Homework 1 Part 1

Stat 435, Spring 2020
Due Friday, April 10, 11:59pm

Gain experience with Kernel smoothing

(a) (20 points) Write a R function

`ksmooth.train(x.train, y.train, kernel = c("box", "normal"), bandwidth = 0.5, CV = False)`

The kernels should be scaled so that their quartiles (viewed as probability densities) are at $\pm 0.25 * \text{bandwidth}$.

The function should produce a list with components `x.train` and `yhat.train`.

If `CV = True`, training observation `i` should not be used in the calculation of `yhat.train[i]`.

Do not assume that `x.train` is ordered. Try to be efficient!

(b) (20 points) Write a R function

`ksmooth.predict(ksmooth.train.out, x.query)`

The function should use linear interpolation inside the range of`x.train` and constant extrapolation outside the range.

Note: Do not assume that `x.query` is ordered. Do not use the R function `ksmooth`.

I have randomly divided the Wage data from ISLR into a training set Wage.train of size 1000 and a test set of Wage.test of size 2000. The data are in the "dump" file `home1-data.R` that you can `source`.

(c) Produce a scatterplot of `wage.train` vs `age.train` and add a kernel smooth for a `normal` kernel with `bandwidth = 3`. Print the residual sum of squares.

(d) Use the smooth computed above to predict `wage.test`. Draw a scatterplot of `wage.test` vs `age.test` and add the smooth. Print the residual sum of squares.

(e) Plot the resubstition estimate of the expected squared prediction error as a function of `bandwidth` for bandwidths = 1, 2,...,10. Print the 10 values.

(f) Plot the LOOCV estimate of the expected squared prediction for the 10 bandwidths. and print the 10 values. What is the bandwidth you would choose?

(g) Plot the test set estimate of the expected suared prediction error for the 10 bandwidths and print the 10 values.

(h) Plot the 5-fold CV estimate of the expected squared prediction error for the 10 bandwidths and print the 10 values.

Use the assignment to training observations to folds defined by the variable `fold` in `home1-data.R`.