STAT 6242

HOMEWORK 3
DUE OCTOBER 17, 2016

October 6, 2016

Assume that Y and X are univariate random variables with

$$y = m(x) + \epsilon$$

where $\mathrm{E}(\epsilon|x) = 0$. Suppose we draw a random sample (x_i, y_i), $i = 1, \ldots, n$ and estimate $\mathrm{E}(Y|X = x) = m(x)$ with

$$\hat{m}(x) = \sum_{i=1}^{n} w(x, x_i, h) y_i = \sum_{i=1}^{n} w_i(x, h) y_i$$

using kernel weights with $\sum_{i=1}^{n} w_i(x, h) = 1$. Assume the kernel function is a second-order symmetric kernel with $R(K) = \int K^2(u)du$ (roughness of K) and $\mu_2(K) = \int u^2 K(u)du < \infty$.

Using second-order Taylor expansion of $\hat{m}(x)$ about $m(x)$ we obtain

$$\begin{aligned} MSE(\hat{m}(x)) &= \mathrm{E}\left[(\hat{m}(x) - m(x))^2\right] \\ &= h^4 \left(\frac{1}{2}m''(x) + \frac{m'(x)f'(x)}{f(x)}\right)^2 \mu_2^2(K) + \frac{\sigma^2(x)R(K)}{nhf(x)} + o(h^4) + o(\frac{1}{nh}) \end{aligned} \tag{1}$$

a. **Aymptotically Optimal Bandwidth:** Find the *optimal* bandwidth h by differentiating (1) with respect to h and setting everything to zero (neglecting the o terms).

b. **Plug-in Bandwidth:** the bandwidth computed by replacing $R(f'')$ in the $h^\star$ optimal formula (10) in the *Smoothing Techniques* handout by $R(g'')$, where g is a reference density. Compute the optimal plug-in bandwidth using $N(0, \hat{\sigma})$ as reference density for $f(x)$, where $\hat{\sigma}$ is the sample standard deviation.

c. **Bandwidth Selection by Cross-Validation:** The usual procedure is to come up with an initial grid of candidate bandwidths, and then use cross-validation to estimate how well each one of them would generalize. The one with the lowest error under cross-validation is then used to fit the regression curve to the whole data.

   Write a function with

1. 4 arguments: the vectors x, y, h, and integer $nfold$. Note that if $nfold = n$, this would result in leave-one-out CV.
2. The return value has three parts. The first is the actual best bandwidth. The second is a vector which gives the cross-validated mean-squared errors of all the different bandwidths in the vector bandwidths. The third component is an array which gives the MSE for each bandwidth on each fold.

v-fold CV algorithm:

1. Divide the data into v equal parts.
2. For each $k = 1, \ldots, v$, fit the model to estimate the smooth $\hat{m}_{-k}$ and compute its mean squared error for predicting the kth part:
$$MSE_k(\hat{m}_{-k}) = \frac{\sum_{\text{i in kth data set}} (y_i - \hat{m}_{-k}(x_i))^2}{\text{\# of points in the kth data set}}$$
The notation $\hat{m}_{-k}$ means that the smoother was based on $(1 - 1/v)$ of the data excluding the kth part.
3. The overall v-fold cross-validation error is
$$MSE = \frac{1}{v} \sum_{k=1}^{v} MSE_k(\hat{m}_{-k})$$
4. For each smoothing method, select the bandwidth with the smallest v-fold CV MSE.

For an initial set of candidate bandwidths, it is often reasonable to start around $1.06\hat{\sigma}_x/n^{1/5}$, where $\hat{\sigma}_x$ is the sample standard deviation of X.

Application: Generate 1000 data points where X is uniformly distributed between -4 and 4, and
$$Y = \frac{e^{7x}}{1 + e^{7x}} + \epsilon$$
with $\epsilon \sim N(0, .01^2)$. Use kernel regression to estimate $m(x) = \mathrm{E}(Y|X = x)$. Using any of the kernels discussed in class, obtain kernel smoothers for the three optimal bandwidths as described above (in CV, use 10-fold). Which of the three bandwidths result in the smallest MSE?