Machine Learning: Homework Assignment 4

E4525 Spring 2018,

IEOR, Columbia University

Due: March 1st, 2019

1. **Bias-Variance For Density Estimation** Let's assume we have $x \in \mathbb{R}$ distributed with unknown probability density $p(x)$. We sample points x_i for $i = 1, \ldots, N$ at random, and consider the following the Bernoulli random variable

$$s_i = \begin{cases} 1, & \text{if } |x_i - x_0| < \frac{h}{2} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

As discussed in class, the probability that $s_i = 1$ is given by

$$\theta = P(s_i = 1) = \int_{-\frac{h}{2}}^{\frac{h}{2}} \mathrm{d}u \, p(x_0 + u). \tag{2}$$

Our sample estimate for the density $p(x_0)$ is given by

$$\hat{p}_h(x_0) = \frac{\hat{\theta}}{h} = \frac{\hat{N}_1}{hN} \tag{3}$$

where $\hat{N}_1 = \sum_i s_i$.

The average square error of our density estimate at point x_0 is

$$\mathcal{E}_h(x_0) = \mathbb{E}_D\left[\left\{\hat{p}_h(x_0) - p(x_0)\right\}^2\right] = \left\{p(x) - \bar{p}_h(x_0)\right\}^2 + \mathrm{Var}\left[\hat{p}\right]$$

where the expectation is taken over all possible data samples $D = \{x_i\}_{i=1}^N$ and

$$\begin{aligned} \bar{p}_h(x_0) &= \mathbb{E}_D[\hat{p}_h(x_0)] \\ \mathrm{Var}\left[\hat{p}_h(x_0)\right] &= \mathbb{E}_D\left[\left\{\hat{p}_h(x_0) - \bar{p}_h(x)\right\}^2\right] \end{aligned} \tag{4}$$

Assuming that h is small, prove that, to leading order on h

(a)

$$\bar{p}_h(x_0) = p(x_0) + \frac{h^2}{24}\frac{\mathrm{d}^2 p(x_0)}{\mathrm{d}x^2} \tag{5}$$

Given equation 3 the expected value of $\hat{p}$ is

$$\bar{p}_h(x_0) = \frac{\theta}{h} \tag{6}$$

expanding p in powers of u up two second order we find

$$p(x_0 + u) = p(x_0) + \frac{\mathrm{d}p(x_0)}{\mathrm{d}x}u + \frac{1}{2}\frac{\mathrm{d}^2 p(x_0)}{\mathrm{d}x^2}u^2. \tag{7}$$

Integrating equation 2 we find

$$\theta = p(x_0)h + \frac{1}{3}\frac{\mathrm{d}^2 p(x_0)}{\mathrm{d}x^2}\left(\frac{h}{2}\right)^3 \tag{8}$$

and, therefore,

$$\bar{p}_h(x_0) = p(x_0) + \frac{h^2}{24}\frac{\mathrm{d}^2 p(x_0)}{\mathrm{d}x^2} \tag{9}$$

(b)

$$\mathrm{Var}\left[\hat{p}_h(x_0)\right] = \frac{p(x_0)}{Nh} \tag{10}$$

From equation 3 we have that

$$\mathrm{Var}[\hat{p}(x_0)] = \frac{\mathrm{Var}[\hat{N}_1]}{N^2 h^2} \tag{11}$$

because $\hat{N}_1$ is the sum of N independent Bernoulli variables, with $p(s_i = 1) = \theta$ we have that

$$\mathrm{Var}[\hat{N}_1] = N\theta(1-\theta) = Nhp(x_0) + O(h^2) \tag{12}$$

and, therefore, to leading order in h

$$\mathrm{Var}[\hat{p}(x_0)] = \frac{p(x_0)}{Nh} \tag{13}$$

(c) that the average square error of the density estimate is

$$\mathcal{E}_h(x_0) = Ah^4 + \frac{B}{Nh} \tag{14}$$

for some coefficients A and B that do not depend on N or h.

[HINT] You don't need to worry about the precise expressions for A and B. Just show there are there are some coefficients that do not depent on N or h.

From the previous two exercises

$$p(x_0) - \bar{p}(x_0) = \frac{1}{24}\frac{\mathrm{d}^2 p(x_0)}{\mathrm{d}x^2}h^2$$
$$\mathrm{Var}[\hat{p}(x_0)] = \frac{p(x_0)}{Nh} \tag{15}$$

where

$$c = \frac{1}{24}\frac{\mathrm{d}^2 p(x_0)}{\mathrm{d}x^2}. \tag{16}$$

Therefore

$$\mathcal{E}_h(x_0) = \{p(x_0) - \bar{p}(x_0)\}^2 + \mathrm{Var}[\hat{p}(x_0)] = Ah^4 + \frac{B}{Nh} \tag{17}$$

where $A = c^2$ and $B = p(x_0)$

(d) Show that the optimal h that minimizes $\mathcal{E}_h(x_0)$ is given by

$$\tilde{h} = FN^{-\frac{1}{5}} \tag{18}$$

for some constant F that does not depend on N.
The optimal h will satisfy

$$\frac{\mathrm{d}\mathcal{E}_h(x_0)}{\mathrm{d}h} = 0 = 4Ah^3 - \frac{B}{Nh^2} \tag{19}$$

solving for h we find

$$\tilde{h} = \left(\frac{B}{4A}\frac{1}{N}\right)^{\frac{1}{5}} = FN^{-\frac{1}{5}} \tag{20}$$

where

$$F = \left(\frac{B}{4A}\right)^{\frac{1}{5}} \tag{21}$$

(e) Show that the expected error at the optimal $\tilde{h}$ is

$$\mathcal{E}_{\tilde{h}}(x_0) = N^{-\frac{4}{5}} \tag{22}$$

for some other constant H.
Substituting $\tilde{h}$ from equation 18 into equation 14 we find

$$\mathcal{E}_{\tilde{h}}(x_0) = AF^4N^{-\frac{4}{5}} + \frac{B}{F}\frac{1}{NN^{-\frac{1}{5}}} = HN^{-\frac{4}{5}} \tag{23}$$

where

$$H = AF^4 + \frac{B}{F} \tag{24}$$

2. **Naive Bayes for Exponential Distribution Data**

   Let's assume

   - $x \in \mathcal{X} = \mathbb{R}_+^D$ so that, for each dimension $d = 1, \cdots, D$, x_d is a possitive real number $0 < x_d$.
   - $y = 1, \ldots, K$ is a categorical variable.
   - **Naive Bayes Assumption**: conditional of the value of y, x_d and $x_{d'}$ are independent provided $d \neq d'$.
   - Conditional on $y = k$, x_d has an exponential distribution

$$p(x_d|y = k) = \begin{cases} \lambda_{d,k} e^{-\lambda d,k x_d} & \text{if } 0 < x_d \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

   [HINT] Using the $z_{i,k}$ the one hot encoding of y_i will probably simplify some of the answers below

   Given a data sample $\{y_i, x_{i,d}\}$ for $i = 1, \ldots, N$ where samples are independent of each other:

   (a) Derive an expression for the maximum likelihood estimate for the parameter $\hat{\lambda}$ of the exponential distribution in terms of the data $\{x_i\}$, where $x_i \in \mathbb{R}_+$.

   The probability of observing x_i, given the parameter λ is

$$p(x_i; \lambda) = \lambda e^{-\lambda x_i} \tag{26}$$

   Therefore the average log likelihood of the observations is

$$\hat{l}(\lambda; \{x_i\}) = \frac{1}{N} \sum_i \log p(x_i; \lambda) = \log \lambda - \frac{\lambda}{N} \sum_i x_i \tag{27}$$

   The maximum will satisfy the first order equation

$$\frac{\hat{\partial l}}{\partial \lambda} = 0 = \frac{1}{\lambda} - \frac{1}{N} \sum_i x_i \tag{28}$$

   solving for λ we find

$$\hat{\lambda} = \frac{N}{\sum_i x_i} \tag{29}$$

   the inverse of the sample average of x_i.

   (b) Write the max likelihood estimate $\hat{\pi}_k$ for the marginal probability that $y = k$

   y is a categorical variable, the max likelihood estimate of the probability is

$$\hat{\pi}_k = \frac{\hat{N}_k}{N} = \frac{1}{N} \sum_i z_{i,k} \tag{30}$$

(c) Using the Naive Bayes assumption find maximum likelihood estimates $\hat{\lambda}_{d,k}$ for the exponential distribution parameter of dimension d, and class k

Each dimension d is independent, so we can consider them separately. Given k, we only need to consider observations were $y_i = k$, the estimate for lambda is then

$$\hat{\lambda}_{d,k} = \frac{\hat{N}_k}{\sum_i x_{i,k} z_{i,k}} \tag{31}$$

(d) Demonstrate that with the naive Bayes assumption the following equation is satisfied

$$p(y = k|x) = \frac{e^{\sum_d w_{d,k} x_d + b_k}}{\sum_{k'} e^{\sum_d w_{d,k'} x_d + b_{k'}}} \tag{32}$$

Write explicit expressions for $w_{d,k}$ and b_k in terms of $\hat{\lambda}_{d,k}$ and $\hat{\pi}_k$

Using Bayes Theorem

$$p(y = k|x) = \frac{p(x|y = k)p(y = k)}{\sum_{k'} p(x|y = k')p(y = k')} \tag{33}$$

where the max likelihood estimate of $p(y = k)$ is $\hat{\pi}_k$.

Using the Naive Bayes assumption we can write

$$p(x|y = k) = \prod_d p(x_d|y = k) = \prod_d \hat{\lambda}_{d,k} e^{-\hat{\lambda}_{d,k} x_d} \tag{34}$$

taking logs we can rewrite this as

$$p(x|y = k) = e^{\sum_d \log \hat{\lambda}_{d,k} - \hat{\lambda}_{d,k} x_d} \tag{35}$$

and therefore

$$p(y = k|x) = \frac{e^{\sum_d \log \lambda_{d,k} - \lambda_{d,k} x_d + \log \pi_k}}{\sum_{k'} e^{\sum_d \log \lambda_{d,k'} - \lambda_{d,k'} x_d + \log \pi_{k'}}} \tag{36}$$

therefore, if we define

$$\begin{aligned} w_{d,k} &= -\hat{\lambda}_{d,k} \\ b_k &= \log \pi_k + \sum_d \log \hat{\lambda}_{d,k} \end{aligned} \tag{37}$$

we finally obtain

$$p(y = k|x) = \frac{e^{\sum_d w_{d,k} x_d + b_k}}{\sum_{k'} e^{\sum_d w_{d,k'} x_d + b_{k'}}} \tag{38}$$