Introduction to Data Science

Course 094201

Lab 6:

KNN & Evaluation of classifiers

Spring 2017

מטרות המעבדה

במעבדה זו תדרשו לעבוד עם מסווג KNN שנלמד בהרצאה
מטרות:

הבנת אופן עבודת ה KNN
הבנת תהליך cross-validation

בהיבט של הנדסת תוכנה:
שימוש באלגוריתמים גנריים של STL ועוד שאלות

The dataset and the code

- The code and the data can be found at:

/mnt/share/students/LAB6

- Copy everything to your **local folder** and unzip the code:

unzip lab6-students.zip

- Your dataset is called: ecoli.processed

This dataset is for learning the location of proteins inside the bacterium cells of ***Escherichia coli*** . Localization of proteins in cells has an important role in biomedical research.

References:

"Expert Sytem for Predicting Protein Localization Sites in Gram-Negative Bacteria", Kenta Nakai & Minoru Kanehisa, PROTEINS: Structure, Function, and Genetics 11:95-110, 1991.

Reference: "A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells", Kenta Nakai & Minoru Kanehisa, Genomics 14:897-911, 1992.

The code – what do we have?

1. בקוד הניתן עם התרגיל ישנן המחלקות הבאות:

מחלקה	תיאור
KNN	המסווג KNN
Evaluation	המחלקה האחראית על העבודה עם המסווג לצורך הערכת הביצועים שלו
DataReader	המחלקה שאחראית על קלט הנתונים. שימו לב מהו הפורמט הנקלט
EvaluationMeasures	מחלקה שאמורה להכיל מדדים לאיכות הסיווג, עד כה למדנו רק accuracy
Point	מחלקה המייצגת אובייקט יחיד לסיווג. שימו לב שאתם מבינים מהם משתני המחלקה ותפקידם

Run the code

- הקוד הינו במצב מוכן להרצה וכדי להתנסות בו אליכם לקרוא לפונקציות המתאימות
- שימו לב: התוכנה מקבלת פרמטר יחיד והוא שם קובץ הקלט
- שאלה1: כתבו פונקציה בקובץ main היוצרת מסווג KNN עם חמישה שכנים הריצו אימון עם כל הנתונים וסווגו את האיבר הראשון בקובץ הנתונים. בידקו מה הפלט של המסווג ומהו ה class הנכון של האובייקט.
- שאלה 2: כתבו פונקציה בקובץ main המריצה 10 fold cross-validation . הדפיסו את הפלט. הבינו את משמעותו.

Assignment – Home - for submission -

בבית עליכם לבצע את המשימות הבאות:

1. כתבו פונקציה בקובץ main המאמנת את המסווג 1NN על כל נתוני הסט ולאחר מכן מסווגת את כל נתוני הסט. דווחו accuracy . הסבירו מה קיבלתם ולמה.
2. כתבו פונקציה בקובץ main היוצרת בלולאה מסווגים עם k שכנים כאשר $1 \leq k \leq 30$. לכל מסווג מריצה leave-one-out cross-validation ומדווחת תוצאות accuracy. מיהו המסווג הטוב ביותר?
3. בפונקציה המבצעת cross-validation הוסיפו הדפסה של איכות ה testing עבור כל אחת מה testing folds. בצעו זאת עבור המסווג הנבחר בשאלה הקודמת.

   כעת הריצו 2-fold-cross-validation ושימרו את הפלט (תוצאה א)

   10-fold-cross-validation ושימרו את הפלט (תוצאה ב)

   leave-one-out cross-validation ושימרו את הפלט (תוצאה ג)

   היכן התקבלו ביצועים עקביים (בעלי שונות נמוכה) על החלוקות השונות? מהי כמות הfolds שהייתם ממליצים עליהם בהינתן שאתם עובדים עם קובץ נתונים קטן יחסית ומדוע?
4. הסבירו מדוע יש צורך במיון רנדומי של הנתונים לפי תהליך ה cross-validation?

C++ questions at home

יש לענות על השאלות הבאות:

1. הסבירו מהן המטרות של שני האופרטורים [] שמומשו ב Point. מדוע לא ניתן להסתפק באחד בלבד?
2. בקוד יש שימוש באלגוריתמים גנריים הבאים של הספריה הסטנדרטית (STL):

std::copy, std::max_element, std::random_shuffle

קראו את תיעוד האלגוריתמים הללו באינטרנט. מה משותף להם? רמז: התייחסו לאופן הקריאה

3. מהו std::pair שנמצא בשימוש ב KNN – קראו ברשת והסבירו במשפט או שניים.
4. הסבירו כיצד ממומשת (מבחינת המגנון) פונקצית predict של KNN?