Midterm (Due July 1)

CS4412: Data Mining
Kennesaw State University
Summer 2022

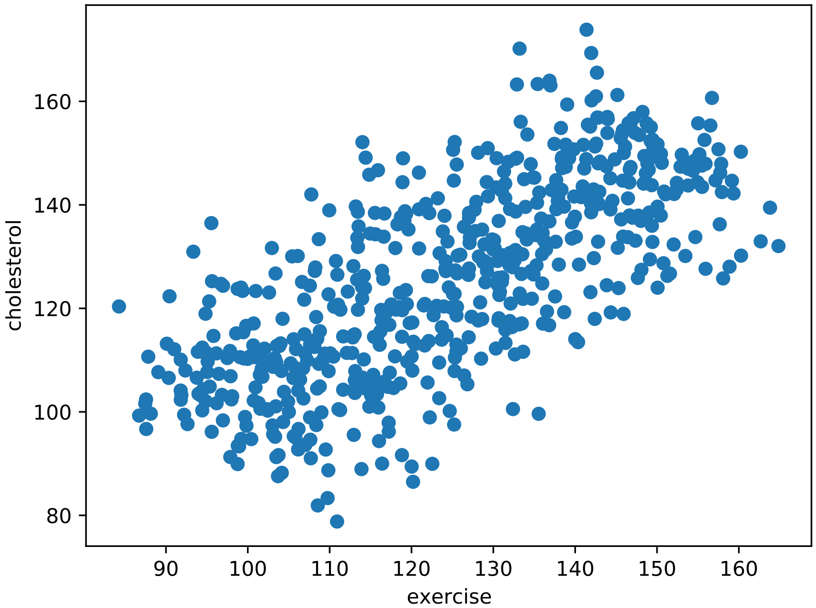
In this midterm, we will examine an example of a paradox, using a simulated dataset of patients, where we have recorded how much exercise they get and their level of cholesterol.

Consider first the dataset file `dataset-one.csv`, which consists of 600 examples, each corresponding to a patient where we have recorded how much exercise they get per week, and their exercise level. Below, we view the first five examples of the dataset:

exercise	cholesterol
113.8885	88.9410
114.8435	101.0020
109.7504	83.3392
106.1830	92.7467
100.7218	102.1816

Note that this is simulated data: a higher number for exercise means the patient performed more exercise, and a higher number for cholesterol means the patient has a higher level of cholesterol (there are no units for these simulated measurements).

Below, we visualize this data using a scatter plot, where each of the 600 patients is plotted as a data point, where the x-axis represents exercise level and the y-axis represents cholesterol level.



Here, we observe that there appears to be a positive correlation between exercise and cholesterol: the more exercise that one performs, the higher your cholesterol level. (Most people would consider this unexpected and counter-intuitive).

Later, we will consider a second dataset file `dataset-two.csv`, which consists of the same 600 examples, except that we include the "age" of each patient, in addition to their levels of exercise and cholesterol.

Questions: Answer the following questions regarding these datasets.

1. Suppose that we learn a linear model $y = ax + b$ to predict a patient's cholesterol level (y) given their exercise level (x). What is the resulting co-efficient a based on the dataset `dataset-one.csv`? Does the co-efficient suggest positive correlation or negative correlation between x and y? (Note that if you use the linear regression module in the python scikit-learn module, this coefficient appears in the `coef_` member variable).
2. Consider `dataset-two.csv` which includes the age of each patient. Suppose that we split the patients into different age groups, in their 20s (20-29), in their 30s (30-39), and so on, up the the age group in their 70s (70-79). Suppose that we again learn a linear model $y = ax + b$ to predict a patient's cholesterol level (y) given their exercise level (x). However, say we learn a different model for each age group, using only that age group's data. What are the resoluting co-efficients a for each of the 6 different age groups in the dataset? Do the resulting co-efficients suggest positive correlation or negative correlation between x and y?
3. The analysis that we performed in part (1) and part (2) uses the same pool of patients, except that in part (2) we incorporated the patients' ages in the analysis. Further, we (should) have observed two different trends in each analysis. Which trend do you intuitively believe should be the correct one, and why? (This is an open-ended question).

Turn in: the answers to the above questions as a text file or as a .pdf file (no Microsoft Word .doc's please), onto the course website under "Assignments" and "Midterm." Midterms are due Friday, July 1 by 11:59pm. Please start early in case you encounter any unexpected difficulties. Late midterms are accepted without penalty until solutions are posted.

Included files:

- `midterm.pdf`: this document
- `midterm.py`: a python script that reads in the datasets, to help you get started
- `data/dataset-one.csv`: the first dataset, over exercise and cholesterol
- `data/dataset-two.csv`: the second dataset, over age, exercise and cholesterol