Assignment 5 – Nonlinear regression & Cross validation

2201STAT3990 Mateen Shaikh

Due: March 26 by 3:30pm

Instructions:

Complete the following assignment, answering all questions and providing plots as appropriate. The total length of the assignment should not be more than 5 pages including plots and excluding code. Use a standard 12pt font with no less than 1 inch margins.

Your code should be part of your single (PDF) document at the very end. *Marks will be deducted for embedded code within the assignment.*

Kernel Prediction with medians

Using the `faithful` data, you will relate the eruption duration to the waiting time until the next eruption. For this analysis, the eruption duration is the independent variable (x) and the waiting time is the dependent variable (y).

1. Code a kernel smoother which uses the median as in class. Use a bandwidth of 1 minute. Use your smoother to estimate the 'true' waiting time after an eruption lasting x =2 minutes.
2. Consider several (at least 5) bandwidths between 0.5 mins and 2.5 mins. Find the prediction errors (using any reasonable distance metric) by repredicting the waiting time y for every unique eruption time, x. Plot corresponding prediction errors with respect to bandwidth in some sensible way. Comment on which bandwidths are best/worst.
3. Randomly partition the data into 8 disjoint subsets of approximately equal size. Repeat the previous question using 8-fold cross validation, this time predicting the left-out fold. Again, plot (including good labels) the prediction errors with respect to bandwidth sensibly and comment on which bandwidths are best/worst.

KNN

k-nearest neighbours is a similar approach to kernel smoothing. Instead of using all of the data points within a fixed certain distance of x, the estimate is made using a fixed number of data points closest to x. Instead of bandwidth, the number of neighbours, k is the parameter.

4. Create a new smoother this time using k-nearest neighbours. Use several (about 5 different) values of k ranging from 5%-40% of the total number of data points. Overlay the corresponding smoothers over the data. Comment on how the shape plotted changes with various values of k.