University of Washington

Stat / Biostat 527 Nonparametric Regression and Classification

Homework 2

Issued: Thursday, April 10, 2014 **Due:** Thursday, April 17, 2014

Abalone are a species of shellfish (mollusks). The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope-a boring and time-consuming task. The dust created through the cutting of abalone shells is also dangerous; it can cause irritant bronchitis, other respiratory irritation responses, and allergic attacks. It is thus of some interest to predict the age of abalone using covariates which are easier and safer to measure. The file **abalone-part.data** on the website contains observations of 150 abalone. In each case, there were 7 measured covariates, and one response - the number of rings.

Problem 2.1

Ridge regression has closed form solutions, as provided in Equation (10.23) of "Bayesian and Frequentist Regression Methods". Write your own code to get the ridge regression parameter estimates for the abalone data. Specifically, set the smoothing parameter λ from 0 to 100 by a 1 difference. For each given λ, store the estimated regression parameters.

(a) Plot the parameter estimates associated with the 7 covariates versus λ. Comment on their relationship.

(b) Perform Leave-one-out Cross Validation (OCV) and Generalized Cross Validation (GCV) on the abalone dataset for every value of λ. Plot the OCV/GCV scores against λ. Report the minimum OCV/GCV score and the corresponding optimal smoothing parameters chosen from each method. (Do not use built-in functions to calculate OCV/GCV)

Hint: To avoid penalizing intercept, you could first center y and all covariates. The intercept $\hat{\beta}_0$ can be estimated at the end by $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \cdots - \hat{\beta}_p\bar{x}_p$. In this problem, you only need to report the 7 parameters associated with the 7 covariates. After centering x and y, the model doesn't need to include intercept, since the fitted line, plane or hyperplane should pass the mean of x and mean of y, which are zero after centering. You may also want to scale *the covariates to have unit variance in order to get parameters of similar magnitude.*

Problem 2.2

Write your own code to implement the "shooting algorithm" for Lasso with the abalone data.

Hint: Again to avoid penalizing intercept and to get parameters of similar magnitude, you need to standardize all covariates and center y.

(a) Run the procedure for 1000 iterations for λ from 0 to 500 by a 10 difference (one iteration means randomly choosing a coordinate j and updating β_j). Treat the parameter estimates after 1000 iterations as the final parameter estimates. For each given λ, store the final parameter estimates and its corresponding shrinkage factor. The shrinkage factor is defined as $\sum_{j=1}^{p} |\beta_j^{Lasso}| / \sum_{j=1}^{p} |\beta_j^{LS}|$.

  (i) Plot the final parameter estimates associated with the 7 covariates versus $\mathbf{log(\lambda)}$ and describe their relationship. How is this different from the result you obtained for Ridge Regression in Problem 1?

  (ii) Plot the final parameter estimates associated with the 7 covariates versus the **shrinkage factor** and describe their relationship.

  Hint: You can check your answers using the `R` *package (lars).*

(b) Suggest a convergence criterion. Run your shooting algorithm again with your convergence criterion.

  (i) Plot the final parameter estimates associated with the 7 covariates versus the **shrinkage factor**. Compare your results with those obtained in Problem 2.2 (a) part (ii).

  (ii) Plot the number of iterations it takes to converge versus the **shrinkage factor**.