Michael Lee

CSE 446

HW1

1. Let T mean “Tested Positive”

Let H mean “Actually have the disease”

* 1. a)

=

=

b)

In our case,

* 1. The multinomial distribution of three variables follows the function:

\*Log of multinomial

Where represent respectively and represent days clear, days cloud, and days rainy respectively. To account for the constraint , we use the method of LaGrange multipliers, where we optimize for a solution under constraint. Hence, we subtract the function to account for our original constraint and solve by taking partial derivatives.

Since

So,

Plugging in:

The actual estimates then are

3. 1) To estimate we need and

2) To estimate we need

3)

By symmetry of the definitions between x and y we attain the expression for y:

4)

=

] -factoring terms out of sum

--we replaced by prior result

Therefore, we have shown

Deriving a similar expression for

5) Two examples where it would be beneficial to do in online learning is when our data is extremely large and training over the data set would be infeasible. The second situation would arise when our data is dynamic and there is a need to update our predictor routinely.

4.1 Lasso Regression

1.

1. The error on the training set will decrease because the focus of the algorithm will be to minimize the residual sum of squares on the training set.
2. The error on the testing set will be large because the model will over fit to the training set and lack prediction accuracy.
3. will have large coefficients because there is no penalty for the magnitude of coefficients.
4. All the elements of will be be non-zero.

2.

a) The error on the training set will be large because the focus of the algorithm will be to penalize the coefficients of our prediction. Hence, our coefficients will go to zero. Therefore, our predictions will be very poor when used on the data set, thus increasing the error.

b) The error on the test set will again be large as the predictor likely has all zero coefficients and our residual sum of squares calculation will be then square of each value in the data set as a result.

c) The coefficients of will be small, if not zero, because the function will penalize large coefficients.

d) There will be many non-zero elements of , if not all the elements.

4.2 Ridge Regression

3. When lasso regression pushes further away and when lasso pushes further away. We can understand this because the function at for ridge is greater when and pushes away more strongly, similarly when lasso has a greater value and pushes away further. This tells us that our predictions for large lambda are more reduced with ridge when than with lasso regression. Although we know lasso regression will enforce its coefficients to zero with large lambda rather than ridge which will make its coefficients approach zero.
4. Programming Question
5. Code is the program Crime\_predictor.py \*Note my data is in an excel sheet within the zip file, I modified my code where necessary to attain it in the first place and then removed the modifications for the final turn-in.
6. Regularization paths graph is on next page.
7. Squared Error Of Training Data Versus Log(Lambda)
8. Squared Error Of Test Data Versus Log(Lambda)

5.

6. One way we might select lambda is how well it performs on the test data as this is where we can assess our model’s accuracy in a real world setting. Though many other factors come into play, increasing lambda will increase the bias to reduce variance which may be something we try to avoid. Another idea all together would be to consider choosing lambda based on principles of cross validation.

7. Optimal Lambda = 18.75

Max coefficient: 0.07741645, with label ‘PctIlleg’

Least coefficient: -0.064852304, with label ‘PctKids2Par’

This max coefficient is the number of children born to unmarried parents and the least coefficient is the percentage of kids who have two parents. Seeing the least coefficient, the percentage of kids to a house with two adults, it seems interpretable that having a high percentage of parents to children would benefit the community and decrease crime (as seen by the negative coefficient). Seeing as that parents are often positive roles in children’s lives. Although, it must be noted an interpretation of the coefficients is not something that should be taken all too seriously. The algorithm is equally likely to drop coefficients of equal magnitude with no interpretation of those variables. Hence, lasso’s output is entirely based on what is put into it and what is left out of it.