Homework 1

STATS205 (Fall 2019–2020)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something's true (by algebra, convexity, etc.). There's no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived. **Up to two additional points will awarded for especially well-organized and elegant solutions.**

Due date: Wed, Oct 9th, 2019, 11pm.

1. Implementation and analysis of nonparametric regression methods (*15 points*)

You will be asked to implement the nonparametric regression methods covered in the first lecture on various datasets. The goal is to help you understand the similarity and difference between these methods and the bias-variance tradeoff effect of changing the bandwidth parameter h.

In the lecture, we mostly focused on estimating $r(x_i)$ for $i = 1, \ldots, n$ because in our theoretical formulation of the problem, the evaluation criteria of the estimator $\hat{r}$ only depend on the estimation on $x_1, \ldots, x_n$.[1] However, empirically, all the methods can be extended naturally to estimating $r(x)$ for $x \neq x_i$. For example, the kernel estimator for $r(x)$ can be written as

$$\hat{r}(x) = \frac{\sum_{j=1}^{n} K(\frac{x-x_j}{h})Y_i}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})} \tag{1}$$

You will be asked to plot $\hat{r}(x)$ as a function of x in various sub-questions in the sequel.
Note: You will need to run several estimators for many datasets and many bandwidth parameters. Therefore, it would be convenient to modularize your code for the efficiency of coding.

a. (*3 points*) Download (A) the data on fragments of glass collected in forensic work from the book website for Exercise 5.3 and (B) the motorcycle data from the book websitefor Exercise 5.4. Implement the following three algorithms with the specified parameters.

1. Regressogram (1 point).
   - For dataset (A): use $[a, b]$ with $a = 0.29$ and $b = 3.50$ as the boundary of the data, and use bin width $h = (b - a)/m = 0.321$ where $m = 10$ is the number of bins. (In other words, you should divide $[a, b]$ into $m = 10$ bins with equal sizes.)
   - For dataset (B): use $[a, b]$ with $a = 2.4$ and $b = 57.6$ as the boundary of the data, and use bin width $h = (b - a)/m = 2.76$ where $m = 20$.
2. Local averaging (local kernel estimator with the boxcar kernel) (1 point).
   - For dataset (A): use bandwidth $h = 0.321$. Compute $\hat{r}(x)$ on an interval $[a, b]$ with $a = 0.29$ and $b = 3.50$.
   - For dataset (B): use bandwidth $h = 2.76$. Compute $\hat{r}(x)$ on an interval $[a, b]$ with $a = 2.4$ and $b = 57.6$.
3. Kernel estimator with Gaussian kernel (1 point).
   - For dataset (A): use bandwidth $h = 0.321$. Compute $\hat{r}(x)$ on an interval $[a, b]$ with $a = 0.29$ and $b = 3.50$.

[1] This is because with a fixed set of datapoints $x_1, \ldots, x_n$, there is no convenient way to conclude something about the estimation on datapoints other than $x_1, \ldots, x_n$ theoretically.

- For dataset (B): use bandwidth $h = 2.76$. Compute $\hat{r}(x)$ on an interval $[a, b]$ with $a = 2.4$ and $b = 57.6$.

For each of the two datasets, plot the following information (with title = the dataset's name and with a proper legend) in a single figure:

- the data points in the dataset (as a scatter graph)
- $\hat{r}(x)$ for the entire range of $[a, b]$. You can use discretization the interval with $N = 200$ evenly-spaced x's for plotting.

So in summary, you should submit two figures, each of which contains three curves and a scatter plot of the data points.

b. (*12 points*) **Understanding the bias-variance tradeoff.**

In this part of the problem, you will be asked to use synthetic data to plot the bias and the variance of the nonparametric regression estimators for various choices of h and various different types of datasets.

Methodology: You will be given the form of the true $r(x)$ and the noise distribution. Throughout the rest of the homework, we will assume the noise ξ_i are i.i.d Gaussian random variable with variance σ^2 for all the x_i's, and the choice of σ will be given to you. Recall that in the class, given $r(x)$ and the choice of σ, we can *analytically* derive the expected estimation $\mathbb{E}[\hat{r}(x)]$ (where the expectation is over the randomness of ξ_i's):

$$\mathbb{E}[\hat{r}(x)] = \text{the result of applying the estimator to the ``clean data''} \ (x_1, r(x_1)), \ldots, (x_n, r(x_n)) \tag{2}$$

In other words, given the function $r(\cdot)$, we can first compute $(x_1, r(x_1)), \ldots, (x_n, r(x_n))$ (numerically in the computer) and then compute $\mathbb{E}[\hat{r}(x)]$ by running the estimator on $(x_1, r(x_1)), \ldots, (x_n, r(x_n))$. Then, we can compute the bias of the estimator, which was defined in the class as:

$$\text{bias} \triangleq \frac{1}{n} \sum_{i=1}^{n} (r(x_i) - \mathbb{E}[\hat{r}(x_i)])^2 \tag{3}$$

Moreover, recall that in the class we have analytically derived the variance of $\hat{r}(x)$

$$\text{var}(\hat{r}(x)) = \frac{\sum_{j=1}^{n} w_j^2}{(\sum_{j=1}^{n} w_j)^2} \sigma^2 \tag{4}$$

where $w_j = K(\frac{x - x_j}{h})$. Recall the variance of the estimator is defined to be

$$\text{variance} = \frac{1}{n} \sum_{i=1}^{n} \text{var}(\hat{r}(x_i)) \tag{5}$$

Therefore, we can use the formula above to numerically compute the variance of the estimator. Recall that MSE of the estimator can be decomposed into

$$\text{MSE} = \text{bias} + \text{variance} \tag{6}$$

Thus, we can compute the MSE of the estimator through the equation (6) and the series of equations above.

We will refer to the quantities above as the analytical bias, the analytical variance, and analytical MSE, etc. (The use of the word analytical is to contrast with a sampling approach described below. However, we note that even though the formulas above are derived by analytical approaches, but the evaluation of the formulas require numerical computations.)

In all of the synthetic datasets in the following sub-questions, for a given n, the inputs/covariants of the n samples are

$$x_1 = \frac{1}{n}, x_2 = \frac{2}{n}, \cdots, x_n = 1 \tag{7}$$

1. (2 points) We consider the following synthetic datasets (C): $r(x) = x$, $\sigma = 1$ and $n = 100$.

   (i) (1 point) Plot the values of $\mathbb{E}\left[\hat{r}(x)\right]$ as a function of x for the **regressogram** and the **local averaging** in the range of $(0, 1]$ using the analytical approach above. In the same figure, also plot the true function $r(x)$. You can discretize the interval $[0, 1]$ into $N = 200$ evenly-spaced points for plotting. Here the number of bins for the regressogram is $m = 10$, and the bandwidth h for the local averaging is $1/10$. We assume the bins of the regressograms are $B_1 = (0, 1/10]$, $B_2 = (1/10, 2/10], \cdots, B_{10} = (9/10, 1]$. Recall the regressogram estimator is: for every $x \in B_i$,

   $$\hat{r}(x) = \frac{1}{n_i} \sum_{j: x_j \in B_i} Y_j \tag{8}$$

   (ii) (1 point) Which method has a smaller bias? Discuss the intuitions for why it has a smaller bias than the other method. (The answer does not have to be fully rigorous. Discussing the intuitions in 1-3 sentences would suffice.)

2. (2 points) Consider the following synthetic dataset (D): $r(x) = \sin(2\pi x)$, $\sigma = 1$, and $n = 99$.

   - (1 point) Compute the MSE of the kernel estimator with Gaussian kernel and bandwidth $h = 0.1$ by using the analytical approach outlined above, and report its value.
   - (1 point) There is another more straightforward way to estimate the MSE by a sampling approach. We can generate $T = 100$ independent datasets randomly from the model $Y_i = r(x_i) + \xi_i$. We run the estimator on all of them, and compute the average of the squared losses on all of these datasets. Recall that the squared loss is defined as:

   $$\frac{1}{n} \sum_{i=1}^{n} (r(x_i) - \hat{r}(x_i))^2 \tag{9}$$

   Report the estimates of the MSE using the sampling methods with $T = 100$. You are supposed to see a number that is quite close to the analytical MSE. (If you like, you can check the convergence of the estimated MSE to the analytical MSE by increasing the number of trials T. But you are not required to do this.)

3 (2.5 points) Consider the synthetic datasets (D'): $r(x) = \sin(10\pi x)$, $\sigma = \mathbf{0.1}$, and $n = 99$. You will be asked to plot the mean of the estimator, $\mathbb{E}[\hat{r}(x)]$, as a function of x, and plot the standard deviation of the estimatte, $\sqrt{\mathrm{var}(\hat{r}(x))}$, as the "error bar". Concretely, in the figure, there should be three curves for a) $\mathbb{E}[\hat{r}(x)]$, b) $\mathbb{E}[\hat{r}(x)] + \sqrt{\mathrm{var}(\hat{r}(x))}$, and c) $\mathbb{E}[\hat{r}(x)] - \sqrt{\mathrm{var}(\hat{r}(x))}$. The curves b) and c) should have a different line color from a), as shown in the Figure 1, which demonstrates the desired format of the figure. Note: You may refer to `matplotlib.pyplot.fill_between()` to fill in the shades if you use python to plot. We are not very strict for the exact format, but your figure should contain an equivalent amount of information and be readable

   - (0.5 point) Generate a figure that contains the curves described above, and the true function $r(x)$, for $h = 0.001$. Also generate $n = 99$ random datapoints and plot them in the same figure. (These data points are not needed to produce any of the curves, and they are included just to give us a sense about what the data look like.)
   - (0.5 point) Generate another figure for $h = 0.05$ with the same specifics as above.

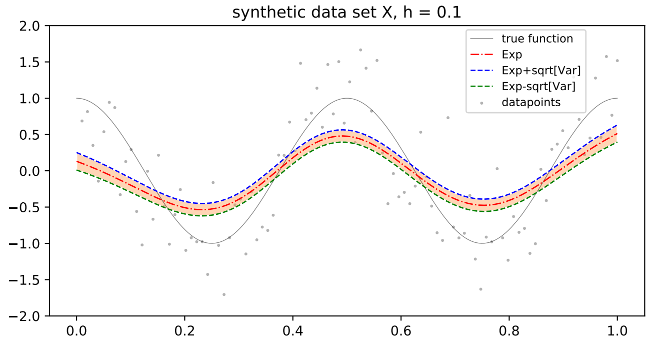

Figure 1: An exemplary figure that demonstrates the desired format. We are not very strict for the exact format, but your figure should contain an equivalent amount of information and be readable.

- (0.5 point) Generate another figure for $h = 0.2$ with the same specifics as above.
- (0.5 point) Generate another figure for $h = 100$ with the same specifics as above.
- (0.5 point) Among the choices of h above, which one has the best bias-variance tradeoff? (You are supposed to be able to tell from the figures.) Explain intuitively why.

4. (2.5 points) Consider the following synthetic datasets (E): $r(x) = 0.5$, $\sigma = 1$, and $n = 99$.

   - (0.5 point) Generate a figure with the same format asked in part b.3. above, for $h = 0.001$
   - (0.5 point) Generate another figure for $h = 0.05$.
   - (0.5 point) Generate another figure for $h = 0.2$.
   - (0.5 point) Generate another figure for $h = 100$.
   - (0.5 point) Among the choices of h above, which one has the best bias-variance tradeoff? (You are supposed to be able to tell from the figures.) Explain intuitively why.

5. (3 points) In this part, we choose the bandwidth to achieve the best bias-variance tradeoff on the synthetic datasets (E): $r(x) = \cos(2\pi x)$, $\sigma = 1$. We will vary the choice of n in $\{5, 20, 80, 320, 1280\}$. For every n, you are asked to choose the best bandwidth among $\{0.02, 0.03, 0.048, 0.063, 0.08, 0.1, 0.12, 0.15, 0.19\}$ that minimizes the analytical MSE. We denote the best choice of h (among the given choices of h) for the datasets with n examples by h_n.

   - (2 points) Report the value of h_n for all n.
   - (1 point) You are supposed to observe that h_n is decreasing as n increases. Use the theory learned in the class to explain why this is the case.