Discussion Session Problem 2
Logistic Regression and Regularization

1) Which of the following methods do we use to best fit the data in Logistic Regression?

A) Least Square Error
B) Maximum Likelihood
C) Both A and B

Logistic regression is based on maximum likelihood. The class section gives an intuition behind for using the maximum likelihood.

2) You are given a coin whose probability of landing on Heads is p. We toss a coin 10 times and get 7 Heads. What is the most-likely value of p?

A) 7/10
B) 5/10
C) 3/10

3) Which of the following evaluation metrics does not make sense if applied to logistic regression output to compare with target?

A) Accuracy
B) Log loss
C) Mean-Squared-Error

We used mean squared error for linear regression. Our cost function was not based on a squared error.

4) What can you say about feature normalization?

A) It is a good practice but is not required to run logistic regression
B) It is required to run logistic regression
C) It is a bad practice and should not be performed to run a logistic regression
D) None of the above

The main goal of normalization is to help us converge faster. We do not need it for logistic regression. normalization is REQUIRED for regularization.

5) Consider a sample of independently and identically distributed (I.I.D.) random variables $x_1, x_2, \ldots, x_m$, that each have Geometric distributions. In other words, $x_i \sim Geo(p)$ for all $1 \leq i \leq m$.

a) Derive the Maximum Likelihood Estimate for the parameter p of the Geometric distribution. Give a numeric answer. Explicitly show all the steps in your derivation.

b) Say that we have a sample of five such I.I.D. Geometric variables with the following values: $x_1 = 4$, $x_2 = 3$, $x_3 = 4$, $x_4 = 2$, $x_5 = 7$. What value of p in the Geometric distribution would maximize the likelihood of these observations? Give a numeric answer.

a. The mass function for the Geometric distribution with given parameter p is $f(X_i \mid p) = p(1-p)^{X_i - 1}$, where $X_i \geq 0$.

The likelihood function to maximize is:

$$L(p) = \prod_{i=1}^{n} p(1-p)^{X_i - 1}$$

So, the log-likelihood function to maximize is:

$$LL(p) = \sum_{i=1}^{n} [\log p + (X_i - 1)\log(1-p)]$$

Taking the derivative of LL(p) w.r.t. p, and setting it to 0, yields:

$$\frac{\partial LL(p)}{\partial p} = \sum_{i=1}^{n} [\frac{1}{p} + (X_i - 1)\frac{-1}{1-p}] = 0$$

Solving for p gives us: $\frac{n}{p} = \frac{1}{1-p}\sum_{i=1}^{n}(X_i - 1) \Rightarrow \frac{1-p}{p} = \frac{1}{n}\sum_{i=1}^{n}(X_i - 1)$

$$\frac{1}{p} - 1 = \left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - 1 \Rightarrow p_{\text{MLE}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}$$

b. We have: $p_{\text{MLE}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1}{\frac{1}{5}(20)} = \frac{5}{20} = \frac{1}{4}$

6) In this problem we will simultaneously estimate the difficulty of problem set questions and the skill level of each student. Consider a set of 200 students and 10 questions where each student answers each question. Let S_{ij} be an indicator variable which is 1 if student i answered question j correctly. You observe all S_{ij}.

We are going to make the assumption that the probability ($p_{i,j}$) that student i answers question j correctly is $p_{i,j}$ = **σ**(a_i - d_j) where:
σ is the sigmoid function,
a_i is a parameter which represents a student's ability
d_j is a parameter which represents a question's difficulty Use MLE to estimate the values for all parameters.

Use MLE to estimate the values for all parameters.

a) Write the log likelihood for a single response S_{ij} in terms of $p_{i,j}$ (hint logistic regression also assumes that its output is a probability of a binary event)
b) What is the partial derivative of LL for a single response S_{ij} with respect to a_i?
c) What is the partial derivative of LL for a single response S_{ij} with respect to d_j?
d) Explain briefly how you can estimate parameters given derivatives of log likelihood with respect to those parameters.

a. The likelihood function is the probability mass function of a Bernoulli with probability p_{ij}:

$$\text{Likelihood} = (p_{ij})^{S_{ij}} \cdot (1 - p_{ij})^{1 - S_{ij}}$$
$$\text{Log Likelihood} = S_{ij} \log(p_{ij}) + (1 - S_{ij}) \log(1 - p_{ij})$$

b. Using chain rule:

$$\frac{\partial \text{LL}}{\partial a_i} = \frac{\partial \text{LL}}{\partial p_{ij}} \cdot \frac{\partial p_{ij}}{\partial a_i}$$

Just like in a deep learning network:

$$\frac{\partial p_{ij}}{\partial a_i} = \frac{S_{ij}}{p_{ij}} - \frac{(1 - S_{ij})}{(1 - p_{ij})}$$

Staring with the equation for p_{ij}

$$\frac{\partial p_{ij}}{\partial a_i} = p_{ij}(1 - p_{ij}) \cdot \frac{\partial}{\partial a_i}(a_i - d_j)$$
$$= p_{ij}(1 - p_{ij})$$

You can optionally reduce your equations further. If you substitute and cancel you will get that:

$$\frac{\partial p_{ij}}{\partial a_i} = S_{ij} - p_{ij}$$

c. Using chain rule:

$$\frac{\partial \text{LL}}{\partial d_j} = \frac{\partial \text{LL}}{\partial p_{ij}} \cdot \frac{\partial p_{ij}}{\partial d_j}$$

This part is the same:

$$\frac{\partial p_{ij}}{\partial a_i} = \frac{S_{ij}}{p_{ij}} - \frac{(1 - S_{ij})}{(1 - p_{ij})}$$

Starting with the equation for p_{ij}:

$$\frac{\partial p_{ij}}{\partial d_j} = p_{ij}(1 - p_{ij}) \cdot \frac{\partial}{\partial d_j}(a_i - d_j)$$
$$= p_{ij}(1 - p_{ij})(-1)$$

You can optionally reduce your equations further. If you substitute and cancel you will get that:

$$\frac{\partial p_{ij}}{\partial d_j} = p_{ij} - S_{ij}$$

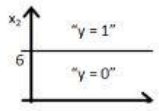
d. Use can estimate the value of all parameters using gradient ascent. Gradient ascent repeatedly takes a step along the gradient with a fixed step size. Just like when we implemented logistic regression, we can program our closed form mathematical solution for gradients to efficiently calculate the gradient for any values of our parameters.

7) Suppose you train a logistic regression classifier and your hypothesis function h is

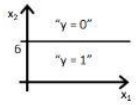
$$f_{w,b}(x) = g(w_1 x_1 + w_2 x_2 + b) \text{ where } b = 6,\ w_1 = 0,\ w_2 = -1.$$

Which of the following figure will represent the decision boundary as given by above classifier?

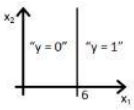
A)



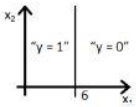
B)



C)



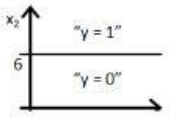
D)

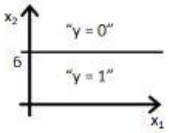


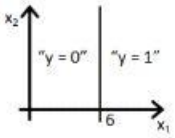
y = g(-6+x2) is shown in the option A and option B. But option B is the right answer because when you put the value x2 = 6 in the equation then y = g(0) you will get that means y= 0.5 will be on the line, if you increase the value of x2 > 6 you will get negative values so output will be the region y = 0.

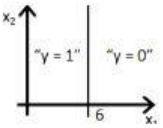
Mention that the decision boundary is given by ThetaX = 0 and then you find the 1 area for the values for which ThetaX is big.

8) If you replace coefficient of x1 with x2 what would be the output figure?

A) 

B) 

C) 

D) 

Similar to 7.

If w is a **constant**, what is the value of the slope of the logistic function at $x = 0$?
Assume that the model is $g(wx)$

A) w

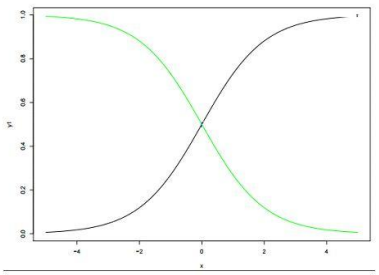
B) $w/4$

C) $1/4$

D) w^2

Answer: $w/4$. Conclusion: the larger the w, the steeper the curve hence the more confident we are in the feature because the function becomes more binary.

9) Below are two different logistic models with different values for b and w_1. Which of the following statement (s) is true about b and w_1 values of two logistics models (Green starts on top)?



Note: consider $Y = g(w_1 X + b)$. Here, b is intercept and w_1 is coefficient.

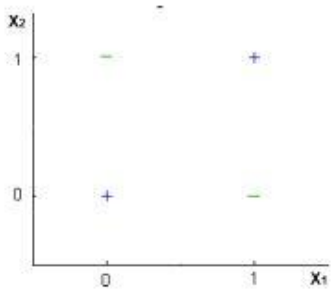
A) w_1 for Green is greater than Black

B) w_1 for Green is lower than Black

C) w_1 for both models is same

D) Can't Say

10) Can a Logistic Regression classifier do a perfect classification on the below data?



Note: You can use only X1 and X2 variables where X1 and X2 can take only two binary values(0,1).

A) True

B) False

C) Can't say

D) None of these

No, logistic regression only forms linear decision surface, but the examples in the figure are not linearly separable.

11) What can you say about regularized logistic regression vs. non-regularized logistic regression?

A) It will perform better on the training set

B) We can expect it to perform better on the training set

C) It will perform better on the testing set

D) We can expect it to perform better on the testing set