The Analytics Edge SUMMER 2022

Model selection concepts

1. Which of the following is correct when comparing LASSO to an ordinary least squares linear regression:

   (a) It is more flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   (b) It is more flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   (c) It is less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   (d) It is less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

2. Which of the following is correct when comparing non-linear regression methods to an ordinary least squares linear regression:

   (a) They are more flexible and hence will give improved prediction accuracy when their increase in bias are less their decrease in variance.

   (b) They are more flexible and hence will give improved prediction accuracy when their increase in variance are less than their decrease in bias.

   (c) They are less flexible and hence will give improved prediction accuracy when their increase in bias are less than their decrease in variance.

   (d) They are less flexible and hence will give improved prediction accuracy when their increase in variance are less than their decrease in bias.

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

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

   for a particular value of s. For parts (a) to (d), indicate which of (i) through (v) is correct and justify your answer.

   (a) As we increase s from 0, the *training* RSS will: (iv) steadily decrease.

   (b) As we increase s from 0, the *test* RSS will: (ii) decrease initially, and then eventually start increasing in a U shape.

   (c) As we increase s from 0, the variance will: (iii) steadily increase.

(d) As we increase s from 0, the (squared) bias will: (iv) steadily decrease.

Justification: As we increase s from 0, we are restricting the β_j coefficients less and less (the coefficients increase to their least squares estimates), and so the model becomes more flexible. Hence *training* RSS and squared-bias will steadily decrease. Moreover, more flexibility variance increases. Test RSS will decrease in the beginning with flexibility, but increases eventually.