Exam 1

CS-6570: Data Science Algorithms I

Name: Dylan Zwick (Solutions)

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?

   A *hyperparameter*

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

   Yes. The ridge regression model biases the coefficients towards 0.

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

   A biased model can be useful in decreasing variance, which can help if the number of parameters in the model relative to the amount of training data is high enough to potentially cause overfitting issues.

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, $k = 10$.

   k-fold cross-validation is a method for determining the value of a hyperparameter by breaking the modeling data into k disjoint segments. One segment is left out, and the model is trained on the $k - 1$ remaining segments, and then tested on the one left out. This is repeated leaving each segment out once, so a total of k times. The success of the model is its average performance on the data that is left out.

   To use this method to determine the value of λ, the performance of the model for different values of λ could be tested using 10-fold cross-validation, and the value of λ that performs the best would be the chosen one.

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?

   As $\lambda \to \infty$ the ridge regression model approaches a constant model. The ridge regression model approaches standard least-squares regression as $\lambda \to 0$.

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?

   It would not be a good idea. Adding more variables will always increase the R^2 value for the training data, so this approach is guaranteed to pick the subset with the most variables.

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

   The advantage to using forward stepwise selection, or backward stepwise selection, is it doesn't check every possible combination of variables like best subset. The number of possible combinations increases exponentially with the number of variables, so if the number of variables is reasonably large, best subset might be infeasible.

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?

   You'd want to use lasso regression. Lasso regression tends to set the coefficients of unimportant variables to 0, while ridge regression tends to make them small but still positive.

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?

The sample created almost certainly won't be the same as the original sample because bootstrapping is done by sampling with replacement, and so its likely that some observations will be repeated, while others won't occur at all, within the second sample.

It is possible that the sample created is exactly the same as the original, but it's very, very, very unlikely. The probability of it happening is $500!/500^{500}$, which is an incredibly small number.

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?

   Because the logistic regression model returns a value between 0 and 1, while the linear regression model has no such constraints.

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?

   With logistic regression we try to find the maximum likelihood, which is the value of the coefficients that makes the probability of the observations the greatest. In practice, we usually try to minimize the negative of the logarithm of the likelihood a.k.a. the log-likelihood.

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.

   The probability of 4 heads is $p^4(1-p)^6$. To find where this is maximized, we can take the derivative and set it to 0:

   $$4p^3(1-p)^6 - 6p^4(1-p)^5 = 0 \Rightarrow 4(1-p) = 6p \Rightarrow 4 = 10p \Rightarrow \frac{4}{10} = p.$$

   So, the probability that maximizes the likelihood of this result is $p = .4$.

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?

   Gradient descent is typically used to attempt to find the optimal value of the model coefficients.

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

   With standard gradient descent, we look at *all* of our data, and move in the direction that minimizes the error over all of it. With stochastic gradient descent, we look at the data one observation at a time, and each step we move in the direction that minimizes the error for the current observation.

   Stochastic gradient descent requires significantly less computation time than standard gradient descent, but it can introduce bias depending on the order in which the observations are processed.