CSM148 Homework 2

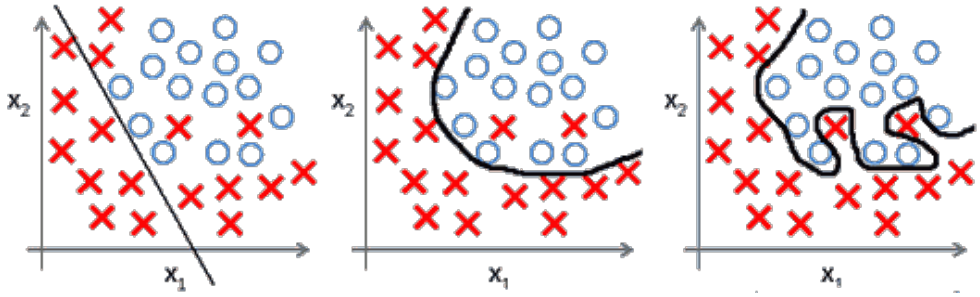
Due date: Wednesday, February 16 at 2PM PST
Instructions: All work must be completed individually.

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.).

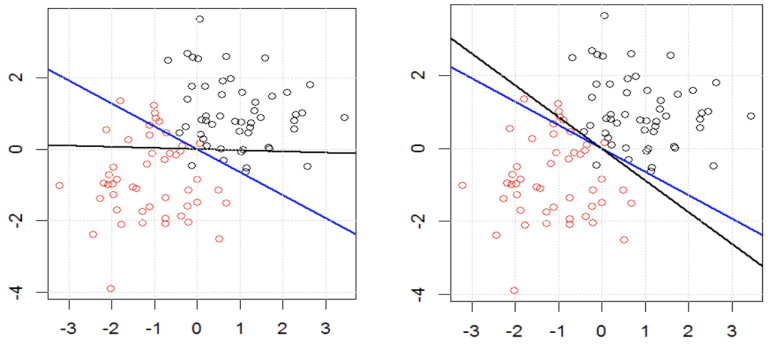
No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

1 Bias, Variance and Regularization

(a) **(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 Maximum Likelihood View of Linear Regression

In class, you have seen how to find a solution for linear regression by minimizing the Mean Squared Error (MSE) loss function. We can show that minimizing the MSE loss is equivalent to maximizing the likelihood function, if we assume that the noise follows a Normal distribution. In this problem, we will derive the solution for linear regression with one feature: $y_i = \beta_0 + \beta_1 x_i$.

(a) **(5 points)** Show the likelihood function assuming that the observations are normally distributed around the regression line. The equation for Normal distribution is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ (*Hint: Use your prediction as the mean value*).

(b) **(5 points)** Show that maximizing the (log) likelihood is equivalent to minimizing the MSE loss.

3 Classification Metrics

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

Correct Label	Predicted Probability
1	0.98
1	0.92
1	0.62
1	0.59
1	0.32
0	0.83
0	0.77
0	0.55
0	0.52
0	0.13

(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.

4 K-Nearest Neighbors for Classification

Load Olivetti Faces dataset using Scikit-Learn:

```
from sklearn.datasets import fetch_olivetti_faces
X,y = fetch_olivetti_faces(return_X_y=True)
```

Each row in X is a vector of size 4096, which represents a flattened 64-by-64 pixel gray-scale image. You can inspect individual images as follows (no need to include images in the answer):

```
import matplotlib.pyplot as plt
plt.imshow(X[0].reshape((64,64)), cmap='gray')
```

Split this dataset to train/test sets (70%/30%), and train a 3-NN classification model.

(a) **(3 points)** List the test metrics using classification_report function. What do the numbered lines in the classification report mean?

(b) **(2 points)** Does your model perform well across all the classes?

(c) **(2 points)** Would your results be different if the lighting or angles of the faces varied more? Would a change in the background affect the results? Explain why or why not.

5 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 $\beta_0 = 3, \beta_1 = 2, \beta_2 = -5$. When $X_1 = X_2 = 0$, what are the odds and probability of the event that $Y = 1$?

(b) **(2 points)** How does one unit increase in X_1 or X_2 change the log odds and odds of the event that $Y = 1$?

(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).

6 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	-0.52
Age	0.04
Frequency	-0.47
Age × Frequency	-0.18

(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 17, 23, 24, 25, 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
17		(0.705, 4.949)
23		(0.325, 1.201)
24		(0.262, 1.036)
25		(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 17, 23, 24, 25, 30, in the table below. Interpret your results and compare your interpretation to part (d).

Age	Difference in probability	95% Confidence Interval
17		(-0.788,0.393)
23		(-0.197,0.088)
24		(-0.232,0.046)
25		(-0.315,-0.016)
30		(-0.540,-0.092)

7 Multinomial Logistic Regression

Using Statsmodels, load the American National Election Studies dataset and train a Multinomial Logistic Regression model (MNLogit). You can load the dataset as follows:

```
import statsmodels.api as sm
anes_data = sm.datasets.anes96.load()
X = sm.add_constant(anes_data.exog, prepend=False)
y = anes_data.endog
```

(a) **(3 points)** Explain why the summary shows multiple values for each coefficient.

(b) **(2 points)** Looking at the p-values, are any of the features insignificant?

(c) **(2 points)** Are there any features that are useful for only few of the classes? If so, list the features and corresponding classes.

(d) **(2 points)** Are there any features that are useful for all the classes? If so, list the features.

8 Support Vector Machine

Using the following dataset:

x_1	x_2	Class
2	1	0
3	2	0
3	0	0
0	1	1
1	0	1
1	2	1

(a) **(3 points)** Plot the points. Draw the decision boundary assuming we are using hard-margin Linear SVM. Your plot should show the exact location of the boundary.

(b) **(3 points)** Which points are the support vectors? Would the boundary be different if one of the support vectors was removed? Explain briefly.

(c) **(3 points)** What is meant by a hard margin or soft margin? In this case will it matter if your Decision Boundary is either? Would the decision boundary be identical?

(d) **(3 points)** In Figure 1, explain which sub-figure (left, middle, right) corresponds to (a) SVM (linear), (b) SVM with polynomial kernel, and (c) SVM with RBF kernel with width (γ) equal to 1?

(e) **(3 points)** For the dataset shown in Figure 1, draw the (approximate) decision boundary for an SVM classifier with RBF kernel with width (γ) equal to 20, without regularization.

(f) **(3 points)** Draw the (approximate) decision boundary for an SVM classifier with RBF kernel with $\gamma = 20$ and $C = 0.1$ (C is the inverse of $L2$ regularization coefficient).

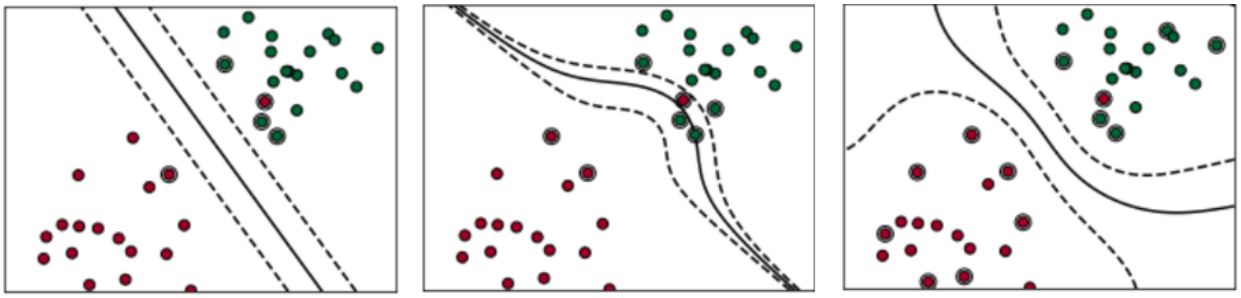


Figure 1: SVM decision boundary