KNN and Cross Validation

Lecture 8: Evaluating Cross-Validation Techniques

- Recall $y = f(x) + \varepsilon$, $E(\varepsilon) = 0 \Rightarrow E(y|x) = f(x)$. (x_i, y_i)
  - We construct/train an estimator $\hat{f}$ for f, based on n iid observations. How well does $\hat{f}$ perform?

- We want to know the expected test MSE $\theta = E_{\text{training data}, x_{new}, y_{new}}\left[(y_{new} - \hat{f}(x_{new}))^2\right]$ ① tends to be too large/pessimistic on average (biased).

Types of estimators for validation

(1) We want to estimate θ by $\hat{\theta}_{val} = $ (of validation) VALIDATION as arithmetic mean of all.

$$\hat{\theta}_{val} = \frac{1}{|\text{Test set}|} \sum_{i \in \text{Test set}} (y_i - \hat{f}(x_i))^2$$, where $\hat{f}$ was trained on the training set.

(2) LOOCV $\hat{\theta}_{LOOCV} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{f}^{(-i)}(x_i))^2$ ~ less biased.

(3) Kfold CV $\hat{\theta}_{KfoldCV} = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{F_k}\sum_{i \in F_k}(y_i - \hat{f}_i^{(-F_k)}(x_i))^2\right]$

$$= \frac{1}{K}\sum_{k=1}^{K} MSE_{Kfold}$$

} which has less variance? (debatable).

- Motivation ex. gene expression data.
  - $\hookrightarrow$ $n \ll p$, where there are many genes (p) and few observations (n)
  - observations ex. binary response variable indicating disease status.
  - Goal: predict disease status based on gene expression
  - Consider a KNN estimator. If response variables are binary, then you can just do a majority vote of neighboring y's.
    ex. 1-NN classifier $\hat{f}(x_0) = \hat{y}_0 = y_i$ for i s.t. x_i is closest to x_0.

Preprocessing and preselection. We take 20 genes with largest marginal correlation with y.
Do 1-NN classification based on these 20 genes (ie fit a classifier)
Do CV to evaluate performance. For any $\hat{f}$, we know

$\theta = E(\text{test error rate of } \hat{f}) = \frac{1}{2}$ (ie if $\exists$ no info on x, y, then it is by random chance)

- Rcode line 151: stopifnot( ) to check data
  38: CV function cv.knn

- It is good to look @, investigate how folds are (randomly) chosen. (Need to be careful)
  Preselect features based on just the training data (?) (fold?) not the whole set!
  (Then, Apply the procedure to evaluate other data (see Rcode line 38)
  see Rcode lines 63-66 for how preselection differs for different num of folds K.

- As we choose K, assess the amount of variability and bias/center pt of the CV error rate. (should be at/close to 0.5).
  $\hookrightarrow$ output of line 116 (is incorrect procedure)

- We should choose folds once per estimate calculation; Then, choose (all folds again) for another estimate. ie choose folds for an estimate

- We want to have the distributions of CV error rates and that they are unbiased; and get info on variance. If each fold is like a different observation, recalling MSE_K, $Var(\hat{\theta}_{KfoldCV}) = Var\left(\frac{1}{K}\sum_{k=1}^{K} MSE_K\right)$

$\Rightarrow$ Then $Var(\hat{\theta}_{KfoldCV}) = \frac{1}{K^2}\sum_{k=1}^{K} Var(MSE_K) = \frac{1}{K} Var(MSE_K)$ This is only true if the K folds are independent

This is a good...