EE P 596 Conceptual Assignment 2: Due by 11:59pm Thursday, January 20

Qingchuan Hou

January 15, 2022

1. For logistic regression, the gradient is given by $\frac{\partial J}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})x_j^{(i)}$. Which of these is a correct gradient descent update for logistic regression with a learning rate of α?
   (A). $w^{(k+1)} = w^{(k)} - \alpha\frac{1}{m}\sum_{i=1}^{m}((w^{(k)})^T x^{(i)} - y^{(i)})x^{(i)}$
   (B). $w^{(k+1)} = w^{(k)} - \alpha\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - (w^{(k)})^T x^{(i)})x^{(i)}$
   (C). $w^{(k+1)} = w^{(k)} - \alpha\frac{1}{m}\sum_{i=1}^{m}(\frac{1}{1+exp^{-(w^{(k)})^T x^{(i)}}} - y^{(i)})x^{(i)}$
   (D). $w^{(k+1)} = w^{(k)} - \alpha\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h_{w^{(k)}}(x^{(i)})))x^{(i)}$

   Answer: C

2. Suppose you train a logistic classifier $h_w(x) = g(w_0 + w_1x1 + w_2x_2)$ where g is sigmoid function, Suppose $w_0 = -6$, $w_1 = 0$, $w_2 = 1$, Which of the following figures represents the decision boundary found by your classifier?

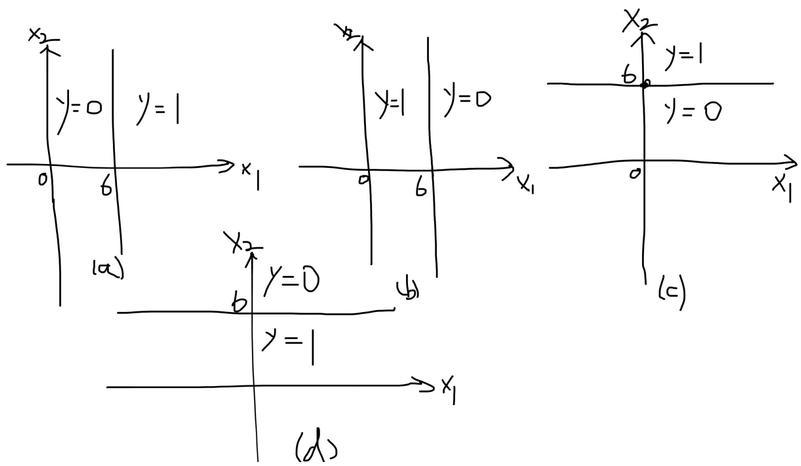

Figure 1: Decision Boundary

(A). Figure 1(a) is correct decision boundary.
(B). Figure 1(b) is correct decision boundary.
(C). Figure 1(c) is correct decision boundary.
(D). Figure 1(d) is correct decision boundary.

Answer: C

3. We aim to apply logistic regression approach for solving the classification problem illustrated below,where "+" means class y = 1 and "O" means y = 0. The data is linearly separable. We assume the $P(y = 1|X,w) = 1+$ ______ $exp_{w0}+^{1}{}_{w1x1}+_{w2}x_2$. The loss function $J(w) = -\sum_{i=1}^{N} log(P(y_i|X_i, w) + \lambda w_j^2$, with regularization of only one parameter j 1,2 and very large λ. Given the data shown above, state whether the training error **increases** or **nearly stays the same (zero)** for each w_j for very large λ.

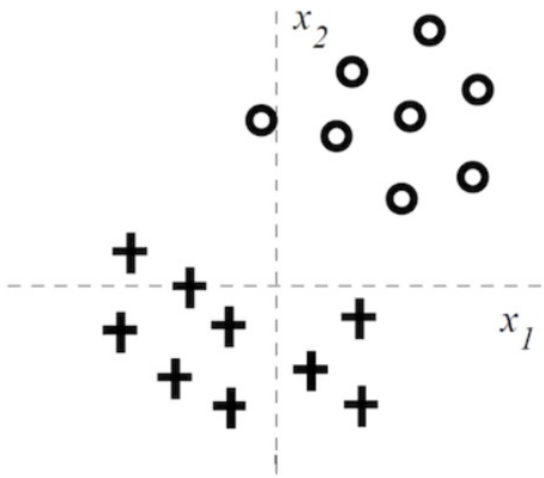

Figure 2: Linear separable data for classification

(A). Only regularize w_1, the training error will increase for larger *lambda* since the result decision boundary will become almost vertical.
(B). Only regularize w_2, the training error will stay the same for larger *lambda* since the result decision boundary will keep staying horizontal.
(C). Only regularize w_1, the training error will stay the same for larger *lambda* since the result decision boundary will keep staying horizontal.
(D). Only regularize w_2, the training error will stay the same for larger *lambda* since the result decision boundary will keep staying vertical.

Answer: C

4. Consider the Problem 3 using Lasso as regularization on w_1 and w_2, then the loss function becomes $J(w) = -\sum_{i=1}^{N} log(P(y_i|X_i, w) + \lambda(|w_1| + |w_2|)$ As we increase the parameter λ, which of the following do you expect? Please explain the reasons.

   (A). First w_1 will become 0, then w_2.
   (B). First w_2 will become 0, then w_1.
   (C). w_1 and w_2 become zero simultaneously. (D).
   None of them will become zero.

   Answer: A

   Explain reasons: *The lasso will sparse the function. It will ignore the most un-important contributions by sequence. For P3, the x1 is not important than x2, so the lasso will remove the x1 first by letting the w1 go to 0. If the λ continues increase, the regularization part will make all the w come to 0. Then the w2 go to 0.*

5. You are training a classification model with logistic regression. Which of the following statementsare true?

   (A). Introducing regularization to the model always results in equal or better performance on the training set.
   (B). Introducing regularization in the model always results in equal or better performance on examples not in the training set.
   (C). Add a new feature to the model are very likely to give you equal or better performance on the training set.
   (D). Add many new features to the model helps prevent overfitting on the training set.

   Answer: C

6. Which of the following is true to logistic regression?

   (A). Logistic regression cannot give you the confidence of a prediction.
   (B). Logistic regression cannot be affected by outliers in the data because the sigmoid function restricted the output between 0 and 1.
   (C). The feature vector X has linear relationship with the logits defined by $log(\frac{P(y|X)}{1-P(y|X)})$.
   (D). Using binary cross entropy loss to train logistic regression is better than mean square error because it can give us closed-form solution.

   Answer: C

7. You are working on housing price prediction problem given 4 features AreaOfHouse, NumberOfRooms, NumberOfFloors, DistanceToTransitCenter. You try to build a linear regression model with Lasso and Ridge regression separately, you tune your model with regularization parameter λ, ranging from 0 to very large number(almost infinity). You know in prior that the importance of 4 features: *AreaOfHouse* > *NumberOfRooms* > *DistanceToTransitCenter* > *NumberOfFloors*, and assume these 4 features are independent of each other. Please sketch approximate plot of absolute value of result coefficient(the weight after training) of each feature with respect to $1/\lambda$(model complexity) in the same figure, one figure for Lasso and the other for Ridge. (Think about what are differences on how these 4 features react to the changes of regularization parameter, and what are differences for lasso and ridge).