CSM148 Homework 2

Due date: Wednesday, November 3 at 11:59 AM PST
Instructions: All work must be completed individually. If you consulted with any classmates for the homework, please note them on the first page.

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 Overfitting

Overfitting is a common problem when doing datascience work.

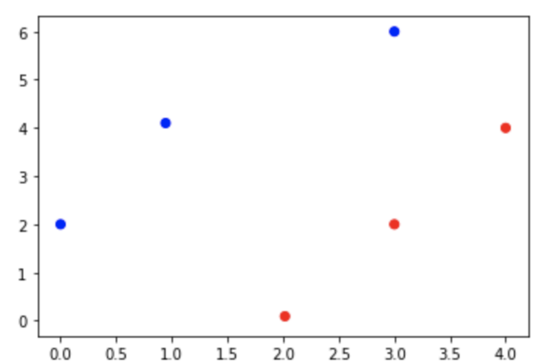
(a) How can you tell if a model you have trained is overfitting?

(b) Why do we want to avoid overfitting?

(c) Explain how L1 and L2 regularization methods can mitigate the problem. How do these two techniques affect the model weights? When would you choose one over the other?

2 K-Nearest Neighbors for Classification

Consider the following two types of data points shown in the Figure. The blue one with coordinates (x, y): (0, 2), (1,4), (3,6), and the red one with coordinates (x, y): (3, 2), (4, 4), (2, 0).



(a) Using k=1, classify points (1, 2), (2, 3), (10, 10). If you can't classify some of the points, explain why and propose a way to solve the problem.

(b) Repeat the same classification with k=3.

(c) If instead our dataset was comprised of blue points: (0, 200), (2, 500), (3,600) and red points: (3, 200), (4, 400), (1, 0), briefly explain what problem we might have with this kind of sampling and how can we address this problem?

(d) Suppose you have a dataset consisting of 1000 samples. Describe a method to select the optimal K to use for KNN?

3 Logistic Regression Interpretation

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

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

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

4 Confusion Table

Suppose we have the following confusion table output by the logistic regression using the probability threshold $P(Y = 1) \geq \pi$.

	Predicted $\hat{Y} = 0$	Predicted $\hat{Y} = 1$
Actual $Y = 0$	735	2
Actual $Y = 1$	50	45

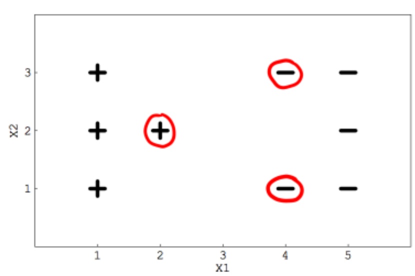
(a) What are false positives, false negatives, true positives, and true negatives.

(b) Compute precision, recall and F1 score.

(c) How would you expect precision, recall and F1 score to change if the threshold was lower? Provide a brief explanation.

(d) Can we compute AUC score with the given information? If not, what else would you need to know?

5 Support Vector Machine



(a) Suppose you have a dataset with 2 classes ('+' and '-'). If you remove one of the points that **is not** circled will that alter your Decision boundary?

(b) What is meant by a hard margin or soft margin? In this case will it matter if your Decision Boundary is either?

(c) How might your Decision Boundary differ when using Linear SVM compared to Radial Basis Function (RBF) kernel SVM?

(d) Explain what the parameters gamma and C of the RBF kernel SVM do.

6 Augmentation

Many methods for making predictions from data, such as linear regression, are limited in terms of the transformations that they can apply to input data before making a prediction. As linear regression assumes that the output is the sum of coefficients multiplied by input features, it is unable to account for cases where the impact of two features together is greater than the sum of their parts. For example, a house that both has > 5 bedrooms and is in California may be worth four times more than would be expected from the learned price impact of each feature on its own.

Feature Crosses are synthetic features you can form by crossing two or more features together, and they can help to improve the predictive power of techniques such as linear regression. Expanding on the above housing example, you could generate a new feature that indicates a combination of both a home's number of bedrooms and location.

(a) Describe two pairs of features from Project 1 that might be interesting to cross together, and explain why.

(b) Come up with an example of a dataset where linear regression would perform poorly without feature crosses. Provide either a table of data points or plot them, and explain why linear regression does not work in that situation. Show how feature crosses solve the problem.