המרכז הבינתחומי
The Interdisciplinary Center

בית ספר "אפי ארזי" למדעי המחשב
The Efi Arazi school of computer science

סמסטר ב' תשע"ו
Spring 2016

מבחן מועד ב בלמידה ממוכנת
Machine Learning Exam B

Lecturer : Prof Ariel Shamir
Time limit : 3 hours
No additional material is allowed for use!

Answer 5 out of 6 from the following questions (each one is 20 points)

מרצה: פרופ אריאל שמיר
משך המבחן: 3 שעות
אין להשתמש בחומר עזר!

יש לענות על 5 מתוך 6 השאלות הבאות
לכל השאלות משקל שווה (20 נקודות)

בהצלחה!
Good Luck!

שאלה 1

ידוע כי מסווג מסוג support vector machine (SVM) בצורתו הפשוטה ביותר פותר את בעיית האופטימיזציה הבאה (המנוסחת בצורה הראשונית =primal):

$$Minimize \frac{1}{2}\|\mathbf{w}\|^2$$
$$subject\ to:\ \forall x^{(d)} \in D, t_d\left(\mathbf{w}\cdot x^{(d)} + w_0\right) - 1 \geq 0$$

א. נתון כי $x^{(d)}$ הוא מופע מקבוצת האימון D. הגדירי מה משמעות כל שאר הסימנים בנוסחאות והסבירי מה מנסה האופטימיזציה להביא לאופטימום?
ב. מה המשותף לאלגוריתם SVM ולאלגוריתם הפרספטרון? הסבירי!
ג. מה המשותף לאלגוריתם SVM ולאלגוריתם KNN? הסבירי!
ד. הסבירי כיצד ניתן להשתמש באלגוריתם KNN לפתרון בעיית רגרסיה
ה. הציעי דרך שבה ניתן להגדיר אלגוריתם SVM לפתרון לבעיית רגרסיה
(רמז: הציעי ניסוח לבעיית אופטימיזציה אותה יש לפתור כדי להגדיר SVM עבור בעיית רגרסיה)

שאלה 2

א. הסבירי מהו ממד רגישות של אלגוריתם למידה?
ב. לאיזה משני המדדים קשורה "רגישות" – bias or variance? הסבירי!
ג. האם עץ החלטות כפי שנלמד בכיתה (עם information gain) הוא אלגוריתם עם רגישות גבוהה או נמוכה? הסבירי!
ד. נניח כי בנינו K עצי החלטה שונים מ-K קבוצות אימון **שונות** הציעי שיטה כיצד ניתן להשתמש ב-K עצים אלה כדי לעשות סיווג של מופע חדש?
ה. האם בשיטה שהצעת ב-ד' הרגישות תעלה או תרד יחסית לעץ החלטה יחיד? הסבירי!
ו. בהינתן קבוצת אימון עם m מופעים ו-n תכונות, כאשר m אינו גדול (כלומר m/K הוא מספר קטן מידי לבנות עץ החלטה), הציעי שיטה לבנות K עצי החלטה שונים בעזרת קבוצת אימון זו.

שאלה 3

משתמשים באלגוריתם k-means עם Euclidean distance לקבץ (cluster) את 6 הדוגמאות הדו-מימדיות הבאות ל-2 מרכזים: A1=(1,7), A2=(4,7), A3=(5,5), A4=(7,3), A5=(8,1), A6=(6,1)

א. צייירי מרחב דו מימדי בגודל 10 על 10 (מ-0 עד 10 בכל ציר) וסמני את 6 הדוגמאות
ב. מלאי את טבלת **רבועי** המרחקים (בלי להוציא שרש) בין כל זוג דוגמאות. שימי לב כי אפשר להשתמש ברבוע המרחקים באלגוריתם k-means במקום המרחק האוייקלדי האמיתי כי יחס הסדר בין המרחקים ישמר:

	A1	A2	A3	A4	A5	A6
A1	0					
A2	X	0				
A3	X	X	0			
A4	X	X	X	0		
A5	X	X	X	X	0	
A6	X	X	X	X	X	0

ג. בהנחה שבאיתחול האלגוריתם k-means אנו בוחרים את שני המרכזים בתור A1 ו-A3. הראי את החישובים בהרצת האיטרציה הראשונה של האלגוריתם ומצאי את שתי הקבוצות (clusters) שנוצרו ואת המרכזים החדשים.
ד. המשיכי להריץ את החישובים באלגוריתם ומצאי את הקבוצות והמרכזים לאחר כל איטרציה עד התכנסות. כמה איטרציות נדרשו עד להתכנסות?
ה. הסבירי כיצד ניתן למדוד את טיב הקיבוץ שנוצר?
ו. האם הקיבוץ שנוצר הוא אופטימאלי? אם כן הסבירי מדוע! ואם לא – הסבירי כיצד ניתן יהיה לשפר את הקיבוץ ותני דוגמא?

שאלה 4

נתונה קבוצת אימון עם m מופעים שלהם n תכונות. אנו רוצים לבנות מסווג בינארי (בין שתי קבוצות A ו-B) מסוג MAP

א. הסבירי איזה הסתברויות עלינו להעריך כדי לסווג מופע חדש x? (כולל נוסחה)
ב. הסבירי מה ההנחה הנאיבית ב-naïve bayes ומדוע צריך אותה? (כולל נוסחה)
ג. הסבירי כיצד מחושבות ההסתברויות שבסעיף א בשיטת parzen window ? (כולל נוסחה)
ד. האם שיטת parzen window גם היא נאיבית? הסבירי!
ה. האם יתכן שהסתברות כלשהי ב-א תהיה 0 בשיטת naïve bayes פשוטה? אם כן הסבירי מתי ואיך פותרים זאת. אם לא הסבירי מדוע.
ו. האם יתכן שהסתברות כלשהי ב-א תהיה 0 בשיטת parzen window? אם כן הסבירי מתי ואיך פותרים זאת. אם לא הסבירי מדוע.

שאלה 5

א. הסבירי מהן הבעיות בלמידה במימדים גבוהים של מרחב הדוגמאות (דוגמאות עם הרבה מאוד תכונות או features) ?
ב. האם יש פתרון אופטימאלי לבחור תת קבוצה של תכונות בו פתרון בעיית סיווג תהיה טובה ביותר? אם כן הסבירי מהו הפתרון, אם לא הסבירי מדוע.
ג. תארי בקיצור את שתי הגישות המקובלות לפתרון בעיית למידה במימדים גבוהים בלמידה חישובית
ד. הסבירי מה ההבדל בין הגישות של סעיף ג לשיטה של סעיף ב.
ה. הסבירי מהם מרכיבים ראשיים (Principal Components), כיצד הם נבחרים באלגוריתם PCA וכיצד הם עוזרים לפתרון הבעיה.
ו. במשימת סיווג, איזו בעיה יכולה להיווצר מכך ש PCA-אינו מתייחס למידע של ה-class בקבוצת האימון? תני דוגמא (ניתן בציור) לקבוצת אימון בה תיווצר בעיה זו.

שאלה 6

א. הגדירי מהו training error ומהו true error?
ב. הסבירי מהו מצב overfitting באלגוריתם למידה?
ג. אנו רוצים אומדן ל-true error באלגוריתם למידה כלשהו. הציעי שתי שיטות שונות לבנות אומדן כזה: שיטה אחת כאשר גודל קבוצת האימון קטן ושיטה שנייה כאשר גודל קבוצת האימון גדול.
ד. נניח שהמדד שבחרנו ב-ג נקרא test error. הסבירי מה יקרה לטעות זו בתלות בגודל קבוצת האימון (מה קורה ל-test error ככל שקבוצת האימון הולכת וגדלה?) הסבירי מדוע! הסבירי מה יקרה ל-training error בתלות בגודל קבוצת האימון? הסבירי מדוע!
ה. הסבירי כיצד ניתן להשתמש במידע בסעיף ד כדי לדעת האם אנו במצב של overfitting ומה (אולי) ניתן לעשות כדי לפתור זאת?

בהצלחה