Homework 5: Smoothing in Multiple Dimensions

36-402, Spring 2019, Section A

Due at 6:00 pm on Wednesday, 20 February 2019

AGENDA: Mostly theory, but also a first go at fitting and understanding an additive model.

1. *Bias of local averaging: one dimension*

   (a) (5) Show that

   $$\frac{1}{2h}\int_{x_0-h}^{x_0+h}(m+t(x-x_0)+c(x-x_0)^2)dx = m+kch^2 \tag{1}$$

   and find the constant k.

   (b) (5) Suppose that X is a one-dimensional variable uniformly distributed on the interval $[x_0-h, x_0+h]$, and $m(x)$ is a smooth function of x. Find an approximate expression for $\mathbb{E}[m(X)]$ which is valid when h is small. *Hint:* Taylor expand m around x_0.

   (c) (5) Suppose that we observe data in the form of (X,Y) pairs, where $Y = \mu(X)+\epsilon$, and $\mathbb{E}[\epsilon|X] = 0$. We try to estimate $\mu(x_0)$ by averaging all the Y_i where $|X_i - x_0| \leq h$. Suppose that the distribution of X is uniform on this interval. Show that the bias of this estimate of $\mu(x_0)$ is $O(h^2)$.

   (d) (5) In the previous problem, suppose that X has the non-uniform pdf $f(x)$. Show that the bias is still $O(h^2)$. *Hints:* You'll need to re-do the equivalents of the first two problems. Taylor expand f as well. (This one is a little more mathematically challenging, and not required for the rest of the assignment.)

2. *Bias of local averaging: two dimensions*

   (a) (5) Let $\vec{x}_0 = (x_{01}, x_{02})$ be a point on the two-dimensional plane, and B_h be the square of side $2h$ centered on $\vec{x}_0$. Let

   $$u(\vec{x}) = m+t_1(x_1-x_{01})+t_2(x_2-x_{02})+c_1(x_1-x_{01})^2+c_2(x_2-x_{02})^2+c_3(x_1-x_{01})(x_2-x_{02}) \tag{2}$$

for some constants m, t_1, t_2, c_1, c_2, c_3. Show that

$$\frac{1}{(2h)^2}\int_{B_h} u(\vec{x})d\vec{x} = m + (k_1c_1 + k_2c_2 + k_3c_3)h^2 \tag{3}$$

for some factors k_1, k_2, k_3.

(b) (5) Suppose that $\vec{X}$ is uniformly distributed on the square B_h around $\vec{x}_0$. Show that $\mathbb{E}\left[m(\vec{X})\right] = m(\vec{x}_0) + O(h^2)$ for small h.

(c) (5) We observe data in the form of $(\vec{X}, Y)$ pairs, where $Y = \mu(\vec{X}) + \epsilon$, and $\mathbb{E}\left[\epsilon|\vec{X}\right] = 0$. We try to estimate $\mu(\vec{x}_0)$ by averaging all the Y_i where $\vec{X}_i$ is in the box B_h around $\vec{x}_0$. Suppose that the distribution of X is uniform on this square. Show that the bias of this estimate of $\mu(\vec{x}_0)$ is $O(h^2)$. (Again, this still holds if X has a non-uniform distribution, and it continues to hold in higher dimensions, but the book-keeping gets annoying.)

3. *Variance of local averaging in p dimensions* Suppose that $\vec{X}$ is a p-dimensional vector, with pdf $f(\vec{x})$. B_h will be the box which extends for a distance of $\pm h$ from a point $\vec{x}_0$.

(a) (5) Explain why, for small h, $\Pr\left(\vec{X} \in B_h\right) \approx f(\vec{x}_0)(2h)^p$.

(b) (5) Explain why, with n samples, the expected number of points in B_h is $nf(\vec{x}_0)(2h)^p$.

(c) (5) Suppose that we estimate $\mu(\vec{x}_0)$ by averaging all the Y_i where $\vec{X}_i \in B_h$. Show that $\mathbb{V}[\hat{\mu}(\vec{x}_0)] = O(n^{-1}h^{-p})$.

4. (15) Exercise 8.3, parts 1–3 (five points each).

5. We return to the Chicago deaths data-set from Homework 1.

(a) (5) Fit, and plot, a non-parametric regression of deaths on temperature. (You can use any technique you like, but be sure to use cross-validation to pick how much smoothing to do, and to explain what technique you are using.) Describe the shape of the plot.

(b) (10) Using the `mgcv` package (introduced in Chapter 8), fit an additive model of deaths on temperature, `pm10median`, `o3median` and `so2median`. Plot the four partial-response functions, and describe their shape in words.

(c) (5) Does the shape of the partial response function for temperature match the shape of the curve you got in Problem 5a? Should the two curves match?

(d) (5) Which model predicts better, the one from Problem 5a or the one from Problem 5b? How can you tell?

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for.