 יש 3 ערכים אפשריים: (A,B ו- C). 2 ערכים לא מופיעים כלל ב- D וערך אחד x_2=A מופיע ב- 20% מהבריאים וב- 10% מהחולים. מאפיין x_1 הוא בינארי וידוע כי ב- 50% מהבריאים וגם ב- 50% מהחולים x_1=1.

a. רשום את שערוכי כל ההסתברויות הנחוצות עבור הסקה של NBC בעזרת ספירה בלבד וללא כל תיקון או החלקה

$$P(+) = \frac{2}{3} = 0.66$$
$$P(-) = \frac{1}{3} = 0.33$$
$$P(B|+) = P(B|-) = P(C|+) = P(C|-) = 0$$
$$P(1|+) = P(1|-) = P(0|+) = P(0|-) = 0.5$$
$$P(A|-) = 0.2$$
$$P(A|+) = 0.1$$

b. חשב את שערוכי ההסתברות של P(B|+) כולל החלקת לאפלאס

$$P(B|+) = \frac{n_{B,+} + 1}{n_+ + |V_{x_2}|} = \frac{0+1}{2000+3} = 4.99 \cdot 10^{-4}$$

c. כיצד תסווגו את הדוגמא (1,A) על פי MAP? הסבירו

נשווה

$$P(+) \cdot P(1|+) \cdot P(A|+) = \frac{2}{3} \cdot 0.5 \cdot 0.1 = \frac{1}{30}$$
$$P(-) \cdot P(1|-) \cdot P(A|-) = \frac{1}{3} \cdot 0.5 \cdot 0.2 = \frac{1}{30}$$

לפי MAP יש הסתברות שווה לסיווג (+) ולסיווג (-)

d. כיצד תסווגו את הדוגמא (1,A) בעזרת MLE (Maximum Likelihood Estimation)? הסבירו

נשווה

$$P(1|+) \cdot P(A|+) = 0.5 \cdot 0.1 = \frac{1}{20}$$
$$P(1|-) \cdot P(A|-) = 0.5 \cdot 0.2 = \frac{1}{10}$$

לפי MLE נבחר לסווג (-)