CSM148 Homework 2 Solutions

Due date: Wednesday, February 16 at 2PM PST
Instructions: All work must be completed individually.

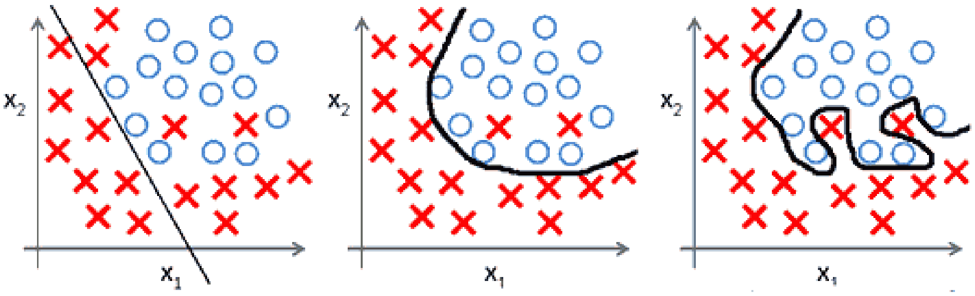
Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

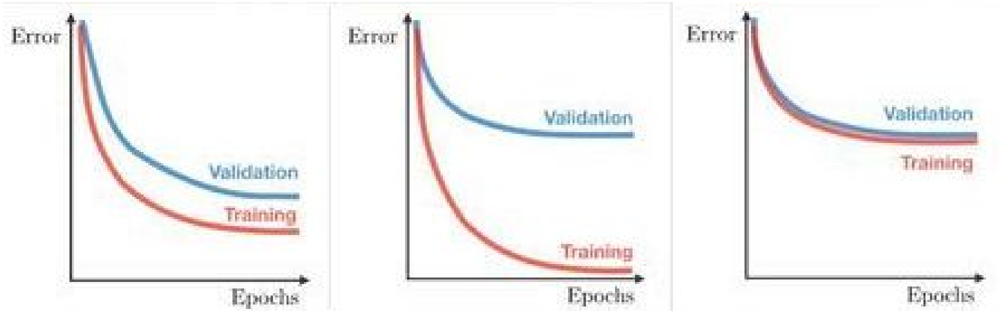
1 Bias, Variance and Regularization

(a) 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.

Solution: Figure (a) has a large bias and Figure (c) has a large variance. See figure below with the corresponding figure numbers.

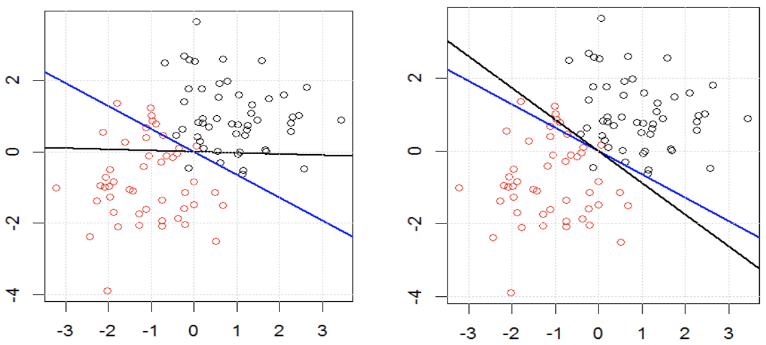


(a) (b) (c)



(b) (c) (a)

(b) 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)

Solution: Figure (a) corresponds to the L1 regularization since the decision boundary is horizontal showing that the coefficients tend to be zero.

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) 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.

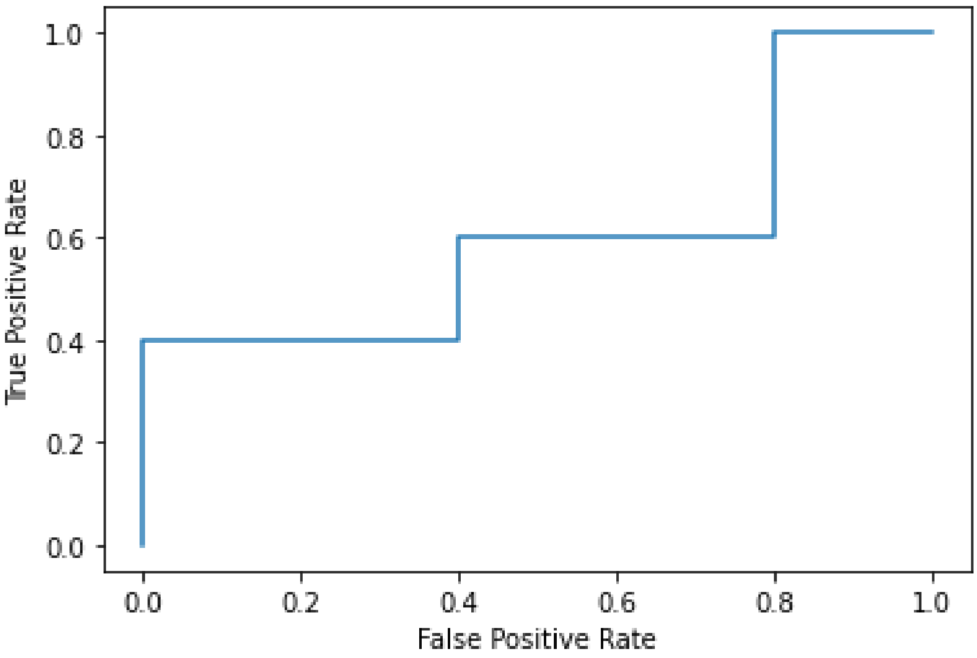


Figure 1: Solution for ROC curve

(b) Compute the AUC score. **Solution:** $AUC = 0.4 * 0.4 + 0.6 * 0.4 + 1 * 0.2 = 0.16 + 0.24 + 0.2 = 0.6$ **(Calculated as the sum of the areas below the three rectangles)**

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

	Prediction=0	Prediction=1
Correct=0	1	4
Correct=1	1	4

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

Solution:

Accuracy = (TP+TN)/(TP+TN+FP+FN) = 5/10 = 0.50

Precicion = TP/(TP+FP) = 4/8 = 0.5

Recall = TP/(TP+FN) = 4/5 = 0.8

F1 = 2*precision*recall/(precision+recall) = (2*(4/8)*(4/5))/(52/40)= 8/13 = 0.62

(e) 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.

E.g. Can improve precision by setting threshold = 0.6. Precision will then improve from 0.5 to $3/(3+2) = 0.6$. But this worsens Recall from 0.8 to 0.6. However, since F1 score with this threshold is better, we are likely to prefer this threshold to the original threshold. **Multiple correct solutions. A correct solution should show a specific threshold or threshold range and explain which metrics will change and which will not. If precision improves and recall worsens or vice-a-versa for change in threshold considered, F1 score must be referenced to explain which threshold is preferred.**

E.g. Can improve precision by setting threshold = 0.6. Precision will then improve from 0.5 to $3/(3+2) = 0.6$. But this worsens Recall from 0.8 to 0.6. However, since F1 score with this threshold is better, we are likely to prefer this threshold to the original threshold.

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) 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 $Y = 1$? **Solution: Odds are $e^{\beta_0} = e^3 = e$, probability is $\frac{e^3}{1+e^3}$.**

(b) How does one unit increase in X_1 or X_2 change the log odds and odds of the event that $Y = 1$? **Solution: One unit increase in X_1 increases log odds by 5 which in turn causes odds to be multiplied by $e^{\beta_1} = e^5$ (odds increase). Similarly, when X_2 is increased by one unit, log odds are decreased by -8 and odds are multiplied by e^{-8}.**

(c) Explain how increasing or decreasing β_0, β_1 or β_2 affect our predictions. **Solution: Increasing β_0 will increase our odds and therefore increase our predicted probability (similarly, decreasing it will decrease odds/probability). If β_1 is increased, our predicted odds/probability will increase if the feature is positive (opposite if the feature is negative). If β_2 is increased, our predicted odds/probability will decrease if the feature is positive (opposite if the feature is negative). Answer could alternatively describe how the probability curve is shifted when β_0 changes and how the curve becomes steeper or less steep based on the other β values.**

(d) What is the formulation of the decision boundary? Which points are on the decision boundary? **Solution: Our decision boundary can be found by setting $P(Y = 1) = 0.5$, so $log(\frac{P(Y=1)}{1-P(Y=1)}) = 0$. Our boundary then becomes $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0$**

(e) 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). **Solution: There is multicollinearity between these features. This makes it problematic to interpret the coefficients, as the coefficients may not be unique. Also, multicollinearity affects our confidence intervals for these features.**

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) Discuss the meaning of each coefficient, and explain what does the coefficient of the interaction term show. **Solution: Assuming that we assign weights in the same order as the features are listed in the table: β_0 shows the log odds for a 23-year-old who doesn't visit physician frequently. β_1 shows how much log odds increase per one year increase in age. β_2 shows**

how much visiting physician frequently decrease the log odds. The ratio of odds ratios for frequent/non-frequent visitors is shown by e^{β_3} (coefficient for the interaction term).

(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. **Solution: Frequent visitors:** $y = \beta_0 + (\beta_1 + \beta_3) * age + \beta_2$**, non-frequent visitors:** $y = \beta_0 + \beta_1 * age$ **(answers with probability or odds are also considered as correct). For frequent visitors, each year multiplies odds by** $e^{\beta_1+\beta_3}$**. For non-frequent visitors, each year multiplies odds by** e^{β_1}**. As a result, increasing age increases odds for non-frequent visitors while it decreases odds for frequent visitors.**

(c) 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.* **Solution:**

Age	Odds Ratio	95% Confidence Interval
18	$\frac{e^{-0.48+0.06*(18-23)-0.45*1-0.19*(18-23)*1}}{e^{-0.48+0.06*(18-23)-0.45*0-0.19*(18-23)*0}} = \frac{0.8607}{0.4317} = 1.99$	(0.705, 4.949)
23	$\frac{0.453}{0.726} = 0.6376$	(0.325, 1.201)
25	$\frac{0.394}{0.755} = 0.436$	(0.262, 1.036)
28	$\frac{0.343}{0.786} = 0.2466$	(0.206, 0.916)
30	$\frac{0.170}{0.960} = 0.1686$	(0.050, 0.607)

(d) 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)?.* **Solution: Odds Ratio being over 1 for 18-year-olds means that the risk is higher for frequent visitors. However, the 95% confidence intervals of the odds ratios include the null value of 1, so we do not have strong evidence of an association between frequent doctor visits and low birth weight for that age range. As women get older, the odds ratio goes down, which means that frequent visits can significantly reduce the risk on older women. While at ages 23 and 25 the confidence interval still includes 1, at ages 28 and 30 we are confident that the risk is actually reduced.**

(e) 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).

As a side note, in general, if a 95% CI doesn't include 0, the p-value would be < .05, which is the conventional cutoff for 'significance'. **Solution: The results match with part (d), so the probability of low weight is higher at 18 years while it is lower for older women. The confidence intervals overlap 0 at 18/23/25, so we do not have strong evidence of an association between low birth weight and frequent physician visits. At 28/30, the confidence interval is entirely negative, so we are confident that the probability of low weight becomes lower with frequent physician visits.**

As a side note, in general, if a 95% CI doesn't include 0, the p-value would be < .05, which is the conventional cutoff for 'significance'.

Age	**Difference in probability**	**95% Confidence Interval**
18	**0.4304-0.3143=0.116**	**(-0.788,0.393)**
23	**0.2829-0.3822=-0.099**	**(-0.197,0.088)**
25	**0.2332-0.4110=-0.1777**	**(-0.232,0.046)**
28	**0.1707-0.4551=-0.2843**	**(-0.315,-0.016)**
30	**0.1370-0.4850=-0.3480**	**(-0.540,-0.092)**