IE590: Homework 2

Instructions: You are supposed to complete Problems 1 manually, while Problems 2-3 should be solved using programming. For questions that need coding, you need to submit your code together with your answer. The format is flexible, but your submission needs to be in an easy-to-follow manner.

Problem 1

In class, we saw that the likelihood associated with logistic regression is:

$$\mathcal{L}(\beta|X, y) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

where $p(x_i)$ is the probability predicted by the logistic model.

1. Compute the log-likelihood function, i.e., $\log \mathcal{L}(\beta|X, y)$.

2. Compute the gradient of the log-likelihood function.

3. Write the logistic regression problem as an optimization problem.

4. Explain why a closed-form solution for the optimal coefficients cannot be reached.

Problem 2

Implement your own KNN function with l_2-norm. You should not use any built-in function. To test your code, you can use the following training and tests data sets: trainKNN.csv and testKNN.csv.
The inputs: data from trainKNN and testKNN

- In the datasets, the first column is your horizontal coordinate (x axis),

- the second column is your vertical coordinate (y axis),

- the third column represents the data label.

The outputs (This is what you need to do):

1. Plot the original data.

2. Plot your classification results after implementing KNN algorithm with K= 1, 2, 5, and 20.

3. Plot your classification results using l_1-norm with K= 1, 2, 5, and 20.

4. What do you observe from these results?

Problem 3

In this homework, you will use classification methods to determine whether you would have survived the Titanic sinking. To find out, we will use the titanic dataset (`titanic_data.csv`), containing the following information about 887 passengers:

1. whether they survived or not (1 = survived, 0 = deceased),

2. passenger class

3. gender (0 = male, 1 = female),

4. age,

5. number of siblings/spouses aboard,

6. number of parents/children aboard, and

7. fare.

	Passenger 1	Passenger 2	Passenger 3	$\cdots$	Passenger 887
Survived	0	1	1	$\cdots$	0
Passenger Class	3	1	3	$\cdots$	3
Gender	0	1	1	$\cdots$	0
Age	22	38	26	$\cdots$	32
Siblings/Spouses	1	1	0	$\cdots$	0
Parents/Children	0	0	0	$\cdots$	0
Fare	7.25	71.2833	7.925	$\cdots$	7.75

Table 1: Titanic dataset sample

Our goal is to construct a classifier that determines/predicts whether an individual would survive or not. The models that you need to use:

- Logistic Regression

- Linear Discriminant Analysis

- Quadratic Discriminant Analysis

- Naïve Bayes

This is what you need to do:

1. Briefly introduce each of the models you are using.

2. Randomly split the data into training data (80%) and testing data (20%). Implement each of the methods using any software. You can use built-in functions.

3. Provide the training and testing errors for each of the models. Discuss your results and conclude.

4. Build your own feature vector $\mathbf{x}$. Would you have survived the Titanic sinking? Briefly explain your conclusion.