Machine Learning: Homework Assignment 4

E4525 Spring 2018,

IEOR, Columbia University

Due: Oct 7th, 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)

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

(b)

$$\operatorname{Var}\left[\hat{p}_h(x_0)\right] = \frac{p(x_0)}{Nh} \tag{6}$$

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

$$\mathcal{E}_h(x_0) = Ah^4 + \frac{B}{Nh} \tag{7}$$

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.

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

$$\tilde{h} = FN^{-\frac{1}{5}} \tag{8}$$

for some constant F that does not depend on N.

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

$$\mathcal{E}_{\tilde{h}}(x_0) = N^{-\frac{4}{5}} \tag{9}$$

for some other constant H.

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

$$p(x_d|y=k) = \begin{cases} \lambda_{d,k} e^{-\lambda d,k x_d} & \text{if } 0 < x_d \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

[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 $x_i \in \mathbb{R}_+$.

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

(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

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

$$p(y = k|x) = \frac{e^{\sum_d w_{d,k} x_d + b_k}}{\sum_{k'} e^{\sum_d w_{d,k'} x_d + b_{k'}}} \tag{11}$$

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