**CS148 Homework 2**

Prithvi Kannan

UID: 405110096

Collaborators: Vanessa Wang

Problem 1:

1. See image
2. See image
3. Our dataset x-values and y-values are on different scales so the impact of the y-term is far more than the x-term when we perform KNN. To solve this, we can scale both X and Y by subtracting the distance to the mean and dividing by the standard deviation of the sample.
4. The best way to find the true optimal k is to plot error rate at a range of k’s and select the best one. A good starting point is to use k=sqrt(n) so we might start with k=31 here and try k-values in the range [25,35].

Problem 2:

1. True. Since the penalty term with Lasso forms a square constraint region and the intercepts are more likely at extrema’s (zeros).
2. False. The constraint region does not have extrema’s so there is no reason for the optimal point to coincide with zero values.
3. True. When we increase lambda in ridge we decrease the variance in the model, which is the same as reducing the magnitude of B.
4. [TODO]
5. False. Many data scientists prefer to use R as their programming language, and may spend time doing other things such as speaking with domain experts, developing experiment/sampling plans, creating visualizations.
6. False. Data engineering is a large part of the data science pipeline since the availability and quality of the input data is crucial to the success of the model building and analysis.

Problem 3:

Problem 4:

1. FP = 66, FN = 150, TP = 45, TN = 801
2. TPR = TP/(TP+FN) = 45/(45+150) = 0.231, FPR = TN/(TN+FP) = 801/(801+66) = 0.924
3. If we increase the threshold pi we would expect …. [TODO]

Problem 5:

1. The circled points are called the support vectors and define the decision boundary. Any of the other points can be removed and the boundary would stay the same.
2. Soft margin SVMs are useful to handle outliers and cases where we cannot perfectly separate the dataset with a decision boundary. [TODO] In this example, a vertical line at x1=3 would separate the points perfectly, so hard margin and soft margin would return the same decision boundary here.

Problem 6: