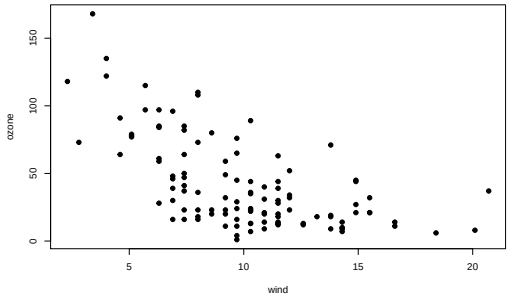
Stat 432 Homework 10

Assigned: Apr 5, 2019; Due: 11:59PM Apr 12, 2019

Question 1 (kernel regression) [5 points]

For this question, we will use the ozone data again. We consider just one variable, `wind` versus the outcome `ozone`.

```
library(ElemStatLearn)
data(ozone)
plot(ozone~wind, data = ozone, pch = 19)
```



One of the popular kernel functions besides the Gaussian kernel is the Epanechnikov kernel. Its kernel function $K(u)$ is defined as follows. Note that u represents $x - x_i$ when you calculate the kernel distance between these two points.

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{if } |u| \leq 1 \\ 0 & \text{o.w.} \end{cases}$$

To incooperate the bandwidth, we have $K_h(u) = K(u/h)/h$. Complete the following questions.

- Use this kernel function in the Nadaraya–Watson kernel estimator, and use the Silverman rule of thumb $h = 1.06\,\widehat{\sigma}\,n^{-1/5}$. Model `ozone` using the `wind` variable. Plot your fitted kernel regression function and calculate the training error.
- Use a smaller and a larger bandwidth (for example, you can try half or twice of the Silverman bandwidth), observe and comment on the changes in the fitted kernel regression function. How does this relate to the bias-variance trade-off?

Question 2 (local quadratic regression) [5 points]

Using this kernel function you defined in Question 1, based on our lecture notes on page 25, fit a local quadratic regression at three testing points: `wind` = 5, 10 and 15.

Extra-Credit Question (multi-dimensional kernel) [4 points]

In this question, we want to model `ozone` using two variables `wind` and `temperature`. We will still use the Nadaraya–Watson kernel estimator hence the only change we need to make from Question 1 is the kernel distance. Since we have two variables, we need to use a multi-dimensional kernel function. One example is the Gaussian kernel defined at page 32. We will further simplify this to be diagonal version such that

$$\mathbf{H} = \begin{bmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{bmatrix}$$

where h_1, h_2 are the bandwidth for the two variables, respectively. Use the Nadaraya–Watson kernel estimator with this two-dimensional kernel function to predict a new target point at `wind` = 10 and `temperature` = 80.

Note that h_1 and h_2 work almost the same as h defined in Question 1, so, pick a value for them that are suitable for their corresponding variable (does not have to be theoretically optimal).