University of Waterloo
ECE 657A: Data and Knowledge Modelling and Analysis
Winter 2019
Homework 1: Data Summarization
Due: January 10th, 2019 11:59pm

Overview

Collaboration: Do your work and report individually. You can collaborate on the right tools to use and setting up your programming environment.

Hand in: One report per person, via the LEARN dropbox. Also submit the code / scripts needed to reproduce your work. Report as a PDF or a python notebook.

General Objective: To study how to apply some of the methods discussed in class on two datasets. The emphasis is on analysis and presentation of results not on code implemented or used.

Specific Objectives:
- Establish your software stack to carry out data analysis homeworks, assignments and the project for the rest of the course.
- Load a simple dataset and compute some basic statistics and plots.

Tools: You can use libraries available in python or R available to you. You need to mention which libraries you are using, any blogs or papers you used to figure out how to set carry out your calculations.

Data sets

This is your first homework so there are two datasets.

- The Breast Cancer Wisconsin (Diagnostic) Data Set
  - `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`
  - Download from Data Folder link, read data set description.
- Class Anonymous Poll (optional)
  - Fill out the poll first with your own answers `https://goo.gl/forms/iFYBlz9H6QsY9GUS2`
  - Download data from `https://docs.google.com/spreadsheets/d/1fOSqZsINRnx8Pp0pEPhNfuFsMMo3sQNn_lA5UoUnHnU/edit?usp=sharing` (will not be available until poll is closed)

Tasks

In the first class we talked about how to summarize single-variable data, how to compute the Pearson Correlation Coefficient for pairs of points

- In the cancer dataset report the mean, mode and skew, standard deviation and variance values for all the continuous valued features.
- In the cancer dataset s a few pairs of features for correlation by computing their PCC and report the resulting numbers and explain what they mean.
- In the cancer dataset plot two histograms for a continuous valued feature of your choice: One for patients with each diagnosis (M or B).