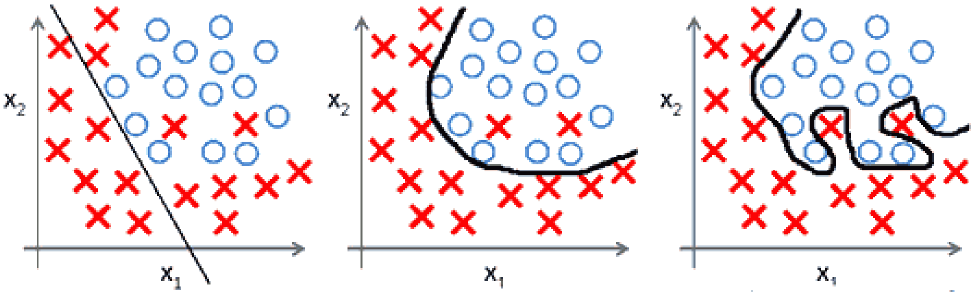
CSM148 Homework 2

Due date: Friday, February 10 at 11:59PM PST
Instructions: All work must be completed individually.

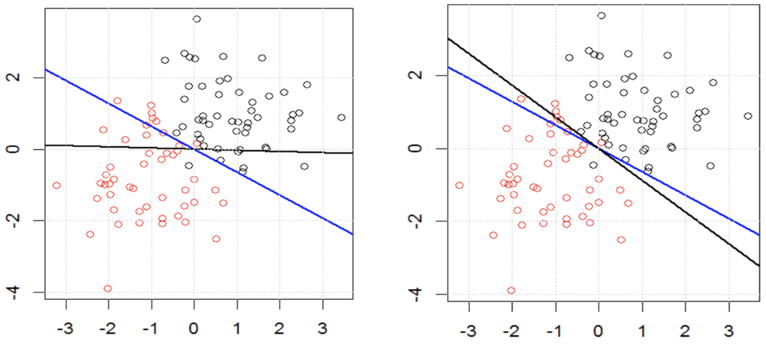
Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

1 Bias, Variance and Regularization

(a) **(5 points)** The figures below show the decision boundary of three different classifiers. In which of the figures below does the classifier have a larger bias and in which figure does the classifier has a larger variance? Draw out an approximate graph where you demonstrate how the training and test error for each classifier will change over time during the course of training the model, and explain how do you expect each classifier to perform (accuracy) on the test set.
Note: You don't need to calculate the errors, but you should show how the training and test errors compare for different classifiers.



(b) **(5 points)** One strategy to reduce variance and improve generalization is regularization. In figure below, the blue lines are the logistic regression without regularization and the black lines are logistic regression with L1 or L2 regularization. In which figure L1 regularization is used and why?



(a) (b)

2 Classification Metrics

You have trained a Logistic Regression classifier which gives you the following predictions on a test set:

Correct Label	Predicted Probability
1	0.97
1	0.93
1	0.65
1	0.52
1	0.37
0	0.89
0	0.71
0	0.56
0	0.54
0	0.16

(a) **(5 points)** Draw the Receiver Operating Characteristic (ROC) curve. Show clearly the points where the curve changes direction by e.g. including ticks on the x- and y-axis in the corresponding locations.

(b) **(2 points)** Compute the AUC score.

(c) **(2 points)** Draw the confusion matrix when the decision threshold is 0.5.

(d) **(2 points)** Using the previous confusion matrix, compute Accuracy, Precision, Recall, and F1 score.

(e) **(5 points)** Can we improve any of the previous scores (without a negative effect on any of the other scores) by changing the threshold? If yes, which threshold value would you choose and why? If not, explain why not.

3 Logistic Regression

Suppose we fit a multiple logistic regression: $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

(a) **(2 points)** Suppose we have $p = 2$, and $\beta_0 = 3, \beta_1 = 5, \beta_2 = -8$. When $X_1 = X_2 = 0$, what are the odds and probability of the event that $Y = 1$?

(b) **(2 points)** Suppose we increase the X_1 value by 2, how does it change the log odds and odds of the event that Y = 1? What if instead, we decrease the X_2 value by 2?

(c) **(2 points)** Explain how increasing or decreasing β_0, β_1 or β_2 affect our predictions.

(d) **(2 points)** What is the formulation of the decision boundary? Which points are on the decision boundary?

(e) **(2 points)** Suppose we fit another two logistic regression models: one with only X_1 and the other one with only X_2, and we observe that the coefficients of X_1 and X_2 in the two models are different than those specified in part (a). Explain what is the potential reason and why it could be problematic that the coefficients are different than those specified in part (a).

4 Logistic Regression with Interaction Term

You are analyzing how the birth weight of a baby (normal weight=0, low weight=1) depends on the age of the mother (number of years over 23, e.g. a 25-year-old will have value 2) and the frequency of physician visits during the first trimester of pregnancy (0=not frequent, 1=frequent). You have also decided to include an interaction term for age and frequency. Your logistic regression coefficients are as follows:

Feature	Coefficient
Intercept	-0.48
Age	0.06
Frequency	-0.45
Age × Frequency	-0.19

(a) **(4 points)** Discuss the meaning of each coefficient, and explain what does the coefficient of the interaction term show.

(b) **(3 points)** Specify the logistic regression models when the mother visited the physician frequently and when they didn't. Explain how the mother's age affects the odds in each scenario.

(c) **(4 points)** Compare how physician visits affect odds of low weight at ages 18, 23, 25, 28, 30, by calculating the odds ratio of low birth weight for mothers with frequent physician visits over those with non-frequent physician visits, in the following table (fill the "Odds Ratio" column in the table below). *Note: for age, you should use number of years over 23.*

Age	Odds Ratio	95% Confidence Interval
18		(0.705, 4.949)
23		(0.325, 1.201)
25		(0.262, 1.036)
28		(0.206, 0.916)
30		(0.050, 0.607)

(d) **(4 points)** Interpret the numbers in the "Odds Ratio" column, considering the listed confidence intervals. *Hint: what does an odds ratio of 1 mean (holding other predictors fixed)?.*

(e) **(5 points)** compare the "difference in probability" of low birth weight for mothers at ages 18, 23, 25, 28, 30, in the table below. Interpret your results and compare your interpretation to part (d).

Age	Difference in probability	95% Confidence Interval
18		(-0.788,0.393)
23		(-0.197,0.088)
25		(-0.232,0.046)
28		(-0.315,-0.016)
30		(-0.540,-0.092)