Homework 5: Smoothing in Multiple Dimensions

36-402, Spring 2024

Due at 6:00 pm on Thursday, 22 February 2024

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

1. *Bias of local averaging: one dimension*

   (a) (3) Show that

   $$\frac{1}{2h}\int_{-h}^{h}(m+tx+cx^2)dx = m+kch^2 \quad (1)$$

   and find the constant k.

   (b) (4) Suppose that X is a one-dimensional variable uniformly distributed on the interval $[x_0-h, x_0+h]$, and $f(x)$ is a smooth function of x. Find an approximate expression for $\mathbb{E}[f(X)]$ which is valid when h is small. *Hints:* Taylor expand f around x_0; use Q1a.

   (c) (5) 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) (2) Show that

   $$\frac{1}{2h}\int_{-h}^{h}(m+tx+cx^2)(1+bx)dx = m+kch^2+rtbh^2+O(h^3) \quad (2)$$

   Find the constant r, and show that k is the same as in Q1a.

   (e) (4) Modify the set-up of Q1c by supposing that X has the non-uniform pdf $p(x)$. Show that the bias is still $O(h^2)$. *Hints:* Taylor-expand both μ and p, and use Q1d. (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) (3) $\vec{x} = (x_1, x_2)$ is a point in the two-dimensional plane; B_h is the square of side $2h$ centered on the origin.

$$u(\vec{x}) = m + t_1 x_1 + t_2 x_2 + c_1 x_1^2 + c_2 x_2^2 + c_3 x_1 x_2 \tag{3}$$

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

$$\frac{1}{(2h)^2} \int_{B_h} u(\vec{x}) d\vec{x} = m + (k_1 c_1 + k_2 c_2 + k_3 c_3) h^2 \tag{4}$$

and find k_1, k_2, k_3.

(b) (4) Suppose that $\vec{X}$ is uniformly distributed on the square of side $2h$ around $\vec{x}_0$. Show that $\mathbb{E}\left[f(\vec{X})\right] = f(\vec{x}_0) + O(h^2)$ for small h. *Hint:* Use Q2a.

(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, the bias being $O(h^2)$ 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 d dimensions* Suppose that $\vec{X}$ is a d-dimensional vector, with pdf $p(\vec{x})$. B_h will be the box which extends for a distance of $\pm h$ from a point $\vec{x}_0$. You may assume that $\mathbb{V}\left[Y | \vec{X} = \vec{x}\right] = \sigma^2$ for all $\vec{x}$.

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

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

   (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 $\mathbb{V}[\hat{\mu}(\vec{x}_0)] = O(\sigma^2 n^{-1} h^{-d})$.

4. *Convergence of local averaging* We continue to try to estimate $\mu(\vec{x}_0)$ by averaging all the Y_i where $\vec{X}_i$ falls within the box of side $2h$ centered at $\vec{x}_0$. We will call the squared bias of this estimate, plus the variance of this estimate, the "total squared error".

   (a) (1) Show that the total squared error is $O(h^4) + O(n^{-1} h^{-d})$. *Hint:* Use the previous problems.

   (b) (4) Show that the total squared error is minimized when $h = O(n^{-1/(4+d)})$.

   (c) (3) Show that the total squared error, at the optimal h, is $O(n^{-4/(4+d)})$.

   (d) (3) Explain why this method is consistent for any d, i.e., why $\hat{\mu}(\vec{x}_0) \to \mu(\vec{x}_0)$ as $n \to \infty$. *Hint:* $\mathbb{E}\left[(\hat{\mu}(\vec{x}_0) - \mu(\vec{x}_0))^2\right] =$ what?

(e) (3) Explain why, with this method, the number of data points n required to reach a given level of total squared error grows exponentially with d.

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

    (a) (3) 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) (7) Use resampling of residuals to create 85% and 95% confidence bands for your curve from Q5a, and add them both to your plot. Describe the shape of the bands, in words. *Hints:* see the HW 4 solutions, and section 7.2.1 of the textbook. Also, if you use the code from chapter 6 in the right way, you can use the *same* set of bootstrap simulations to get both confidence bands.

6. Still with the Chicago deaths data-set.

    (a) (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 shapes in words.

    (b) (5) Does the shape of the partial response function for temperature match the shape of the curve you got in Q5a? *Should* the two curves match?

    (c) (5) Which model predicts better, the one from Problem 5a or the one from Problem 6a? How can you tell?

7. (1) How long, roughly, did you spend on this assignment?

PRESENTATION 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. 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 plots and tables are generated by code included in the R Markdown file. 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. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.