Homework 1

DS4400: Machine Learning/Data Mining 1

Question 1

Install Python and Jupter Notebook. If you have any problems with python coding, check with TA. Install "sklearn" module in the python (update any module if you need). Check the help page: https://scikit-learn.org/stable/install.html

Question 2 (10 points)

Load the Diabetes Dataset in the sklearn module through Jupter Notebook. Feel free to use the following lines:

- from sklearn.datasets import load_diabetes
- data = load_diabetes()
- X, y = data.data, data.target
- print(data.DESCR)

Read the data descriptions. Write a Paragraph to introduce the dataset. You need to include:

- Topic of the data
- Size of the data
- Which is the target feature
- What kind of plots you want to include in the EDA
- Describe three steps that you think are necessary to pre-process the data

Question 3 (15 points)

Write two matrix A and B. A is a 4 * 2 matrix and B is a 2 * 4 matrix (Fill in an number you like as long as the dimensions match).

- Write your A and B
- Can we calculate $A + B$? If not, why?
- Can we calculate $A^T + B$? If not, why?
- Calculate AB and BA.
- Can we calculate A^{-1}? How about AB^{-1}? Write out the answer if you can.

Question 4 (15 points)

The human genome is composed of the four DNA nucleotides: A, T, G, and C. Some regions of the human genome are extremely G-C rich (i.e. a high proportion of the DNA nucleotides there are guanine and cytosine), and other regions are A-T rich (a high proportion there are adenine and thymine). Suppose a region of the DNA sequence is selected at random from a G-C rich region: 60% of the nucleotides are G-C.

a. What is the probability that two successive nucleotides in this region are G-C?

b. What is the probability that two successive nucleotides in this region are A-T?

c. What is the probability that two successive nucleotides in this region are not the same?

Question 5 (20 points)

The gene that controls white coat color in cats, KIT , is known to be responsible for multiple phenotypes such as deafness and blue eye color.

A dominant allele W at one location in the gene has complete penetrance for white coat color; all cats with the W allele have white coats. There is incomplete penetrance for blue eyes and deafness; not all white cats will have blue eyes and not all white cats will be deaf. However, deafness and blue eye color are strongly linked, such that white cats with blue eyes are much more likely to be deaf. The variation in penetrance for eye color and deafness may be due to other genes as well as environmental factors.

Suppose that 30% of white cats have one blue eye, while 10% of white cats have two blue eyes. About 73% of white cats with two blue eyes are deaf and 40% of white cats with one blue eye are deaf. Only 19% of white cats with other eye colors are deaf.

a. Calculate the probability of deafness among white cats.

Hint: Let B be the event that the cat has one blue eye, let BB be the event that the cat has two blue eyes and let N be the event thta the cat does not have blue eyes. Let D be the event that the cat is deaf. Based on the question, write out: $P(B)$, $P(BB)$, $P(N)$, $P(D|BB)$, $P(D|B)$ and $P(D|N)$. Then use the law of total probability.

b. Given that a white cat is deaf, what is the probability that it has two blue eyes? What is the probability that it does not have a blue eye?