Biostat 830 Assignment 1

Due: Tuesday, Feb 6, in class

You don't *have to* work on all the problems: 1 to 3 are mandatory, you can select one of the 4 and 5 to work on.

1. Lecture Note 2, Exercise 1.
2. Lecture Note 2, Exercise 4. Compare the weights used by knn prediction algorithm
3. Implement a bootstrap method to estimate the EPE of the linear classifier that we used in the class and compare it to the K-fold cross-validation results for $K = 2, 5$ and 10.
4. For linear prediction function, generalized cross-validation (GCV) provides a convenient approximation to leave-one-out cross-validation. Consider a linear smoothing function,
$$\hat{\boldsymbol{f}} = \boldsymbol{H}\boldsymbol{y},$$
i.e., each fitted value is a linear combination of observed outcomes in the training data. An example of this is the least square fit. The matrix $\boldsymbol{H}$ is used to construct a prediction function subsequently. Here we focus on the cross-validation problem with the training data.
   (a) Show that if $\boldsymbol{H}$ is obtained from least squares algorithm,
   $$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - H_{ii}},$$
   where H_{ii} denotes the i-th diagonal element of $\boldsymbol{H}$.
   (b) Use above result to show that
   $$|y_i - \hat{f}^{-i}(x_i)| \geq |y_i - \hat{f}(x_i)|.$$
   (c) Show that the generalized cross-validation result, using a squared error loss, can be approximated by
   $$\text{GCV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - \text{trace}(\boldsymbol{H})/N}\right]^2$$
   (d) Conduct a numerical study to compare the results of GCV and leave-one-out cross-validation for a linear prediction function of your choice.

5. For the knn classifier that we discussed in the class

    (a) Implement a method of your choice to select the tuning parameter k, i.e., the "optimal" number of nearest neighbors.

    (b) Estimate the EPE for your optimal k using the training data

    (c) Simulate new data according to the true generative model and re-estimate the EPE for the estimated optimal k.