**Q1. Logistic Regression – Decision Boundary:**

**** We consider the following model of logistic regression for binary classification with a sigmoid function

**Model:**

X2

8

2

4

6

2

4

6

x1

8

Suppose the trained parameter values are Ɵ0 = -8, Ɵ1 = 2 AND Ɵ2 = 2

**Predict ‘’y = 1’’ if h(x) >= 0.25**

Calculate and Draw the decision boundary according to the threshold given above. Show your working here. If you just draw the boundary without working, you will not get any point.

**Q2. Model Selection:** A friend of yours is faced with a regression problem with two possible inputs, X1 and X2. he/she considers three linear regression models:

(Model 1) h(x) = ɵ0 +ɵ1x1

(Model 2) h(x) = ɵ0 +ɵ1x1+ɵ2x2

(Model 3) h(x) = ɵ0 +ɵ1x1+ɵ2x2 +ɵ3x1x2

The data set is given in the following table:

|  |  |  |
| --- | --- | --- |
| x1 | x2 | y |
| 2 | 2 | 18.1 |
| 3 | 3 | 24 |
| 1 | 4 | 15 |
| 5 | 3 | 32.9 |
| 4 | 2 | 20.1 |
| 6 | 6 | 23 |
| 5 | 4 | 12 |
| 10 | 8 | 25 |

**Training Data: {(x(1), y(1)), (x(2), y(2)), (x(3), y(3)), (x(4), y(4))}**

**Validation Data: {(x(5), y(5)), (x(6), y(6))}**

**Test Data: {(x(7), y(7)), (x(8), y(8))}**

All models are fitted to a training data set using mean-squared-errors ( ), resulting in the three prediction models respectively:

(Prediction model 1): h(x) = 12.9 + 3.2X1,

(Prediction model 2): h(x) = 11.6 − 1.4X1 + 1.7X2

(Prediction model 3): h(x) = 10 + 2X1 + 3X2 - 0.5X1X2

Your friend is puzzled by these results and comes to you for advice.

1. What do you think, which model will be the best?
2. How well does the model generalize?

**Q3. Multiclass Classification:**

In the table given below, we have labeled data for patients. The output is classified into classes as given below:

y = 1 if patient is “**not- ill**”, y = 2 if patient has “**cold**”, and y = 3 if patient has “**Flu**”

|  |  |  |
| --- | --- | --- |
| **x1** | **x2** | **y** |
| 2 | 2 | 1 |
| 4 | 2 | 1 |
| 2 | 8 | 2 |
| 3 | 7 | 2 |
| 7 | 6 | 3 |
| 8 | 8 | 3 |

1. Draw the plot for the training data, where each class should be represented by a different symbol. (keep **x1** on x-axis and **x2** on y-axis)
2. How we will train logistic regression classifiers for this data?
3. On new input x (new patient), how we will predict if the patient has “flu”, “cold” or is “not-ill”.

**Q4.**  Suppose you train a logistic regression classifier in order to predict if the aircraft engine is faulty or not. Our model predicts 1 if h(x) > 0.7. Given the test data (mtest = 300), we already know that 20% of the aircrafts have actually fault. On testing, our hypothesis predicted that 30% of the aircrafts have fault. Only 50% of the predicted ones (it’s not 50% of the total), which actually have fault.

* 1. Create a confusion matrix with actual number of true positive, true negative, false positive and false negative examples. Moreover, Calculate the Precision, Recall, F score, and Accuracy.
  2. If we don’t want to miss too many faculty engines, what we will do (How we will change the threshold)?
  3. If we want to predict faculty engines only if we are very confident, what we will do (How we will change the threshold)?

**Q5.**  **Short Questions**:

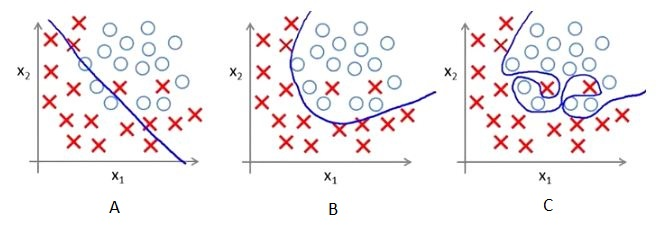
1. **[**2 points**]** In classification problems, why do we use equation (ii) for cost instead of equation (i) given below. Here in both equations h(x) is a sigmoid logistic function.

|  |  |
| --- | --- |
|  | Equation (i) |
|  | Equation (ii) |

1. What is the **overfitting problem** and what can be the possible cause for this problem?Write down all possible options for addressing the overfitting problem.
2. Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 98% of times in the training data. Your model has 98% accuracy after taking the predictions on test data. Which of the following is true in such a case?
   1. Accuracy metric is not a good idea for imbalanced class problems.
   2. Accuracy metric is a good idea for imbalanced class problems.
   3. Precision and recall metrics are good for imbalanced class problems.
   4. Precision and recall metrics aren’t good for imbalanced class problems.
3. Suppose you execute logistic regression twice, once with ʎ = 1 , and once with ʎ = 0. One of the times, you got parameters Ѳ , and the other time you got Ѳ . However, you forgot for which values ʎ corresponds to which value of Ѳ. Which one do you think to ʎ = 1?
4. Ѳ b) Ѳ

**Reason:**

1. Below are the three scatter plot(A,B,C left to right) and hand drawn decision boundaries for logistic regression.



**Suppose, above decision boundaries were generated for the different value of regularization. Which of the above decision boundary shows the maximum regularization?**

A) A B) B C) C D) All have equal regularization

**Reason:**

1. Consider a following model for logistic regression: P (y =1|x, w)= g(w0 + w1x)

where g(z) is the logistic function.

In the above equation the P (y =1|x; w) , viewed as a function of x, that we can get by changing the parameters w.

What would be the range of p in such case?

A) (0, inf) B) (-inf, 0 ) C) (0, 1) D) (-inf, inf)

**Reason:**