Practice Exam 1

CS-6570: Data Science Algorithms I

Name:

Model Selection and Tuning (15 points) Ridge regression is a form of multivariate regression in which a quadratic penalty is placed upon the regression coefficients. So, instead of minimizing RSS, we minimize:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

1. (2 points) The parameter λ is not determined when the model is fit, but determined before the model is fit. What is that type of a parameter called?

2. (3 points) Is the ridge regression model biased? If so, in what way?

3. (4 points) Why would you possibly want to use a biased model?

4. (4 points) Explain what k-fold cross-validation is, and how it could be used to determine the value of λ with, for example, $k = 10$.

5. (2 points) What type of model does ridge regression become as λ gets very large? What type of model does ridge regression become as λ gets very small?

Variable Selection and Dimension Reduction (10 points) Suppose I have a multiple linear regression model, and I want to build a simpler model using fewer variables. I want to drop variables in a way that best maintains the predictive power of my model.

1. (4 points) Suppose I decided to pick the best model over all possible subsets of variables by determining which subset gives me the highest R^2 value. Would this be a good idea? If so, why? If not, why not?

2. (3 points) What would be the advantage of using forward stepwise selection to choose my variables as compared to best subset selection?

3. (3 points) If I wanted to completely eliminate some of my input variables, would I likely want to use lasso or ridge regression? Why would I prefer one to the other?

Bootstrapping (5 points) If I have a sample of 500 observations, and I use the bootstrapping approach to create another sample of 500 observations from the original sample, why isn't the other sample guaranteed to be exactly the same as the original? Could it be exactly the same as my original?

Logistic Regression (20 points) Please note that for all these questions by *logistic regression* we mean *binary logistic regression*. With linear regression we use a predictive model of the form:

$$\hat{y} = c_1 X_2 + c_2 X_2 + \cdots + c_n X_n.$$

With logistic regression we use a predictive model of the form:

$$\hat{y} = \frac{1}{1 + e^{-(c_1 X_1 + c_2 X_2 + \cdots + c_n X_n)}}.$$

1. (3 points) Why can the logistic regression model be interpreted as a probability while the linear regression model cannot?

2. (4 points) With linear regression we try to minimize the residual sum of squares (RSS). What do we try to optimize with logistic regression?

3. (4 points) If I flip a coin 10 times and 4 of those flips come up heads, derive the maximum likelihood estimate for p, the probability that the coin comes up heads on a given flip.

4. (4 points) With logistic regression, unlike linear regression, there's no closed-form solution to optimize the model coefficients. Instead, what method is typically used to search for optimal coefficients?

5. (5 points) Describe stochastic gradient descent, and how it differs from standard gradient descent.