Example Sheet 1: Solutions

(Several of these are modified from Hastie, Tibshirani, and Friedman)

1. Suppose there are n points, i.i.d. in the unit sphere in $\mathbb{R}^p$. What is the median distance from the origin to the closest data point? Comment on the implications as p increases.

Define r as the median distance from the origin to the closest data point. Then

$$\Pr(\text{all } n \text{ points have distance} \geq r) = \frac{1}{2}.$$

For each point x_i in the unit ball, denote the distance to the origin d_i,

$$\Pr(d_i \geq r) = 1 - \Pr(d_i < r) = 1 - \frac{c\pi r^p}{c\pi 1^p} = 1 - r^p$$

.Therefore, we have

$$\begin{aligned} \Pr(\text{all } n \text{ points have distance} \geq r) &= \Pi_{i=1}^n \Pr(d_i \geq r) \\ &= (1 - r^p)^n = 1/2 \end{aligned}$$

Solve the equation, we get $r = (1 - \frac{1}{2}^{1/n})^{1/p}$

2. Consider a training sample of n points, i.i.d. $N(\mathbf{0}, \boldsymbol{I})$ in $\mathbb{R}^p$. Let $\boldsymbol{x}_0$ be a point at which a new prediction is to be made, drawn from this same distribution. Let $\boldsymbol{e} = \boldsymbol{x}_0/\|\boldsymbol{x}_0\|$ be its associated unit vector, so that $z_i = \boldsymbol{e}'\boldsymbol{x}_i$ is the projection of the ith point in the training sample in the $\boldsymbol{x}_0$ direction.

Show that the z_i are i.i.d. $N(0, 1)$ with expected squared distance 1 from the origin, whereas $\boldsymbol{x}_0$ has expected squared distance p from the origin. What is the implication?

For fixed n, as $p \to \infty$, what is the limiting geometric configuration of the n points?

Let $a = \frac{x_0}{||x_0||}$, $z_i = a^T x_i$, and then

$$\mathbf{E}(z_i) = \mathbf{E}[a^T x_i]$$

$$
\begin{aligned}
&= a^T \mathbf{E}[x_i] \\
&= a^T \cdot 0 = 0
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{Var}(z_i) &= \mathbf{Var}[a^T x_i] \\
&= a'^2 \mathbf{Var}[x_i] \\
&= a'^2 \cdot 1 \\
&= \sum_{k=1}^{p} a_k^2 = 1
\end{aligned}
$$

The Linear combination of normal rvs is still normal, so $z_i \sim N(0,1)$, and thus the squared distance from origin is 1. While for the target point, $x_i \sim No(0, \mathbf{I}_p)$, and so the squared distance has a χ^2 distribution with mean p.

From the previous result, the expected squared distance of a test point from the expected center point of the training data, which is origin, is p. The expected distance is thus about $\sqrt{p}$, while along direction a, all training points have expected distance 1. The implication is that the test point tends to be far from the training points.

The limiting configuration is a simplex with random orientation; see Hall, Marron, and Neeman, *JRSS-B*, **67**, 2005 for more details.

3. Why is cross-validation typically biased, and what is the direction of that bias? How does that depend upon v when using v-fold cross-validation?

The cross-validation error estimate typically understates the predictive accuracy since the fitted regression function is based on a large fraction of the data, but not the full data. As one moves from 2-fold cross-validation to leave-one-out cross validation, the bias gets smaller.

4. For squared error loss, define "in-sample error" as

$$
\mathrm{Err}_{\mathrm{in}} = n^{-1} \sum_{i=1}^{n} \mathbb{E}_{Y^N} \mathbb{E}_{\boldsymbol{y}} (Y_i^N - \hat{f}(\boldsymbol{x}_i))^2
$$

where Y^N represents a new observation at $\boldsymbol{x}_i$ and $\boldsymbol{y}$ denotes the response values in the training data. Thus in-sample error describes the average error one would get if one measured new response values at each of the original vectors of explanatory variables.

The training error is just

$$\text{err} = n^{-1}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x}_i))^2.$$

This is the average observed error in the fit to the training data.

Define the "optimism" as the expected difference between the in-sample error and the expected training error:

$$\text{Opt} = \text{Err}_{\text{in}} - \mathbb{E}_{\boldsymbol{y}}\text{err}.$$

Show that the optimism is equal to $(2/n)\sum \text{cov}(y_i, \hat{y}_i)$.

Let $\hat{f}(\boldsymbol{x}_i) = \hat{y}_i$. Then

$$\begin{aligned}
\text{Opt} &= \text{Err}_{\text{in}} - \mathbb{E}_{\boldsymbol{y}}\text{err} \\
&= n^{-1}\sum_{i=1}^{n}\mathbb{E}_{Y^N}\mathbb{E}_{\boldsymbol{y}}(Y_i^N - \hat{f}(\boldsymbol{x}_i))^2 - \mathbb{E}_{\boldsymbol{y}}n^{-1}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x}_i))^2 \\
&= n^{-1}\sum_{i=1}^{n}\mathbb{E}_{Y^N}\mathbb{E}_{\boldsymbol{y}}(Y_i^N - \hat{y}_i)^2 - \mathbb{E}_{\boldsymbol{y}}(y_i - \hat{y}_i))^2 \\
&= n^{-1}\sum_{i=1}^{n}\mathbb{E}_{Y^N}\mathbb{E}_{\boldsymbol{y}}(Y_i^{N2} - 2Y_i^N\hat{y}_i - \hat{y}_i^2) - \mathbb{E}_{\boldsymbol{y}}(y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2) \\
&= n^{-1}\sum_{i=1}^{n}\mathbb{E}_{Y^N}Y_i^{N2} - 2\mathbb{E}_{Y^N}\mathbb{E}_{\boldsymbol{y}}(Y_i^N\hat{y}_i) + \mathbb{E}_{\boldsymbol{y}}\hat{y}_i^2 - \mathbb{E}_{\boldsymbol{y}}y_i^2 + 2\mathbb{E}_{\boldsymbol{y}}(y_i\hat{y}_i) - \mathbb{E}_{\boldsymbol{y}}\hat{y}_i^2) \\
&= n^{-1}\sum_{i=1}^{n}\mathbb{E}_{Y^N}Y_i^{N2} - 2\mathbb{E}_{Y^N}\mathbb{E}_{\boldsymbol{y}}(Y_i^N\hat{y}_i) - \mathbb{E}_{\boldsymbol{y}}y_i^2 + 2\mathbb{E}_{\boldsymbol{y}}(y_i\hat{y}_i) \\
&= n^{-1}\sum_{i=1}^{n}2\mathbb{E}_{\boldsymbol{y}}(y_i\hat{y}_i) - 2\mathbb{E}_{Y^N}\mathbb{E}_{\boldsymbol{y}}(Y_i^N\hat{y}_i) \\
&= n^{-1}\sum_{i=1}^{n}2\mathbb{E}_{\boldsymbol{y}}(y_i\hat{y}_i) - 2\mathbb{E}_{Y^N}(Y_i^N)\cdot\mathbb{E}_{\boldsymbol{y}}(\hat{y}_i) \\
&= n^{-1}\sum_{i=1}^{n}2\mathbb{E}_{\boldsymbol{y}}(y_i\hat{y}_i) - 2\mathbb{E}_{\boldsymbol{y}}(Y_i)\cdot\mathbb{E}_{\boldsymbol{y}}(\hat{y}_i) \\
&= \frac{2}{n}\sum_{i=1}^{n}\text{cov}(y_i, \hat{y}_i).
\end{aligned}$$

5. For a linear smoother $\boldsymbol{H}$ (so $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$), show that

$$\sum_{i=1}^{n} \mathrm{cov}(y_i, \hat{y}_i) = \mathrm{tr}(\boldsymbol{H})\sigma_\epsilon^2.$$

This supports our use of the trace of $\boldsymbol{H}$ as the effective number of parameters in the smooth.

Let $\boldsymbol{h}_i'$ be the ith row of the smoothing matrix $\boldsymbol{H}$. Then

$$\begin{aligned} \mathrm{cov}(y_i, \hat{y}_i) &= \mathrm{cov}(y_i, \boldsymbol{h}_i'\boldsymbol{y}) \\ &= \mathrm{cov}(y_i, \sum h_{ij} y_j) \\ &= h_{ii}\mathrm{cov}(y_i, y_i) \\ &= h_{ii}\sigma^2 \end{aligned}$$

and the result follows trivially.

6. Asymptotically in the sample size, what fraction of the original sample is not used in a bootstrap sample?

The fraction excluded is about $1/e$.

$$\begin{aligned} \mathbb{P}[\text{ observation } i \text{ is in bootstrap sample }] &= 1 - (1 - 1/n)^n \\ &\approx 1 - e^{-1}. \end{aligned}$$

7. Efron famously introduced the bootstrap by setting a confidence interval on the correlation coefficient between the average LSAT scores and the average GPAs for entering classes at a sample of law schools. Download the population data (use law82 in R, or go to our class website) and use the bootstrap to set the confidence interval for 5 random samples of size 15 at level 95%. Compare your results to Efron's confidence interval (e.g., in the SIAM monograph) and comment.

Efron's initial sample was surprisingly lucky. Under repeated draws from the population, there is only about a 7% chance of getting a sample for which the sample correlation is so

close to the population correlation. Further, the confidence interval generated by Efron's initial sample under the bootstrap finds a standard error very close to that appropriate for this population.

8. For the k-nearest-neighbor smooth in $\mathbb{R}^1$, what are the degrees of freedom that it expends? For the linear-spline smooth in $\mathbb{R}^1$, what are the degrees of freedom expended? (Assume that alternating observations are used as knots.)

   For the k-NN smooth, each diagonal element of the smoothing matrix has value $1/k$, so the trace is n/k, the degrees of freedom used in that smoothing operation.

   For the linear spline smooth, note that for this specification the smooth need not be continuous at the knots. Thus we are fitting lines to the triplets of points corresponding to consecutive values of x. For convenience, assume that $n = 2K + 1$ is odd and that the x_i values have been ordered.

   There are two ways to find the df. The easy way is to note that between any pair of knots, there is one unused degree of freedom in the linear fit, and no degrees of freedom below the first knot and above the last. So the df are $K - 1$.

   The harder way is to find the trace. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ be $3\times$ matrices whose first column is $(1, 1, 1)'$ and whose second column is $(x_{2k-1}, x_{2k}, x_{2k+1})'$, for $k = 1, \ldots, K$. Then the hat matrices for the regression lines between the successive knots x_1 and x_3, x_3 and x_5, etc., are given by $\boldsymbol{H}_k = \boldsymbol{X}_k(\boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\boldsymbol{X}'$.

   To handle the edge effects, suppose we include x_{2k-1} and x_{2k} in the block, but not x_{2k+1}. The the full hat matrix consists of $\mathrm{diag}(\boldsymbol{H}_1^*, \ldots, \boldsymbol{H}_K^*)$ where the asterisk indicates that we only use the first two rows and columns of $\boldsymbol{H}_k$. And so on.

9. The mean integrated squared error (MISE) of an estimated function $\hat{g}$ is $\mathbb{E}[\int(\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x}))^2\, d\boldsymbol{x}]$. Derive the bias-variance decomposition for MISE.

Let $f(\boldsymbol{x}) = \mathbb{E}[\hat{g}(\boldsymbol{x})]$. Then

$$\begin{aligned}
\text{MISE} &= \mathbb{E}[\int (\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x}))\, d\boldsymbol{x}] \\
&= \mathbb{E}[\int [(\hat{g}(\boldsymbol{x} - f(\boldsymbol{x})) - (f(\boldsymbol{x}) - g(\boldsymbol{x}))]^2\, d\boldsymbol{x}] \\
&= \mathbb{E}[\int [(\hat{g}(\boldsymbol{x} - f(\boldsymbol{x}))^2 d\boldsymbol{x}] \\
&\quad -2\mathbb{E}[\int [((\hat{g}(\boldsymbol{x} - f(\boldsymbol{x}))(f(\boldsymbol{x}) - g(\boldsymbol{x}))]\, d\boldsymbol{x}] \\
&\quad +\mathbb{E}[\int [(f(\boldsymbol{x}) - g(\boldsymbol{x})^2]\, d\boldsymbol{x}] \\
&= \int \text{Var}[\hat{g}(\boldsymbol{x})] + \int \text{Bias}^2(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned}$$

10. Fit a multiple linear regression model, a stepwise linear regression model, an additive model, a generalized additive model, a projection pursuit regression model, a neural network model, an ACE model, an AVAS model, a regression tree model, and a MARS model to the Los Angeles Ozone data posted on our website's homework section. For the GAM fit, first transform the response value (groundlevel ozone concentration) using the transformation found from ACE. Comment briefly on the comparative fits.

We can easily compare the RMSE's for these different models. You can do this in R, Matlab, or with the code at the website given in class.