הנדסת נתונים ומידע
תרגיל בית 2
תאריך הגשה 11.6.17

מטרות התרגיל

במישור של הנדסת תוכנה תרגול שימוש בירושה ופולימורפיזם.
במישור של הנדסת נתונים: תרגול ומימוש אלגוריתמי סיווג שנלמדו בהרצאות.

נתונים

בתרגיל זה נעבוד עם נתונים רפואיים:
1. מאגר של נתונים על חולי סכרת (**מאגר 1**). התיאור של הנתונים הוא בקובץ pima-indians-diabetes.names. הנתון אותו אמור המסווג ללמוד הוא האם האדם הוא חולה סכרת או לא.
2. מאגר של נתונים על מחלת לב (**מאגר 2**). התיאור של הנתונים הוא בקובץ heart.doc. כאן המטרה היא לחזות מחלת לב.

שימו לב ששני קבצי הנתונים עברו נרמול Min-max וכל המשתנים נמצאים על סקלה זהה.

חלק רטוב + ניתוח תוצאות (100%)

על מנת להתחיל, עליכם ליצור פרויקט חדש ולייבא אליו את הקוד שניתן לכם. וודאו כבר בשלב זה שהסטנדרט של C++ שאתם עובדים אתו הוא 98.

רשימת המחלקות שיש להגיש עם התרגיל (**יש להקפיד מאוד על השם המדויק של הקבצים**):

בנוסף יש להגיש את קובץ ה main **מבלי לערוך בו שינויים**.

שם המחלקה	המטרה של המחלקה
Tests	מחלקה שמכילה את הבדיקות של התוכנית. אין לשנות את קוד המחלקה, פרט לפונקציה metaClassifierTest. עליכם להבין מהם הדרישות של הקוד הזה מהקוד שאתם צריכים לממש
Classifier	מחלקה אבסטרקטית המייצגת מסווג כלשהו
KNNClassifier	מחלקה שמייצגת מסווג מסוג KNN
Perceptron	מחלקה שמייצגת מסווג מסוג Perceptron
Point	מחלקה שמייצגת נקודה
DataReader	מחלקה שאחראית על קריאת נתונים בפורמט: Feature1,Feature2,…,FeatureN,class
Distance	מחלקה שמייצגת מטריקת מרחק כלשהי לשימוש על ידי המסווג KNN
EuclideanDistance	מחלקה שמייצגת מטריקה אוקלידית
ManhattanDistance	מחלקה שמייצגת מטריקת L1 (מנהטן)
MetaClassifier	מחלקה שמייצגת מסווג שלכם שמשתמש במסווגים שלמדתם

1. ממשו את המסווג KNN כפי שנדרש בבדיקות *TestKNNEuclidean* ו *TestKNNManhattan* . על מנת לעשות זאת קחו את המחלקה ממעבדה 6 והבינו מהם השינויים הדרושים. לפניכם פירוט הדרישות עם רמזים למימוש:

1.1 מימוש היררכיה של מחלקות המרחק (10 נקודות).

1.2 מימוש KNN כיורש ממסווג כללי. (5 נקודות)

1.3 למחלקה KNN צריכה להיות האפשרות לעבוד עם כל מרחק מבלי לדעת מהו בדיוק. (5 נקודות)

1.4 הוסיפו פונקציה PrintResults במחלקת Classifier אשר מקבלת את סט הבדיקה לאחר התיוג ע"י המסווג ומדפיסה את עבור כל נקודה: את מספרה הסידורי בסט הבדיקה, את התיוג שלה ע"י המסווג (prediction) ואת התיוג האמתי שלה (class), מופרדים בפסיקים . היעזרו בקבצי הפלט בסעיף הבא לצורך הספקת הפורמט המדויק של הפלט. (3 נקודות)

1.4 יש לבדוק את הנכונות הטסטים *TestKNNEuclidean* ו *TestKNNManhattan* לפי הטבלה הבאה (כל טסט שווה 3 נקודות, סה"כ 12 נקודות):

קלט	פלט
מאגר סכרת TestKNNManhattan	diabetes_KNN_M
מאגר מחלת לב TestKNNManhattan	heart_KNN_M
מאגר סכרת TestKNNEuclidean	diabetes_KNN_E
מאגר מחלת לב TestKNNEuclidean	heart_KNN_E

2. מימוש ה Perceptron. הבינו מה המטרות של המתודות הנתונות לכם וממשו את המתודות החסרות. **שימו לב ש class של האובייקט הראשון בסט האימון קובע מיהו ה class החיובי ומיהו ה class השלילי (זה לא אילוץ של האלגוריתם אלא בחירה שלי). למשל כתוצאה מכך עבור קובץ הסוכרת מי שמקבל class=1 הם אנשים בלי סכרת ומי שמזוהה עם class=-1 הם דווקא אנשים עם.** הדבר אינו משפיע על טיב ההפרדה שמוצא המסווג, אך כן משפיע על הסימנים של המשקולות (עלולים להתקבל משקולות זהים בגודלם, אך הפוכים בסימנם).

שימו לב כי לצורך הבדיקה של האלגוריתם המשקולות ההתחלתיים מוגרלים רנדומלית, אך ה srand מאותחל לערך שאין לשנותו, כדי הפלט שלכם יהיה תואם לזה שניתן לכם.

2.1 לאחר מימוש ה Perceptron נבדוק אותו תחילה על הפונקציות הלוגיות and ו xor. הראשונה פרידה לינארית בעוד השנייה לא. הריצו את הטסט *testPerceptronLogic* (12 נקודות).

קלט	פלט
and.data	and.out
xor.data	xor.out

2.2 שרטטו גרפים עבור הפונקציות הלוגיות and ו xor בהם ציר ה x הוא l_1 וציר ה y הוא l_2 ובו מופיעים הנקודות של כל class ו decision boundary שנלמד (6 נקודות)

2.3 הריצו את האלגוריתם עבור הקלטים כמפורט בטבלה וראו שמקבלים פלטים זהים (6 נקודות):

קלט	פלט
מאגר סכרת testPerceptron	diabetesPerceptron
מאגר מחלת לב testPerceptron	heartPerceptron

2.4 הסתכלו על המשקולות שהאלגוריתם למד. מהם 2 המשתנים **הכי משפיעים** על הסיווג לפי מה שנלמד **בכל אחד ממאגרי הנתונים** ומהם 2 המשתנים **הכי פחות משפיעים**. דווחו את שמות המשתנים לפי קבצי התיאור של המאגרים. הסבירו כיצד קבעתם זאת ומהי כיווניות הקשר בין המשתנה ל class (8 נקודות)

2.5 הסבירו מדוע אין צורך לקבוע את הפרמטרים של Perceptron באמצעות validation set (4 נקודות)

3. ניתוח תוצאות המסווגים

3.1 באמצעות הפלטים שקיבלתם אפשר היכן טעה כל אחד מהמסווגים. לכל מסווג דווחו מטריצת בלבול (confusion matrix) . (9 נקודות).

3.2 סכמו בטבלה את הביצועים של כל המסווגים עבור שני המאגרים. (4 נקודות)

3.3 הסבירו את תהליך אימון ה KNN בפונקציה evaluateKNN (6 נקודות)

4. מימוש מסווג שלכם

4.1 עליכם להוסיף מחלקת MetaClassifier **שהוא מסווג** שאתם מחליטים כיצד הוא מסווג את הנתונים בסט הבדיקה. הדרישה המינימלית היא שהקוד בקובץ הבדיקות יתקמפל וירוץ ושהמסווג יעשה שימוש <u>לפחות</u> באחת מהשיטות בתרגיל זה. רצוי שהשיטה שאתם מציעים לסיווג תהיה הגיונית ותעיד על הבנת החומר. **יש להקפיד על כללי הנדסת תוכנה**. (10 נקודות)

4.2 בונוס (אין חובה על ביצוע סעיף זה):

שיפור ביצועי ה MetaClassifier מעל ביצועי המסווג הטוב ביותר (לכל מאגר) מאלו שדיווחתם בסעיף 3.2. **רצוי** להציע שיטה כללית (ולא ספיציפית למאגר מסוים) ולממש אותה בתוך ה MetaClassifier, אם הדבר דרוש מותר לכם לשנות את פונקצית הבדיקה של ה MetaClassifier בקובץ הבדיקות לצורך בחירת מודל (כמו עם KNN) אם תחליטו שה MetaClassifier שלכם צריך זאת.

<u>מה מותר לעשות (+רמזים):</u>

- להשתמש בכל שיטה שלמדנו בקורס (לא רק בקשר לסיווג)
- להוסיף עוד פונקציות למחלקות הקיימות לפי הצורך (ורק לפי הצורך)
- לעשות שילוב של תוצאות המסווגים השונים שכבר מימשתם בתרגיל זה ואפשר גם להשתמש בעקרון הפעולה של Rocchio שנלמד בכיתה
- לחשוב כיצד ניתן להשתמש בידע של משקולות ה perceptron לצורך שיפור ה KNN
- לחשוב כיצד ניתן לבדוק האם KNN או ה perceptron "בטוחים" בסיווג שהם נותנים וכיצד ניתן להשתמש בזה
- להוריד/להוסיף משתנים לנתונים **<u>אחרי</u>** שהם מתקבלים לאימון ע"י יצירת סט נקודות אימון "חדשות" במימד המתאים. לדוגמא: אתם חושבים שרמת הסוכר בדם אינה תורמת לאבחון סכרת- ניתן לא להשתמש במידע זה יותר, אלא שהתאמה למאגר מסוים עלולה לפגוע בביצועים במאגר השני, כי אותו קוד ירוץ על שניהם (אז צריך

לחשוב על פתרון כללי יותר). ניתן גם להוסיף משתנים שהם צירוף של המשתנים שנתונים כבר. למשל להוסיף משתנים שהם מכפלות של כל זוג משתנים נתונים, כאשר מכפלה תופסת קשר בין שני משתנים – מה שהמסווג לא מקבל בצורה מפורשת.

- ועוד הרבה....

מה אסור לעשות:

בגדול אסור לרמות. למשל לשנות את נתוני הקלט או את חלוקת הנתונים לסט אימון וסט בדיקה, לאמן על סט בדיקה. אסור להשתמש במידע חיצוני. למשל הגדלת נתוני האימון ברוב המקרים תשפר את הביצועים, אבל זה לא יהיה הוגן ביחס לביצועי המסווגים שכבר דיווחתם. שימו לב שאין איסור להתאים את המסווג למאגר מסוים, אבל רצוי שהשיטה תהיה כללית, חכמה ומוסברת היטב. אם אתם לא בטוחים --- שאלו בפורום (או במייל מפאת שמירת סודיות הפתרון המקורי מאוד שלכם ☺)

כיצד יקבע ציון הבונוס:

גובה הבונוס מוגבל ל 10 נקודות בציון תרגיל זה והציון הכולל בתרגיל מוגבל ל 100.

1. שיטה מקורית וחכמה תזכה לבונוס לפי שיקול הבודק (על השיטה עצמה). יש להסביר את השיטה בכתב בקובץ התשובות.
2. שיפור בביצועים על כל מאגר יזכה בבונוס ביחס ישר לגודל השיפור היחסי מעל לביצועי המסווג הכי טוב שדיווחתם בסעיף 3.2. מהו בדיוק היחס יקבע בזמן הבדיקה.

דרישות טכניות

יש לבדוק את זהות הפלטים שלכם לאלה שניתנו לכם באמצעות פקודת diff – זלזול בדרישה זו יביא להורדת ניקוד משמעותית

תיעוד הקוד

מעל כל פונקציה בקוד יש לציין:

1. מה המטרה של הפונקציה
2. מהם משתני הקלט (שאותם הפונקציה חייבת לקבל כדי לבצע את מטרתה)
3. מהם משתני הפלט (שהם התוצאה של הריצה של הפונקציה)

דוגמה לפורמט של התיעוד ישנה במודל. **יש לתעד רק בקבצי H**

דגשים למימוש

הקפידו על:

1. כתיבת קוד בהתאם למוסכמות שפורסמו באתר.
2. העברה נכונה של פרמטרים לפונקציות.
3. שימוש נכון ב const.
4. איתחול משתנים.
5. תיעוד בהתאם לדרישות.
6. הרשאות נכונות למתודות ומשתנים במחלקות.
7. חלוקה של הקוד לפונקציות ומחלקות.
8. הקפדה על כללי הירושה והקישור הדינמי.
9. ניהול זכרון נכון והקפדה על בניה והריסה של אובייקטים.
10. מתן שמות משמעותיים למשתנים ופונקציות.

אופן בדיקת התרגיל

התרגיל יבדק בדיקה יבשה ורטובה. חלק מהקלטים והפלטים של התרגיל אינם נתונים לכם (כלומר יש לדאוג לנכונות האלגוריתמים מעבר למה שבדקתם כאן). **תרגיל שאינו מתקמפל עם הקומפיילר של CLion ב Linux – חלקו הרטוב לא יבדק ולא יבדקו התשובות לשאלות היבשות שהן תוצאה של הרצה של הקוד.**
יש לוודא את פורמט הפלטים ע"י diff כמתבקש.

<u>הגשת התרגיל</u>

- לצורך התרגיל יפתח פורום במודל. **מתן תשובות לשאלות בנוגע לתרגיל יתבצע דרך הפורום בלבד.**
- התרגיל להגשה בזוגות או ביחידים (ולא בשום הרכב אחר)
- לפני ההגשה, חובה לקמפל ולבדוק את התרגיל במעבדת הוראה ולא בסביבה אחרת .
- ההגשה חייבת להכיל קובץ **ZIP** יחיד בלבד (ולא קובץ RAR וכדומה) המכיל: **כל קבצי קוד המקור, ללא קבצי הרצה וללא קבצי הפרויקט. אין ליצור מבנה תיקיות בתוך ה ZIP – יש לכווץ את הקבצים בלבד, כולם ביחד.**
- בקובץ ZIP זה יש לכלול מסמך word או pdf המכיל את התשובות לחלק היבש. יש לציין ברור את מספר השאלה עליה ניתנת התשובה. יש להדביק את הגרפים שביקשתי בתוך קובץ התשובות ולא להגישם בקבצים נפרדים.
- שם הקובץ חייב להיות , **hw1_yyyyyyyyyy_xxxxxxxxxx.zip** כאשר xxxxxxxxxx ו - yyyyyyyyyy הם מספרים תעודות הזהות של המגישים, כולל ספרת ביקורת .
- ההגשה היא אלקטרונית בלבד, דרך אתר ה moodle-של הקורס. תרגילים שיוגשו בכל דרך אחרת לא ייבדקו.
- אין להגיש את אותו הקובץ פעמיים. התרגיל יוגש על ידי אחד מבני הזוג.
- שימו לב שההגשה תיחסם בדיוק בשעה 23:55 ביום ההגשה. מומלץ להגיש לפחות שעה לפני המועד האחרון.
- ניתן להגיש כמה פעמים. רק ההגשה האחרונה תישמר.
- תרגיל בית שלא יוגש על פי הוראות ההגשה - לא ייבדק (כלומר יקבל ציון אפס).

<u>בהצלחה !!!</u>