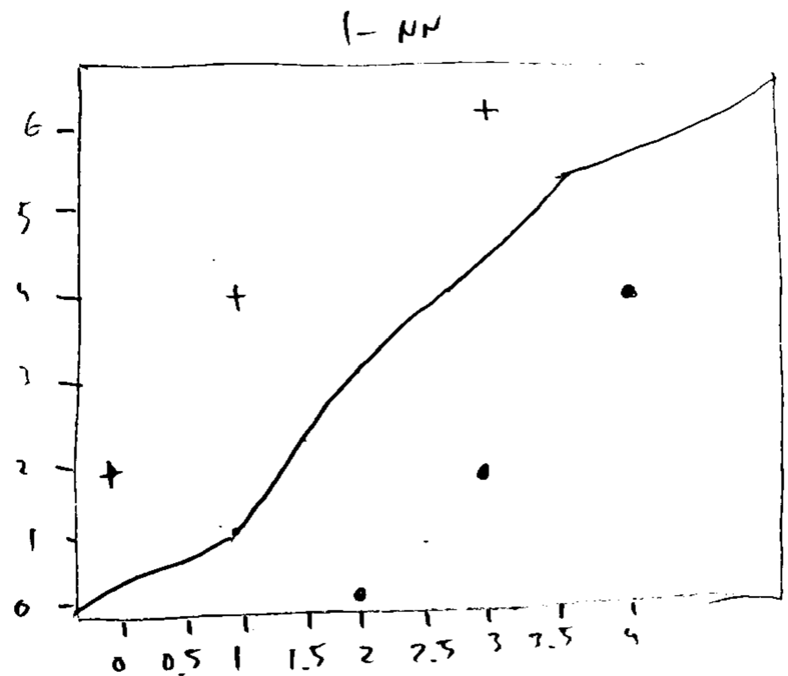
CS148 Homework 2

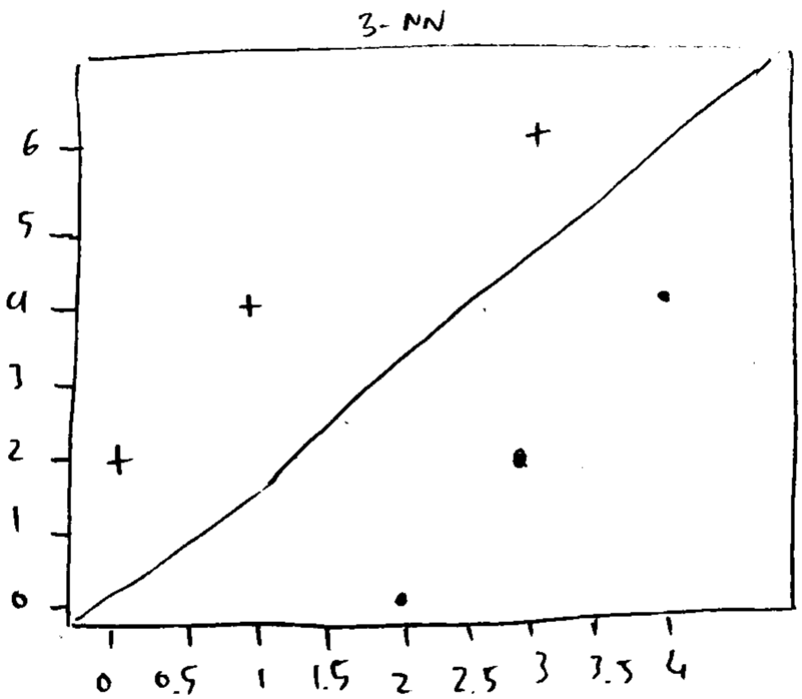
Prithvi Kannan
UID: 405110096

Collaborators: Vanessa Wang, Samuel Alsup

Problem 1:

a.





b.

c. 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.

d. 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:

a. False. Since the penalty term tends to shrink data toward the center, we expect to see more sparse beta.
b. False. The constraint region does not have extrema's so there is no reason for the optimal point to coincide with zero values.
c. True. When we increase lambda in ridge we decrease the variance in the model, which is the same as reducing the magnitude of B.
d. True. Increasing lambda in lasso will lower the coefficients which would make it more likely for more coefficients to end up as zero in B.
e. 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.
f. 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:

a. Xp = -(B0 + B1x1 + B2x2 + … + B(p-1)x(p-1))/Bp

b. When p = 2 we have X2 = - (B0 + B1x1)/B2
c. The coefficients B0, B1, B2 represent the relation between x1 and x2 to the labels. B0 is the bias term and indicates the curve is to the right. B1 and B2 define the shape of the curve and indicates that the parameter x2 has twice the weight of x1 in the overall shape. When x1=x2=0 then we have log((P(Y=1)/(1-P(Y=1))) = -1 + 0 + 0 = -1. P(Y=1) = 1/(1+e^*-(B0+B1x1+B2x2)) = 1/(1+e). A unit increase in x1 or x2 increases the odds that Y=1 by a factor of e.

Problem 4:

a. FP = 66, FN = 150, TP = 45, TN = 801
b. TPR = TP/(TP+FN) = 45/(45+150) = 0.231, FPR = TN/(TN+FP) = 801/(801+66) = 0.924
c. If we increase the threshold pi we would expect to reduce the number of false positives, which would raise precision. However since precision and recall are inversely related, this will likely lower recall.

Problem 5:

a. 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.
b. Soft margin SVMs are useful to handle outliers and cases where we cannot perfectly separate the dataset with a decision boundary, so introduce a special penalty term to allow for misclassification. Hard margin SVMs will attempt to find the largest decision boundary without misclassifying any points. 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.
c. In the RBF kernel, gamma defines the influence of a single training example, where low values indicating less important and smaller values indicating more important. C defines the amount of misclassification allowed versus the size of the margin. Low C will allow for a larger margin with less accuracy while a very large say may not allow any misclassification but with a small margin.

Problem 6:

a. Rather than just considering x-coord (lat) and y-coord (long) separately, we can take the cross product of those two to form a true location variable. This would allow for the additional complexity of considering distinct lat/long locations and increase the important of similar location points, which would matter significantly when it comes to Airbnb prices.
b. To make feature crossing more simple, we can discretize the latitude and longitude into bins.