Introduction to Data Science

Course 094201

Lab 2:

Basic C++ and Data Analysis

Spring 2017

Welcome to CLion

- Copy the following to your local folder and unzip plots.zip :

cp –r /mnt/share/students/LAB2/lab2_for_students your_folder

cp /mnt/share/students/LAB2/plots.zip your_folder

unzip your_folder/plots.zip

- CLion will be our workspace for C++ this semester
- Open CLion (Applications->Programming->CLion)
- Open the project you just copied.
- Compile the project with CLion.
- Open the folder of the project and find the executable file.
- Run the executable outside of CLion (like in lab1).

The Data

- **<u>רקע:</u>**
- לרשותכם קובץ נתונים על איכות היין.
- הנכם מעוניינים לבצע חקר של הקשרים בין המשתנים השונים בקובץ: כגון רמת חומציות, רמת הסוכר והאלכוהול וכו'. איכות היין (המשתנה quality) היא מדד סובייקטיבי הניתן על סמך שיפוט אנושי.
- שימו לב שכאשר מדובר בבעיות חיזוי סטטיסטיקאים נוטים לחלק את המשתנים (variables) למשתנה תלוי (dependent) - שזה אותו משתנה שמנסים לחזות ומשתנים מסבירים (explanatory). למשל אם המטרה הייתה חיזוי איכות היין, אזי המשתנה של האיכות היה התלוי ומשתנים אחרים בקובץ היו המסבירים. לעיתים נקרא המשתנה התלוי גם response variable .
- אנשי Machine Learning משתמשים במינוח אחר (בד"כ) למשתנה התלוי קוראים class או label או outcome variable. ואילו למשתנים המסבירים קוראים: features (תכוניות) או predictors. כאמור לעיתים משתמשים במושגים אלו במעורב.

Reading the data

- בטרם אתם עובדים עם קובץ נתונים כלשהו מומלץ לאסוף עליו כמה שיותר מידע
- הסתכלות על השורות הראשונות והאחרונות של הנתונים כבר נותנת לנו קצת מידע על פורמט הקובץ (מידע מצומצם כמובן)
- אתם יכולים לעשות זאת על ידי שימוש בפקודות bash: head ו tail שלמדנו במעבדה הקודמת
- בקובץ main.cpp מצויה פונקציה שטוענת את קובץ הנתונים
- ראו שאתם מבינים מהם מבני הנתונים שבהם משתמשים בתרגיל
- שימו לב לאופן העברת הפרמטרים לפונקציה
- הבינו מתי הפונקציה מחזירה true ומתי היא מחזירה false ולמה הבדיקות הללו נחוצות (ולמה זה לא מספק עבור קובץ חדש)

Debug with CLion

debug

Break point

Compile

Step in

Step over

Step out

Variables

lab2 – [~/CLionProjects/lab2] – .../main.cpp – CLion 2016.3.3

```
28                  getline( s, temp, ',' );
29                  dataNames.push_back(temp);
30                  data.push_back(vector<double>());
31              }
32          }
33          else{
34              double val;   val: 7
35              while ( s >> val ){
36                  if (s.peek() == ',')
37                      s.ignore();
38                  vector<double> tempVec((vector<double> &&) data.front());
39                  data.erase(data.begin());
40                  tempVec.push_back(val);
41                  data.push_back(tempVec);
```

Debug: val = {double} 7
s = {std::stringstream}
infile = {std::ifstream}
line = {std::string} "7,0.27,0.36,20.7,0.045,45,170,1.001,3,0.45,8.8,6"
flag = {bool} false
fileName = {const std::string &} "/mnt/share/winequality.csv"

Build finished in 336ms (moments ago)

Debugging with CLion

- Let's see what happens if we want to stop the run in middle and see the current state of our program.

- **Set a few "break points" in main.cpp and debug the program**
- **Set a break points to see :**

1. The value of the 4th feature in the 3th iteration.

_____

2. The name of the 6th feature.

_____

Data Analysis

- The files Statistics.h and .cpp should contain some statistical measures we learned at class
- Implementation of the Median function is given to you, your task is to complete the other measures
- The Median function contains call to standard library (STL) algorithm: sort, we learn more STL algorithms later on

Data Analysis

- Do it yourself: implement the following:
  - Mean: a function that receives a const reference to vector and returns its mean.

Mean $(\bar{x})$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Data Analysis

- Do it yourself: implement the following:

  – Variance: a function that receives a const reference to vector and returns its variance.

The **sample variance**:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$