Math 650 HW 4

Mose Wintner

1.a) Best subset has the smallest training RSS.
b) This depends on the problem.
c) F, T, F, F, F

3.a) iv, because the algorithm will perform as well as possible on the training set, which is steadily growing monotonically.
b) ii, since as s (flexibility) increases from 0, the model will come to hit the value of s minimizing test error and then the test error will proceed to rise.
c) iv. I assume we mean test variance, which will steadily decrease as s increases since more flexible models are capable of modeling data with higher variance.
d) ii, since the bias will be high for a model which is not flexible enough, then the bias will decrease as s approaches a more appropriate parameter, then it will increase again with s as the model overfits its training data.
e) v; it is independent of the model.

4.a) iv. The method will fit the training set as closely as possible for smaller values of λ; increasing λ allows for bias in order to control variance, thus decreasing performance on the training set.
b) ii, because as λ increases from 0, eventually it will hit a value minimizing test error and then the test error will proceed to rise.
c) iii. As λ increases, bias becomes increasingly allowed in order to decrease variance. Therefore variance will steadily decline. d) iv. As λ increases, flexibility decreases, so bias is introduced.
e) v; it is independent of the model.

5. a)

$$\min_{\beta_1,\beta_2} \left[(y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2)\right].$$

Under the conditions, this simplifies to

$$\min_{\beta_1,\beta_2} \left[2(y_1 - (\beta_1 + \beta_2)x_{11})^2 + \lambda(\beta_1^2 + \beta_2^2)\right].$$

b) The expression to be minimized is symmetric in β_1 and β_2, so the estimates must be equal.
c) Under the conditions, it is

$$\min_{\beta_1,\beta_2} \left[2(y_1 - (\beta_1 + \beta_2)x_{11})^2 + \lambda(|\beta_1| + |\beta_2|)\right].$$

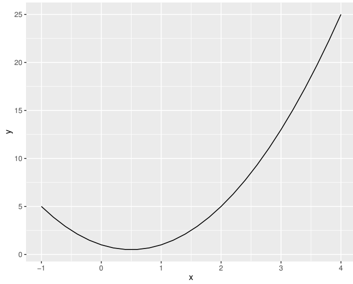
We can also formulate it as

$$\min_{\beta_1,\beta_2} 2(y_1 - (\beta_1 + \beta_2)x_{11})^2$$

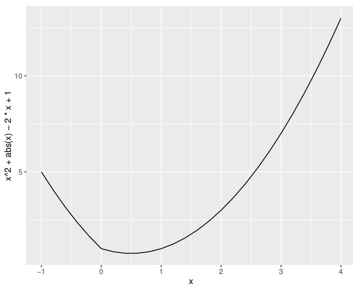
subject to $|\beta_1| + |\beta_2| \leq s$.
d) Under the second formulation in c, it is easy to see that the problem can only minimize with re-

spect to the quantity $\beta_1+\beta_2$. In particular, there will be a family of solutions described by $\beta_1+\beta_2 = c$.

6.a) For $p = 1$, this is $(\lambda+1)\beta_1^2 - 2y_1\beta_1 + y_1^2$. For $\lambda = y_1 = 1$, the plot is



which clearly has a minimum at $\beta_1 = 0.5 = y_1/(1+\lambda)$.
b) For $p = 1$, this is $\beta_1^2 + \lambda|\beta_1| - 2\beta_1 y_1 + y_1^2$. for $\lambda = y_1 = 1$, the plot is



which is minimized at $1/2 = y_1 - \lambda/2$.

7.a)

$$L = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right).$$

b) The joint pdf for β and x is

$$(2b)^{-p}\exp\left(-\frac{1}{b}\sum_{j=1}^{p}|\beta_i|\right)\cdot(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right).$$

Letting $x_{i0} = 1$ for all i, the posterior takes the form

$$p(\beta|x) = \frac{\exp\left(-\frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2\right)}{\int \exp\left(-\frac{1}{b}\sum_{j=1}^{p}|\beta_j|\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2\right)\, d\beta}.$$

c) The lasso estimate is the argmax of the posterior. Taking logarithms of the original form of the posterior and discarding terms not depending on β, we see that we seek the argmax of

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{p} x_{ik}\beta_k\right)^2 - \frac{1}{b}\sum_{j=1}^{p}|\beta_j|,$$

which, after multiplying through by $-2\sigma^2$, becomes the problem

$$\text{argmin}_\beta \sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{p} x_{ij}\beta_k\right)^2 + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|,$$

which is equivalent to finding the lasso estimate.

d) Letting $x_{i0} = 1$ for all i, the posterior takes the form

$$p(\beta|x) = \frac{\exp\left(-\frac{1}{c}\sum_{k=1}^{p}\beta_k^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2\right)}{\int \exp\left(-\frac{1}{c}\sum_{k=1}^{p}\beta_k^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2\right) d\beta}.$$

e) We again seek the argmax of the posterior. Again taking logarithms, ignoring constants, and multiplying through by $-c$, we get that the mode is equal to

$$\text{argmin}_\beta \sum_{k=1}^{p}\beta_k^2 + \frac{c}{2\sigma^2}\left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2$$

which is equivalent to finding the ridge regression estimate. It is the mean as well because the posterior is an ordinary Gaussian, for which the mean and mode are equal.

9,11. See attached .r file.