CptS 475/575: Data Science
Mid-Term Exam
November 13, 2019, 9:10–10:00am

Name: Yu-Chieh Wang

WSU ID: 11641327

Instructions:

1. Make sure that your exam has 8 pages (including this cover page) and is not missing any sheets, then write your full name and WSU ID number on this page.
2. There are 7 questions in this exam; some of the questions have multiple parts. The questions are of "short-answer/essay" type. The point each question carries is shown in parenthesis. The points add up to 100. Read each question carefully and give a brief but complete answer in the space provided beneath each question. Please write legibly.
3. The exam is closed book and closed notes.

Question	Points	Your Score
Q1 (a–b)	12	9
Q2	12	10
Q3	14	12
Q4 (a–b)	14	14
Q5	20	18
Q6 (a–b)	14	12
Q7 (a–g)	14	8
Total	100	83

1. (**12 points**) This question has two parts:

(a) (**5 points**) Describe in plain terms what Exploratory Data Analysis (EDA) means.

-1

EDA is a method that we can use on analyze data and get some plot and graph results. Using these results, people can find some relationships between predictors easier.

(b) (**7 points**) Enumerate the sort of things that can be achieved by carrying out EDA.

-2

Through doing EDA, we can get some plots to show the relationship between data. In addition, we can also find the coefficient of each predictors and use some machine learning algorithm to make predictions.

2. (**12 points**). This question has two parts:

a) (**6 points**) Explain briefly (in one or two sentences) the notion of "tidy data". What are the advantages of data being tidy?

-2

Tidy data means that we only keep the information we are going to use and delet the rest of information, which make the data clear and smell.

You should have defined "tidy" data as a data structure to organize data

each observation → row

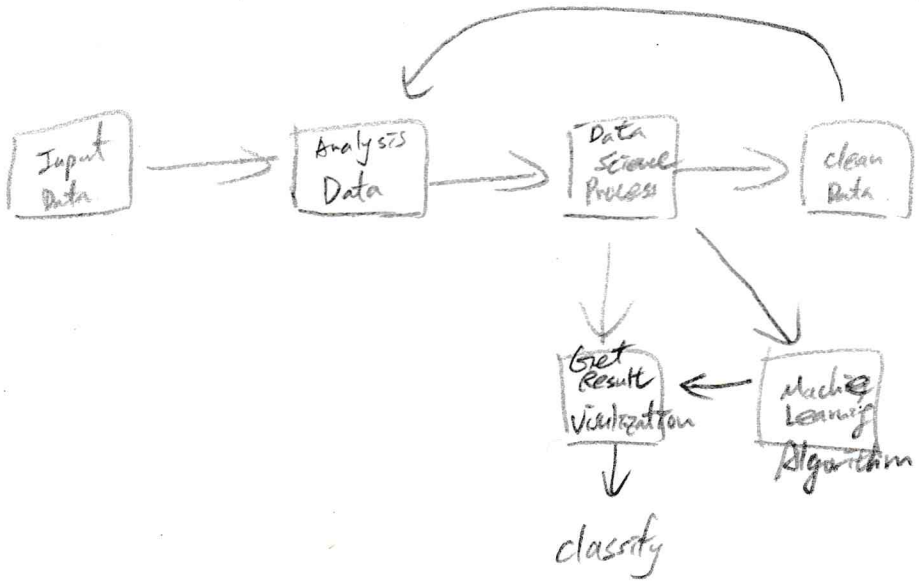
each variable → column

Also you should have discussed its advantages

b) (**6 points**) For many reasons, there could be missing values in data. List a few general methods that can be used for dealing with missing values.

1. Delet the examples with missing values.
2. Count the average value of the column and replace the missing value by the coculate result.

3. **(14 points)** Give a schematic description of the "Data Science Process". Most entries in your drawing will be self-explanatory. Feel free to elaborate on the ones that you think need explanation.



4. **(14 points)** This question has two parts

a) **(6 points)** State clearly and succinctly the difference between these two machine learning classes: supervised learning and unsupervised learning.

Supervised Learning: Use labeled data and assign pattern to train. We usually use it when we know what we are looking for. For example, classify the result is 0 or 1.

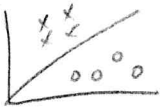
Unsupervised Learning: We use this method when we don't know what we look for, so there is no label data. We just give a computer all data and wish the computer can give us a interesting pattern.

b) **(8 points)** Characterize when a supervised learning problem is called a *classification* problem and when it is a *regression* problem. Give an example of a classification problem. Give an example of a regression problem.

Classification Problem: We only want to get the result as classes. For example, binary class only shows the result as 0 or 1.

Regression Problem: It shows the probebility of the expected result. For example. When we assume a result is 1, and we count the probebility of the result equal to one. If we get the probebility is larger than 0.5, then we can say the result is 1. Otherwise, the result is 0.

The Classification problem is used to classify data



but the regression problem is used to find the relationships between continue data

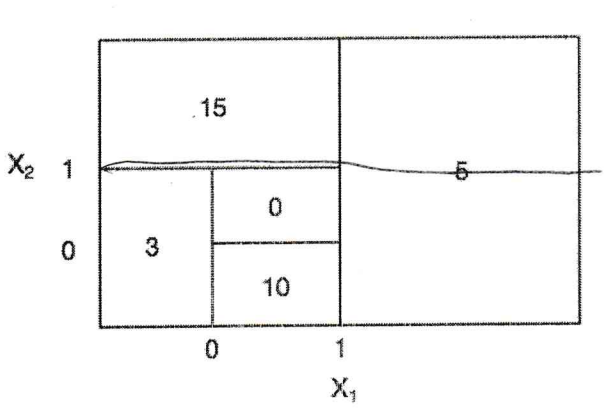
5. **(20 points)** Explain how the k-Nearest Neighbors algorithm works. The k-NN process involves picking an *evaluation metric* (e.g. accuracy). Give one other example of an evaluation metric that could be used in the context of k-NN and define it.

-2

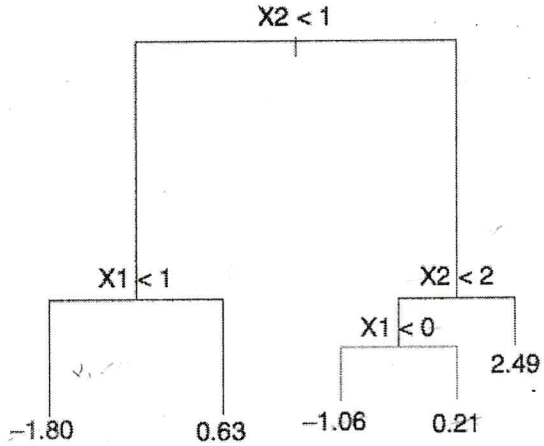
1. Give a distodance matrix
2. split the data to train and validation data
3. Give an evaluation matrix
4. Run kNN for times on train data
5. change k value and run it again
6. Test the validation data
7. Get the accuracy.

Evaluation metric?

6. (**14 points**) This question relates to the plots shown below. It has two parts.



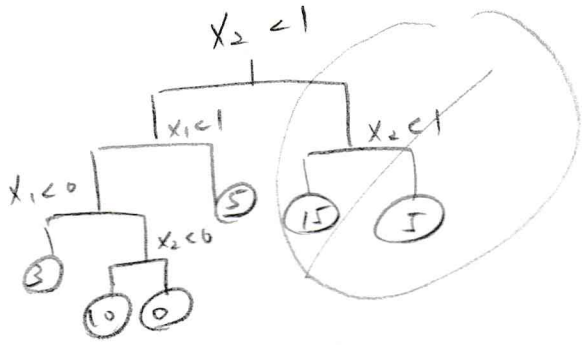
(A)



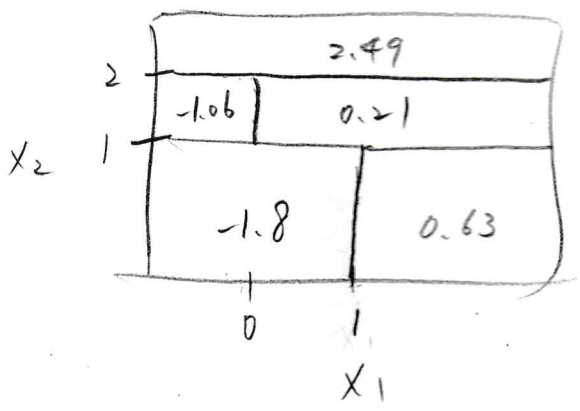
(B)

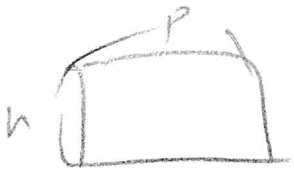
a) (**7 points**) Sketch the decision tree corresponding to the partition of the predictor space illustrated in the left-hand panel of the figure. The numbers inside the boxes indicate the mean of Y within each region.

-2



b) (**7 points**) Create a diagram similar to the left-hand panel of the figure, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.





7. **(14 points)** A number of statements are listed below regarding Principal Components Analysis (PCA) of a data set X with n observations and p features. For each statement, state whether the statement is true or false. If your answer is false, provide a brief justification. (If your answer is true, no justification is needed.)

a) PCA finds a low-dimensional representation of the data set that contains as much as possible of the variation.

b) Each of the dimensions found by PCA is a non-linear combination of the p features in the data set.

False

c) The loading vector ϕ_1 (with elelemnts $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$) correspoding to the first principal component defines a direction in the feature space along which the data vary the most.

d) The loading vector ϕ_2 corresponding to the second principal component is orthogonal to the vector ϕ_1 corresponding to the first principal component.

e) Results obtained when we perform PCA on the data set X are independent of whether or not the variables have been individually scaled.

False : Scaling matters

f) Two different software packages will yield the same principal component loading vectors for the data set X, although the signs of those loading vectors may differ.

g) Consider the first four principal components of the data set X. The third principal component could have a larger *proportion of variance expalined* (PVE) compared to that of the second.

False