CS4342 Assignment #4

Conceptual and theoretical questions

1. Using basic statistical properties of the variance, as well as singlevariable calculus, derive (1). In other words, prove that α given by (1) does indeed minimize $Var(\alpha X + (1 - \alpha)Y)$.

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_Y^2 + \sigma_X^2 - 2\sigma_{XY}} \quad (1)$$

Points: 5

2. We now review k-fold cross-validation.

(a) Explain how k-fold cross-validation is implemented.

(b) What are the advantages and disadvantages of k-fold crossvalidation relative to:

i. The validation set approach?

ii. LOOCV?

Points: 5

3. Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

Points: 5

4. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain *p + 1* models, containing *0, 1, 2, . . . , p* predictors. Explain your answers:

(a) Which of the three models with k predictors has the smallest training RSS?

(b) Which of the three models with k predictors has the smallest test RSS?

(c) True or False:

i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by forward stepwise selection.

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by backward stepwise selection.

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by forward stepwise selection.

iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

Points: 5

4. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \qquad \text{subject to} \qquad \sum_{j=1}^{p}|\beta_j| \leq s$$

for a particular value of s – s is postitive. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

Points: 5

5. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase λ from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error.

Points: 5

Applied Questions

1. We can use the logistic regression to predict the probability of default using income and balance on the Default data set. We will now estimate the test error of this logistic regression model using the validation set approach.

(a) Fit a logistic regression model that uses income and balance to predict default.

(b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

i. Split the sample set into a training set and a validation set.

ii. Fit a multiple logistic regression model using only the training observations.

iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.

iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

(c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

(d) Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.

Hint:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Points: 15

2. We will now perform cross-validation on a simulated data set.

(a) Generate a simulated data set as follows:

> x: create 100 random samples from normal distribution with mean 0 and variance 1

> $y = x - 2x^2 + noise$ noise are samples from normal distribution with mean 0 and variance 1

In this data set, what is *n* and what is *p*? Write out the model used to generate the data in equation form.

(b) Create a scatterplot of X against Y . Comment on what you find.

(c) Compute the LOOCV errors that result from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + \epsilon$

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon.$

(d) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

(e) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

Hints:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html
- Check extra for cross-validation https://scikit-learn.org/stable/modules/cross_validation.html

Points: 10

3. We will now consider the Boston housing data set.

(a) Based on this data set, provide an estimate for the population mean of medv. Call this estimate $\hat{\mu}$.

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.

(c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of medv.

Hints:

- Check https://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html
- We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.
- You can approximate a 95% confidence interval using the formula [$\hat{\mu}$ – 2 SE($\hat{\mu}$), $\hat{\mu}$ + 2 SE($\hat{\mu}$)].

Points: 10

4. Here, we will generate simulated data, and will then use this data to perform best subset selection.

(a) Generate a predictor X of length n = 100 from a normal distribution with mean 0 and variance 1, as well as a noise vector ϵ of length n = 100.

(b) Generate a response vector Y of length n = 100 according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

where β_0, β_1, β_2, and β_3 are constants of your choice. For X and ϵ, use the data being generated in (a).

(c) Perform best subset selection in order to choose the best model containing the predictors X, X^2, . . ., X^{10}. What is the best model obtained according to Cp, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.

(d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

(e) Now fit a lasso model to the simulated data, again using X, X^2,. . . , X^{10} as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

(f) Now generate a response vector Y according to the model

$$Y = \beta_0 + \beta_7 X^7 + \epsilon,$$

and perform best subset selection and the lasso. Discuss the results obtained.

Hints:

- Check this link for the best subset selection https://nbviewer.jupyter.org/github/pedvide/ISLR_Python/blob/master/Chapter6_Linear_Model_Selection_and_Regularization.ipynb#6.5.1-Best-Subset-Selection You can change the code for different metrics.
- Check this for forward and backward subset selection http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/
- Check this link for Ridge https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
- Check this link for Lasso https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

Points: 20

5. Here, we will predict the number of applications received using the other variables in the College data set.

(a) Split the data set into a training set and a test set.

(b) Fit a linear model using least squares on the training set, and report the test error obtained.

(c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

(d) Fit a lasso model on the training set, with λ chosen by crossvalidation. Report the test error obtained, along with the number of non-zero coefficient estimates.

Points: 15