San Diego State University **CS5XX Machine Learning**

Assignment -2

Due: Sep 26, 2021 11:59 PM

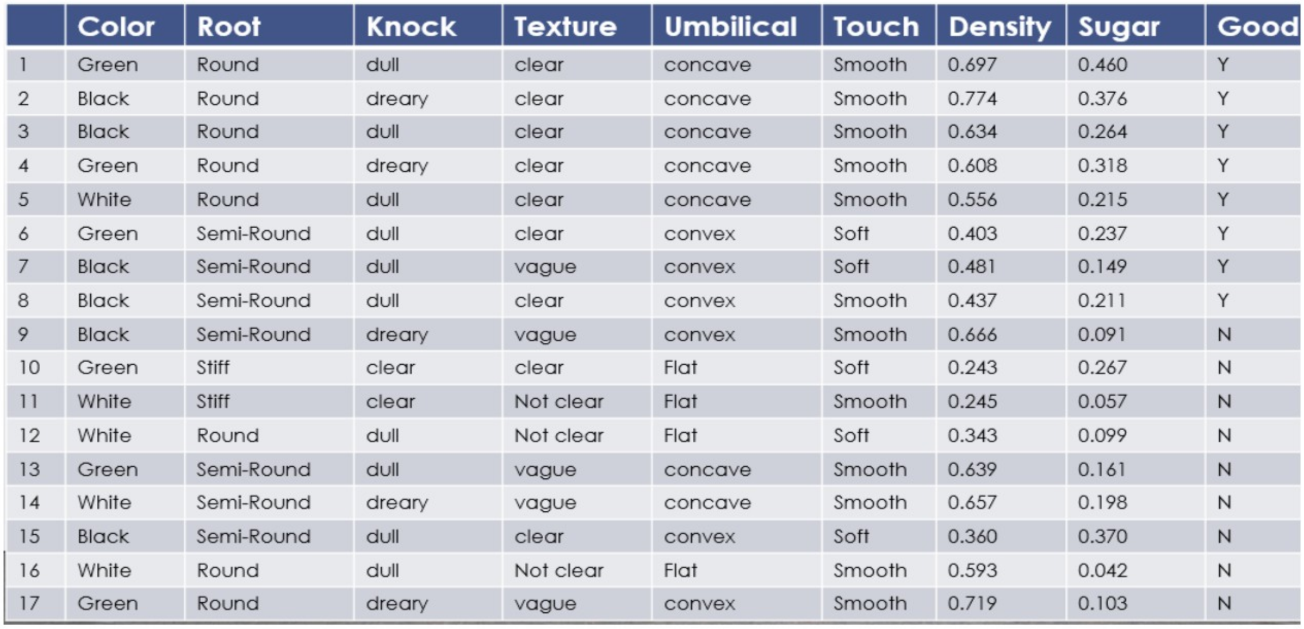
- Please type the solutions using a word processor such as MS Word, Latex, or write by hand neatly and upload the scanned copy of it.

- **I, __Aayushi Dubey______ (sign your name here), guarantee that this homework is my independent work and I have never copied any part from other resources. Also, I acknowledge and agree with the plagiarism penalty specified in the course syllabus.**

- Turn in your assignment before the deadline. Penalty will be applied to late submission.

A. Exercise:

1. Given the watermelon dataset, use MLE to compute the conditional probability for feature Color: $\mathrm{P(C = Black|Good)}$, $\mathrm{P(C = Green|Good)}$, $\mathrm{P(C = White|Good)}$.



	Color	Root	Knock	Texture	Umbilical	Touch	Density	Sugar	Good
1	Green	Round	dull	clear	concave	Smooth	0.697	0.460	Y
2	Black	Round	dreary	clear	concave	Smooth	0.774	0.376	Y
3	Black	Round	dull	clear	concave	Smooth	0.634	0.264	Y
4	Green	Round	dreary	clear	concave	Smooth	0.608	0.318	Y
5	White	Round	dull	clear	concave	Smooth	0.556	0.215	Y
6	Green	Semi-Round	dull	clear	convex	Soft	0.403	0.237	Y
7	Black	Semi-Round	dull	vague	convex	Soft	0.481	0.149	Y
8	Black	Semi-Round	dull	clear	convex	Smooth	0.437	0.211	Y
9	Black	Semi-Round	dreary	vague	convex	Smooth	0.666	0.091	N
10	Green	Stiff	clear	clear	Flat	Soft	0.243	0.267	N
11	White	Stiff	clear	Not clear	Flat	Smooth	0.245	0.057	N
12	White	Round	dull	Not clear	Flat	Soft	0.343	0.099	N
13	Green	Semi-Round	dull	vague	concave	Smooth	0.639	0.161	N
14	White	Semi-Round	dreary	vague	concave	Smooth	0.657	0.198	N
15	Black	Semi-Round	dull	clear	convex	Soft	0.360	0.370	N
16	White	Round	dull	Not clear	Flat	Smooth	0.593	0.042	N
17	Green	Round	dreary	vague	convex	Smooth	0.719	0.103	N

2. The following data set shows the different parameter depending on which a person may or may not buy a computer. Use Naïve Bayes classifier to find out if a person with age <=30, income = medium, student = yes, credit_rating = fair, will or will not buy a computer. Show individual probablity calculated as well as the probablities for the final class.

Age	Income	isStudent	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31-40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31-40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31-40	medium	no	excellent	yes
31-40	high	yes	fair	yes
>40	medium	no	excellent	no

3. The prediction of an Naive Bayes Classifier can be described as

$$\mathrm{h}_{\mathrm{nb}}(\mathrm{x}) = \underset{c \in Y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i|c).$$

   When the data dimension is high ($\mathrm{d} \rightarrow \infty$), the production of the probability of any $\mathrm{c} \in Y$ will quickly coverage to zero. Propose your solution to prevent this from happening.

4. What is sigmoid function and its use in logistic regression? Explain it with a graph. Explain the cost and gradient for logistic regression.

B. <u>Coding Assignment NBC:</u>

In this assignment, we will build an Naive Bayes Classifier to recognize spam emails.
Data: The dataset from Kaggle competition and can be downloaded here: https://www.kaggle.com/c/adcg-ss14-challenge-02-spam-mails-detection/overview. It already splits the training set and the testing set.
Self-evaluation: File “spam-mail.tr” contains the true label for the training set. You can further partition the training set into training and testing parts in order to evaluate the model.
Testing set evaluation: To verify the performance upon the testing set, you can output the predicted labels to a csv file in the format of “Id, Prediction”. Upload the csv file to the link above and finds out your score.