**CS480 Computational Statistics II**

**Homework #7**

1. 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?

Although it is also possible that forward & backward stepwise selection might lead to the same model, the model utilizing best subset selection would perform the best. This is because using the p, the best subset selection will fit all 2^p potential models. There is no subset of predictors that could be identified by forward/backward selection that would not also be identified by best subset selection if the selection criteria is minimizing the training RSS.

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

We don't know, but best subset selection is more likely to produce better results. The cross-validated prediction error or some other type of penalized statistic (e.g., AIC, BIC, Adjusted R2) will typically be used to select the final model from the p+1 candidate models after the training RSS is minimized (to choose between models of the same size). There are no assurances regarding test set performance because everything is based on training performance.

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. **True, with one additional predictor.**

ii. The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the (k + 1)- variable model identified by backward stepwise selection. **True, with one predictor removed.**

iii. The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the (k + 1)- variable model identified by forward stepwise selection. **False, forward and backward stepwise are not identical, might be identical for some scenarios.In general, it is not true.**

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. **False, as said forward stepwise selection and backward stepwise selection are not identical.**

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. **False, as there is no guarantee.**

2. For parts (a) through (c), indicate which of i. through iv. is correct. Justify your answer.

a. The lasso, relative to least squares, is:

i. More flexible and hence will give improved prediction ac- curacy when its increase in bias is less than its decrease in variance.

ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

b. Repeat (a) for ridge regression relative to least squares.

c. Repeat (a) for non-linear methods relative to least squares.

a). Iii, for lambda > 0. If the rise in bias is offset by the decline in variance, predictions will be more accurate.

This is since lasso chooses the that minimizes RSS+λ∑pi=1|βi| rather than just the RSS in least squares. This tends to decrease the estimates towards zero since the optimal lasso will be closer to zero than the least squares for a given lambda > 0, and the shrinkage penalty λ∑pi=1|βi| is modest for β1,β2,…,βp close to zero. The shrinkage grows for larger because the importance of the shrinkage terms is higher than that of the RSS.

This shrinking, at the expense of a slight increase in bias, is what lowers the variance of the forecasts. Usually, this compromise is worthwhile.

b) similar to a. But the difference is in the ridge objective function RSS+λ∑pi=1β2i, where The shrinkage term for ridge regression differs slightly from the lasso's shrinkage term.

c) ii. When its rise in variance is less than its decrease in bias, it will be more flexible and, as a result, provide better prediction accuracy.

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

∑i=1n(yi−β0−∑j=1pβjxij)2subject to∑j=1p|βj|≤s

for a particular value of s. 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.

a) iv Steadily decrease. Another way to determine the lasso parameters is to minimize the RSS (subject to the restriction that ∑pj=1|βj|≤s). The least squares solution will meet the requirement once is big enough. The least squares solution in this case will always be the one that minimizes RSS=∑ni=1(yi−β0−∑pj=1βjxij)2 and also complies with this restriction. The training RSS will gradually reduce up until that time.

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

C) iii. Steadily increase. This is due to the fact that when the size of the constraint zone grows (s grows from zero), shrinkage decreases, improving model flexibility and causing an increase in variance. The variance will cease to rise if s is big enough for to fall under the constraint zone since the chosen will always be the least squares estimate.

d) iv. Steadily decrease. Simliar to c.

e) v. Remain constant. The error caused by innate uncertainty or noise in the system being approximated is known as the irreducible error. It is constant regardless of model flexibility because, regardless of how well-specified the model is (basically, it is completely independent of s), there may be unmeasured variables not in X that are needed to explain it or unmeasurable variation in y that cannot be predicted with the variables in X.

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

∑i=1n(yi−β0−∑j=1pβjxij)2+λ∑j=

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.

1. iii. Steadily increase. When λ = 0, the ridge regression β^  
    will be the same as the least squares estimate for β. Since the shrinkage term is gone, the training RSS will already be reduced. This training RSS can only grow as rises, and it will grow as shrinking rises.
2. ii. Decrease initially, and then eventually start increasing in a U shape. It is hoped that as (shrinkage) rises, the benefits of the decrease in variance will exceed the expense of decreasing the's towards zero. In general, this will result in a decrease in the test RSS until the model simply becomes underfit and loses predictive accuracy (where the increased bias outweighs the decreased variance), at which point the test RSS will begin to rise once more.
3. iv. Steadily decrease. By shrinking the's towards zero and lowering the variance, increasing reduces flexibility. When the variance is arbitrarily close to zero, we can arbitrarily increase the shrinkage to further reduce the variance (as the's approach zero, the model approaches the null model, and our predictions approach zero variance).
4. iii. Steadily increase. Increase in lambda will increase the bias as β shrinks to zero.
5. v. Remain constant. Models are unrelated to the irreducible mistake.