STAT/BIOST 527
Homework 1
Out Thursday April 6, 2023
Due Thursday April 13, 2023

©Marina Meilă
mmp@stat.washington.edu

Problem 1 - Estimating h by cross-validation

For this problem, submit your code.
In this problem you will compute and plot a kernel density estimate of the corresponding densities f and g given below.

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

$$g(x) = \begin{cases} 4x & 0 \le x \le 0.5 \\ 4(1-x), & 0.5 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

a. Sample a training set $\mathcal{D}$ consisting of $n = 1000$ samples from f and a validation set $\mathcal{D}_v$ of $m = 300$ samples. Use the Gaussian kernel and find the optimal kernel width h by cross-validation. For this, construct $p_h(x)$ the density estimated from $\mathcal{D}$ with kernel width h. Then compute the log-likelihood $l_v(h)$ of the data in $\mathcal{D}_v$ under p_h. Also compute $l(h)$, the likelihood of the training set $\mathcal{D}$ under p_h. Repeat this for several values of h and plot $l_v(h)$ and $l(h)$ as a function of h on the same graph. (Suggested range of h: 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5). Save the training set $\mathcal{D}$.

b. Let h^* be the h that maximizes $l_v(h)$. Make a plot of $p_{h^*}(x)$ (by, for instance, computing the $p_{h^*}(x)$ values on a grid $x = -0.5 - 0.49, -0.48, \ldots 1.49, 1.5$). Plot the true $p(x)$ on the same graph. Make sure that the x axis extends left and right of the $[0, 1]$ interval and contains the entire region where $p_h \not\approx 0$.

c., d. Repeat questions **a, b** for G and g.

e.,f. Repeat questions **a,c** on the same samples $\mathcal{D}$ from **a, c**, this time with 5-fold CV. Plot the new $l_{5cv}(h)$ on the same graphs as in **a,c**, together with its standard deviation. Use either Rule 1 (argmax of $l_{5cv}(h)$) or Rule 2 to obtain h^*_{5cv}. Make sure your graphs are clearly labeled and readable. Make separate graphs for f and g.

The homework you hand in should contain: the formula(s) you used for p_h, the formula(s) you used to compute $l_v(h)$ and $l(h)$ and the required graphs. It is OK to replace log-likelihoods with likelihoods in the plots and equations. Clarification: because there are 2 true distributions here, we denote them f, g instead of p_X. The estimators will be both denoted p instead of $\hat{p}_X$, as in the notes.

g. Compare the optimal h's and the quality of the plots in **b, d**. Which of the densities looks easier to approximate? Which of the optimal kernels widths is larger, the one used for f or the one used for g? Can you suggest an explanation why?

[f. – Extra credit] Observing the bias and variance. Implement a sampler from f. Repeat $B = 10$ times: (1) draw a sample $\mathcal{D}^b$ of size $n = 100$ from F; (2) use the h found in **a** to estimate f from $\mathcal{D}^b$, denote this particular estimate of f by p^b_h, (3) use the value $h' = 2h$ to estimate f from $\mathcal{D}^b$, denote this particular estimate of f by f^b_{2h}, Plot $p^{1:B}_h$ on the same plot and $f^{1:B}_{2h}$ on a separate plot.

Compare the two plots in terms of bias and variance. In which of the plots do you observe higher variance? In which of the plots do you observe higher bias? Explain your answer. *To convince us that you understand these concepts, please use the terms correctly and precisely.*

Problem 2 – variation of h with n

Wenyu[1] has a data set $\mathcal{D}_0$ with size $n_0 = 1,000,000,000$ and he wants to compute a kernel density estimator based on this data. He decides to select h by cross-validation (CV). For clarity, in this problem we will always denote

$$p_{h,\mathcal{D}}(x) = \frac{1}{|\mathcal{D}|h} \sum_{i \in \mathcal{D}} k\left(\frac{x - x^i}{h}\right)$$

Wenyu will sample $\mathcal{D}_v \subset \mathcal{D}_0$ and use it as validation set, while using $\mathcal{D} = \mathcal{D}_0 \setminus \mathcal{D}_v$ for constructing $f_{h,\mathcal{D}}$. After CV is finished, Wenyu will use the optimal h^* obtained from CV to construct the final density estimator $f_{h^*,\mathcal{D}_0}$.

a. Recommend a value n_v for the size of the validation set $|\mathcal{D}_v|$. Explain your choice.

b. Wenyu has followed your advice in **a**, and now has $\mathcal{D}_v$ of size n_v and $\mathcal{D}$ of size n, with $n + n_v = n_0$. He has chosen a range of h with $m = 100$ possible values. He would like to know how many times the function $k(\,)$ would have to be computed to complete the entire CV procedure. For example, to obtain $f_{h,\mathcal{D}}(x)$, the kernel function $k(\,)$ is computed n times, one for each term $k(\frac{x-x^i}{h})$.

c. Denote by N the value computed in **b**. Wenyu discovers that on his computer it will take too long to run the entire procedure and he will miss his homework deadline. Hence, he subsampled a data set $\mathcal{D}_{\text{small}} \subset \mathcal{D}$ of size $n_{\text{small}} = 10,000$, and obtained h^*_{small} by CV with $\mathcal{D}_{\text{small}}$ and $\mathcal{D}_v$.

But what he really needs is h^*, the optimal kernel width for $f_{h^*,\mathcal{D}_0}$. Can he obtain h^* by a simple calculation from h^*_{small} and the other information available?

Problem 3 – k-NN and Kernel Regression on a toy data set

For this problem, feel free to make the graphs by hand or computer, as you wish. I recommend, though, that you do them by hand. If you make them by computer, please submit the code. *As always, you are required to implement all the functions by hand, and this toy problem is no exception.*

The data set is $\mathcal{D} = \{(1,1),(2,2),(3.5,1),(4.0)\}$, $n = 4$. The task is to perform (by hand, preferably) non-parametric regression on these data.

a. Let b_1 be the square kernel with $h = 1$

$$b_1(z) = \begin{cases} 1, & \text{if } |z| \le 0.5 \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

and let $b_{1/2}$ be the square kernel with $h = 0.5$. Complete the formula below (no proof required), in a way that ensures $\int_{\mathbb{R}} b_{1/2}(z)dz = 1$.

$$b_{1/2}(z) = \begin{cases} ?, & \text{if } |z| \le ? \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

b. Let $\hat{y}_1(x)$ be the Nadaraya-Watson kernel regression result for $\mathcal{D}$ with b_1. Write the formula of $\hat{y}_1(x = 4.1)$, once as a literal expression, and once with all the numbers plugged in. *Example: If D={(-1,2),(-1.1,-3), (-7,2)}, and $x = -1$, $\hat{y}_1(x = -1) = \frac{1\times(-2)+1\times(-3)}{1+1}$.* Draw $\hat{y}_1(x)$ for $x \in [-1, 6]$.

c. Let $\hat{y}_{1/2}(x)$ be the Nadaraya-Watson kernel regression result for $\mathcal{D}$ with $b_{1/2}$. Write the formula of $\hat{y}_{1/2}(x = 4.1)$, once as a literal expression, and once with all the numbers plugged in. Draw $\hat{y}_{1/2}(x)$ for $x \in [-1, 6]$.

d. Let $\hat{y}_{1NN}(x)$ be the 1-NN regression result for $\mathcal{D}$. Write the formula of $\hat{y}_{1NN}(x = 4.1)$, with all the numbers plugged in. Draw $\hat{y}_{1NN}(x)$ for $x \in [-1, 6]$.

e. Let $\hat{y}_{2NN}(x)$ be the 2-NN regression result for $\mathcal{D}$. Write the formula of $\hat{y}_{2NN}(x = 4.1)$, with all the numbers plugged in. Draw $\hat{y}_{2NN}(x)$ for $x \in [-1, 6]$.

f. What is the set $\operatorname{supp} \hat{y} = \{x \in \mathbb{R},\, \hat{y}(x) \neq 0\}$ for $\hat{y} \in \{\hat{y}_1, \hat{y}_{1/2}, \hat{y}_{1NN}, \hat{y}_{2NN}\}$? (no proofs required)

[1] Names in problems are often those of past 391 TA's.