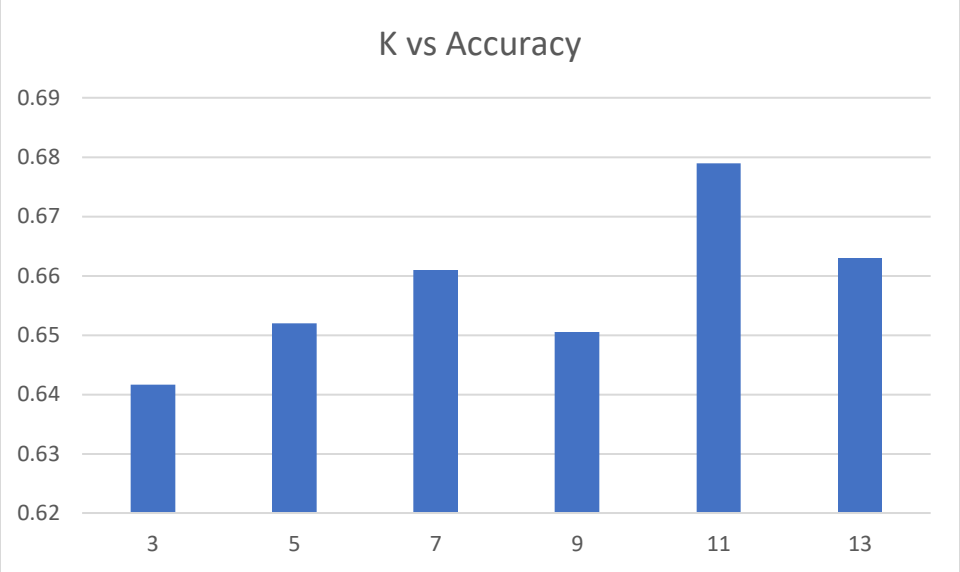
Brandon Peck
Bjp9pq
Machine Learning Homework 2

1)
2.5 –

	Accuracy
3	0.64167916042
5	0.652000576288
7	0.660984072528
9	0.650508829669
11	0.678994586791
13	0.663002582793

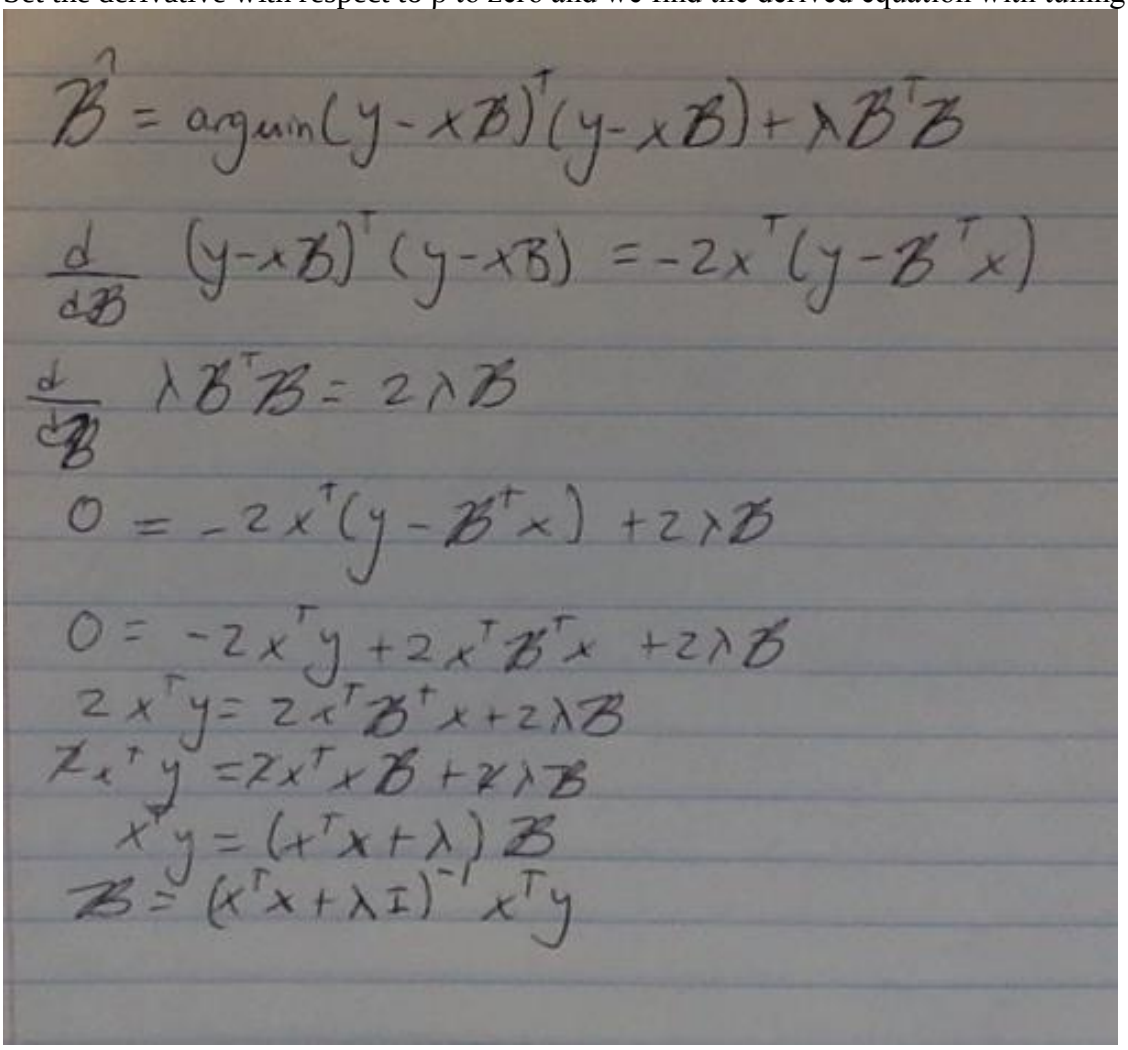
As K increases, the accuracy increases. This is because with an increase in K, there is a larger training size for each testing sample. More training data means there is a greater potential to find the true closest neighbors (not the sampled ones) to a testing record and therefore have a more accurate prediction for that record.



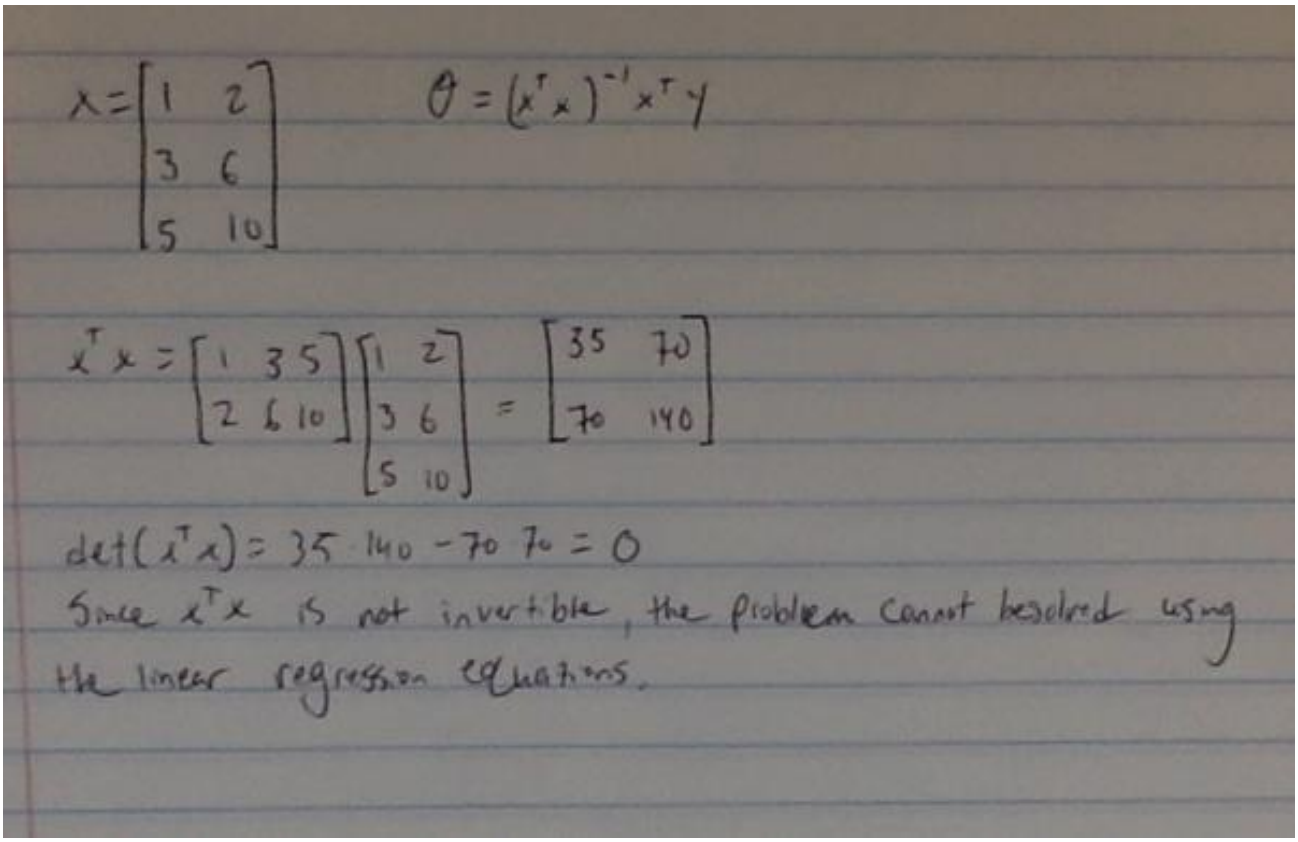
2)

1.1 –

Set the derivative with respect to β to zero and we find the derived equation with tuning lambda.



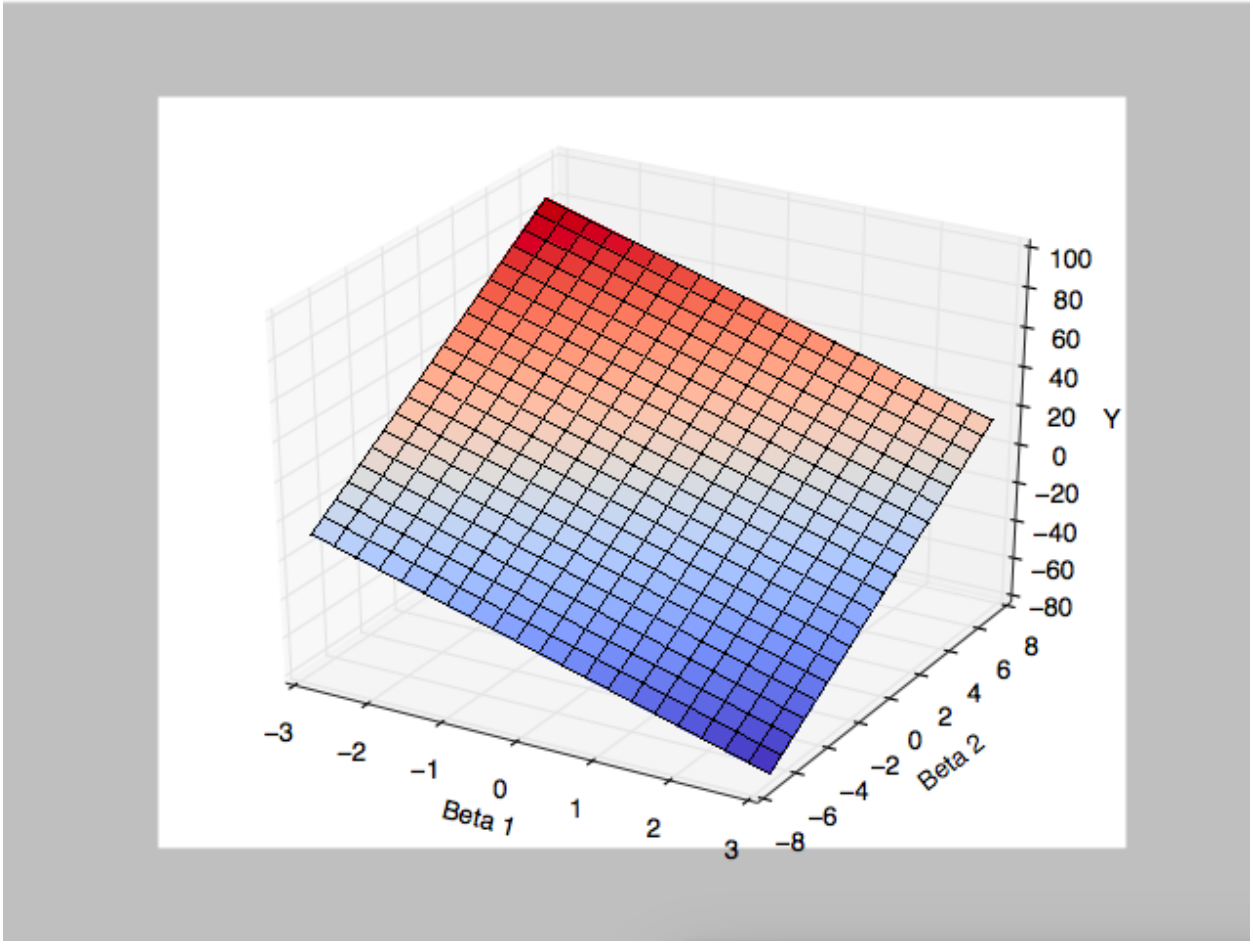
1.2 –



1.3 -

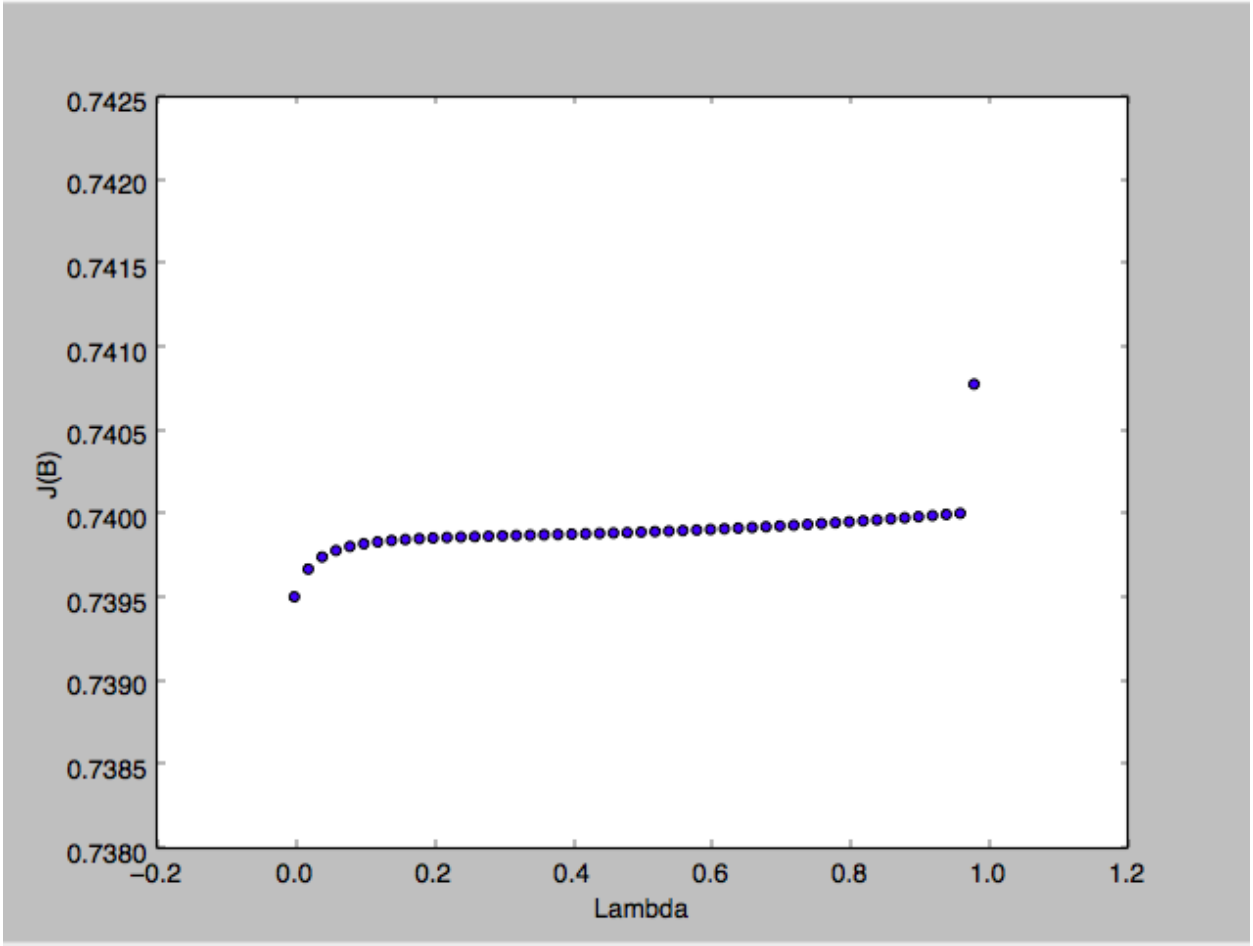
lasso regression, since it intends for the regularized beta norm values to equal zero.

1.4.1 –

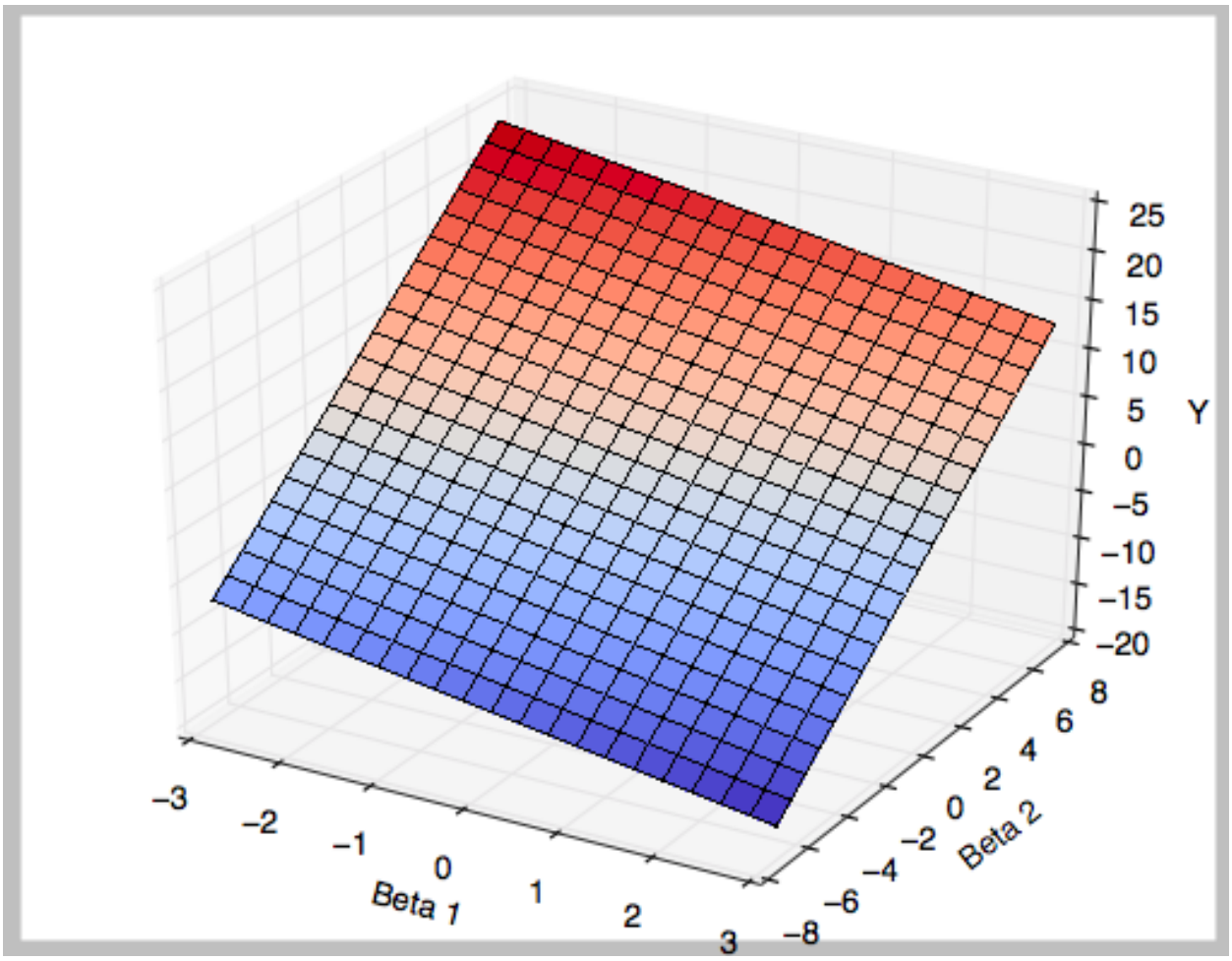


Learned plane when lambda equals 0.

1.4.3 –



Lambda vs J(B) graph. A smaller lambda value minimizes the loss function. This is expected since the smaller lambda allows for the beta value to more closely model the linear regression minimized loss function value.



Improved learned plain for the best ridge regression (lambda = 0.02)

1.5 –

$\beta_{RidgeRegression} = (2.97235068\ -1.54524704\ \ 2.23364926)^T$

$\beta_{LinearRegression} = (2.97139801\ -11.00332214\ \ \ 6.96229098)^T$

If the true coefficient is s $\beta = (3, 1, 1)^T$, it is clear that the coefficient created by the ridge regression is much closer to true β, and it performs better than the linear regression coefficient on this data.

This is likely because the data we use has variance on x_1 points. Since the ridge regression restricts the β value, the error from the variance in the ridge regression model is minimized.

3)
1.
A)
True
Ridge regression puts a constraint on the β coefficient. This creates a model that may not minimize the loss function when lambda is greater than zero, adding bias to account for variance in the data.

B)
False
A counter example is that the training data set does not contain data points with high variance, making the data model positively biased.

C)
True
As the training set size increases, regardless of the testing set size, the training error increases and testing error decreases moving the errors closer together (Slide 40 lect 10).

D)
True
A smaller set of training data will create a model that is tuned closer to that data, causing it to be overfitted.

E)
False
A smaller hypothesis space means there the model is more ridged and has a greater bias. This means overfitting is less likely as a greater bias means there is less variance in the model.

F)
False
β is inversely proportional to lambda, so as lambda grows infinitely β approaches zero.

G)
True
If feature variables are linearly dependent, X^TX would not be invertible. Since a lambda value can be added to the diagonal of the matrix, a lambda can be chosen that makes the columns of the feature variables linearly independent so the matrix can be invertible.

2)
A)
Not enough info

B)
Not enough info

C)
P(A) = P(A B) U P(A ~B)
P(A B) = P(A|B)*P(B) = 2/9
P(A ~B) = P(A|~B) * P(~B) = 1/3 * 2/3 = 2/9
P(A) = 4/9
P(B) = 1/3
P(A|B) = 2/3
P(B|A) = P(A|B)P(B)/P(A)
P(B|A) = 2/3 * 1/3 / (4/9) = 1/2
= 1/2

D)
P(B|A) = P(A|B)P(B)/P(A)
P(B|A) = 2/3 * 1/3 / 4/9
= 1/2