San Francisco Bay University

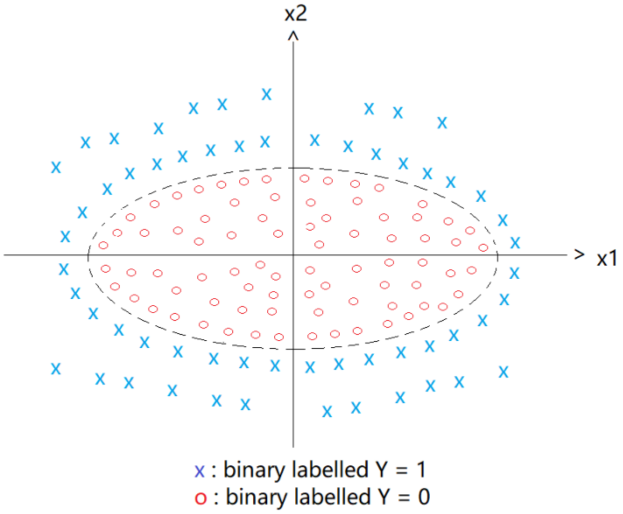
CS483 - Fundamentals of Artificial Intelligence
2022 Summer Quiz #1

Student Name: Khoi Duong **Student ID: 19610**

Instruction:

A. Put your answer right after each question in the answer sheet

1. Assuming two features ($x_1$and x_2) binary labeled Y dataset and the boundary line of classification shown as follows, try to figure out the hypothesis function, loss function, and cost function for the training model and explain why



Since we have to classify data with two kinds of labels, this is a binary classification. Furthermore, we have 2 features ($x_1$and x_2), which means the classification is based on both 2 features. Thus, we can use logistic regression with 2 features for this dataset. Hence, we have a hypothesis function as below:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Where g(z) is a logistic function, which can be shown as $g(z) = \frac{1}{1 + e^{-z}}$

We have the cost function as below:

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if} \quad y = 1 \\ -\log(1 - h_\theta(x)) & \text{if} \quad y = 0 \end{cases}$$

When combined together, it forms a cost function like this:

$$Cost(h_\theta(x), y) = -y\log\left(h_\theta(x)\right) - (1 - y)\log\left(1 - h_\theta(x)\right)$$

We have the loss function as below:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i\log(p_i) + (1-y_i)\log(1-p_i))$$

Where p_i is the probability of class binary labelled Y = 1, and $1 - p_i$ is the probability of class binary labelled Y = 0.

2. Given 1,000 samples dataset with more features and labeled Y's values, 80% training set, and 20% validation set will be taken as splitting into data preprocessing. If the training model takes the KNN algorithm, what is K's value to create the overfitting? Explain why, and how to avoid it. Based on the rule of thumb, what is the appropriate K's value to make the error rate minimum?

80% training set = 800 training samples
20% validation set = 200 test samples
If the training model takes the KNN algorithm, a small K's value can create the overfitting. Overfitting occurs when a model presents very well with the training set but becomes significantly less accurate when the new data from the validation set is added in. The model will do much better on the training set than the test set. In this case, if K's value is small (e.g. K's value equals to 1,2, or 3), it can create the overfitting.
Based on the rule of thumb, K's value should equal to sqrt(N) or sqrt(N/2), where N is the number of test set. In this case, there are 200 test sets. Therefore, K's value should be:
$K = \sqrt{200} = 10\sqrt{2}$ or $K = \sqrt{200/2} = 100$