CSCI 567 Machine Learning HW1

- Ankit Kothari
USC ID:4138854460

1 Density Estimation

$(a)(i)$

$Beta\ Distribution\ f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \qquad where\ B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$and\ \Gamma(t) = (t-1)!$

$Therefore\ Beta(\alpha,\beta) = \frac{1}{\alpha}$

$f(x) = x^{\alpha-1}\alpha$

$L(\alpha|x) = \prod_{1}^{N} x^{\alpha-1}.\ \alpha$

$Taking\ log\ likelihood$

$l(\alpha|x) = \sum_{1}^{N} log(x^{(\alpha-1)}.\ \alpha)$

$l(\alpha|x) = \sum_{1}^{N} log(x^{(\alpha-1)}) + N\ log(\alpha)$

$Taking\ derivative\ wrt\ \alpha$

$= \frac{dl(\alpha)}{d\alpha} \sum_{1}^{N} \frac{1}{x^{\alpha-1}}.x^{\alpha-1} log(x) + \frac{N}{\alpha} = 0$

$= \sum_{i}^{N} log(x) + \frac{N}{\alpha} = 0$

$therefore, \alpha = \frac{-N}{\sum_{i}^{N} log(x)}$

$(a)(ii)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}.\exp\frac{-(x-\mu)^2}{2\sigma^2}$$

$Given\ \ \mu = \Theta,\ \ \sigma = \Theta$

$$f(x) = \frac{1}{\sqrt{2\pi\theta}}.\exp\frac{-(x-\theta)^2}{2\theta}$$

$Likelihood\ function\ is\ given\ by,$

$$L(\theta|x) = \prod_{1}^{N} f(X_i|\theta)$$

$Log\ Likelihood\ function\ is\ given\ by,$

$$l(\theta) = \sum_{i=1}^{N} log(\frac{1}{\sqrt{2\pi\theta}}.\exp\frac{-(x_i-\theta)^2}{2\theta})$$

$Differentiating\ wrt\ \theta,$

$$\frac{dl(\theta)}{d\theta} = \frac{-1}{2}[-\frac{N}{\theta} - \sum\frac{x^2}{\theta^2} + N]$$

$$\frac{dl(\theta)}{d\theta} = \frac{-1}{2}[-\frac{N}{\theta} - \sum\frac{x^2}{\theta^2} + N] = 0$$

$$\theta = \frac{-N \pm \sqrt[2]{N^2 + 4N\sum x^2}}{2N}$$

$$\therefore \theta = \frac{1}{2}\ \left[-1 \pm \sqrt[2]{1 + \frac{4\sum x^2}{N}}\ \right]$$

(b)

(i)

$$To\ Show\ E_{X_{i,\dots,n}}[\hat{f}(x)] = \frac{1}{h}\int K(\frac{x-t}{h})f(t)dt$$

$$\hat{f}(c) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-X_i}{h})$$

$$E\hat{f} = \frac{1}{nh}\sum_{i=1}^{n}E[K(\frac{x-X_i}{h})] \qquad As\ E\sum_{x_i} = sumE(x_i)$$

$$= \frac{1}{nh}[nE(K(\frac{x-X_i}{h}))]$$

$$= \frac{1}{h}\int K(\frac{x-X_i}{h})f(X_i dX_i) \qquad E(g(x)) = \int g(x)f(x)dx$$

$$= \frac{1}{h}\int K(\frac{x-t}{h})f(t)dt$$

(ii)

$$E[\hat{f}(x)] = \frac{1}{n}\sum_{i=1}{}"nE[\frac{1}{h}K(\frac{x-X_i}{h})]$$

$$= E[\frac{1}{h}K(\frac{x-X}{h})]$$

$$= \int dt\frac{1}{h}K(\frac{x-t}{h})f(t)$$

$$since\ z = \frac{x-t}{h} \quad \therefore -hdz = dt$$

$$= \int dzK(z)f(x-hz)\text{...(changing limits and then eliminating h and changing limits again)}$$

$$= \int dzK(z)[f(x) - hzf'(x) + \frac{h^2z^2}{2}f''(x) - \frac{h^3z^3}{3!}f'''(x) + ...]$$

$$= f(x) + \frac{h^2f''(x)}{2}\int dzK(z)z^2 + \frac{h^3f'''(x)}{3!}\int dzK(z)z^3 + ...$$

by definition $\int dzK(z) = 1\ \ and\ \ \int dzK(z)z = 0$. If we call $\int dzK(z)z^2 = \sigma_k^2$, then bias of KDE is

$$E[\hat{f}(x)] - f(x) = \frac{h^2\sigma_k^2 f''(x)}{2} + o(h^2)$$

2 DENSITY ESTIMATION

(*a*)

$$P(X_{1\ldots n}|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Using Naive Bayes Thm

$$P(Y = y_k|X_1 \ldots Xn) = \frac{P(Y = y_k)\prod_{i=1}^{n} P(X_i|Y = y_j)}{\sum P(Y = y_y)\prod_{i=1}^{n} P(X_i|Y = y_j)}$$

$$P(Y = 1|X) = \frac{\pi\prod_{i=1}^{n} P(X_i|Y = 1)}{\pi\prod_{i=1}^{n} P(X_i|Y = 1) + (1-\pi)\prod_{i=1}^{n} P(X_i|Y = 0)}$$

$$= \frac{1}{1 + \frac{(1-\pi)\prod_{i=1}^{n} P(X_i|Y=0)}{\pi\prod_{i=1}^{n} P(X_i|Y=1)}}$$

$Given\ \ P(X_i|Y = y_k)$ follows a Gaussian distribution of the form $N(\mu_{jk}, \sigma_j^2)$

$for\ \ X_{i=0}$

$$P(X_0|Y = 0) = \frac{1}{\sqrt[2]{2\pi\sigma_0^2}} \exp^{-(\frac{(x-\mu_{00})^2}{2\sigma_0^2})}\, dx$$

$$P(X_0|Y = 1) = \frac{1}{\sqrt[2]{2\pi\sigma_0^2}} \exp^{-(\frac{(x-\mu_{01})^2}{2\sigma_0^2})}\, dx$$

$ForX_i$

$$\frac{\prod P(X|Y = 0)}{\prod P(X|Y = 1)} = \exp^{-}\Big[\sum \frac{(x_i - \mu_{i0})^2 - (x_i - \mu_{i1})^2}{2\sigma_i^2}\Big]$$

$$= \exp^{-}\Big[\sum \frac{2x_i(\mu_{i1}^2 - \mu i1^2)}{2\sigma_i^2} - log(\frac{1-\pi}{\pi})\Big]$$

$$= \exp^{-}\Big[log(\frac{1-\pi}{\pi}) - \sum \frac{x_i(\mu_{i1} - \mu_{i0})}{\sigma_i^2} + \sum \frac{(\mu_{i1}^2 - \mu_{i0}^2)}{2\sigma_i^2}\Big]$$

$$\therefore w_0 = -[\frac{log(1-\pi}{\pi}) \ + \ \sum_{i=1}^{n} \frac{(\mu_{1i}^2 - \mu_{i0}^2)}{2\sigma_i^2}]$$

$$w_i = \frac{(\mu_{i0} - \mu_{i1})}{\sigma_i^2} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_k \end{bmatrix}$$

(b)

$P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$

$= P(Y = y) \prod_{d=1} DP(X_d = x_d|Y = y)$

$P(Y = k) = \pi_k \ \ for \ \ k \ E \ 0,1$ given y follows Bernoulli Distribution

$P(x_j|Y = y_k) \ N(\mu_{jk}, \sigma_{jk})$

$$log(P(x_i|Y = k)) = \frac{log(2\pi\sigma_{jk})}{2} - \frac{(x_j - x_{jk})^2}{2\sigma_{jk}}$$

Using Naive Bayes

$P(X_i = x, Y = y) = P(Y_i = y)P(X_{i1} = x_1, X_{i2} = x_2 ... X_{iD} = x_D|Y = y_k)$

$= P(Y_i = y) \prod + j = 1^D P(X_{ij} = x_i|Y = y_i)$

log likelihood function $$log(L) = \sum_{i=1}^{N} log(P(Y_i = yi)) + \sum_{i=1}^{N} \sum_{j=1}^{D} log(P(X_i = x_{ij}|Y = y_i))$$

$$= \sum k = 1^K P(Y = k)N_k + \sum_{k=1}^{K} \sum_{j=1}^{D} log(P(X_{ij} = x_{ij}|Y = k))N_k$$

Differentiating w.r.t the params $\pi, \ \mu \ and \ \sigma$

$$For \ \ \pi, \sum_k \pi_k = 1, \ \ we \ \ get \ \ \pi_k \frac{N_k}{N}$$

$$\frac{\partial \sum_{k=1}^{K} \sum_{j=1}^{D} log(P(X_i = x_{ij}||Y = k))N_k}{\partial \mu_{jk}} = 0$$

$$\mu_{jk} = \frac{\sum_{i;Y_{i=k}} x_{ij}}{N_k}$$

$$\frac{\partial \sum_{k=1}^{K} \sum_{j=1}^{D} log(P(X_i = x_{ij}||Y = k))N_k}{\partial \sigma_{jk}} = 0$$

$$\therefore, \sigma = \frac{\sum_{i;Y_{i=k}} (x_{ij} - \mu_{jk})^2}{N_k}$$

3 NEAREST NEIGHBOR

(a)

$$\bar{x} = \frac{1}{N}\sum_{1}^{13} x_i = \frac{166}{13} = 12.769$$

$$\bar{y} = \frac{1}{N}\sum_{1}^{13} y_i = \frac{160}{13} = 12.307$$

$$\sigma(x) = \frac{1}{N-1}\sum_{1}^{13}(x_i - \bar{x})^2 \quad = 20.717$$

$$\sigma(y) = \frac{1}{N-1}\sum_{1}^{13}(y_i - \bar{y})^2 \quad = 25.931$$

Subject	Normalized x	Normalized y	L2 distance	L1 distance
Mathematics	-0.616	1.415	1.899	2.585
	-0.954	0.959	1.622	2.269
	-1.051	1.338	2.088	2.943
Electrical Engineering	0.783	-0.012	0.474	0.627
	1.749	0.721	1.680	2.326
	1.170	0.991	1.453	2.017
Computer Science	-0.230	-0.127	0.585	0.657
	0.011	-0.513	0.453	0.646
	-0.906	-0.590	1.316	1.64
	-1.630	-0.012	1.990	2.172
Economics	0.687	-1.709	1.538	1.842
	0.300	-1.014	0.801	0.858
	0.687	-1.246	1.094	1.379

Distance column is the distance between new point (20,7)
normalized to (0.349,-0.205) other points (Normalized)
From the above table, For L2 distance, the new point is classified as Computer Science for K=1 and 5
For L1 distance, it is classified as Electric Engineering for K=1 and Computer Science fro K=5

(b)

(i)

$$P(x) = \sum_{i=1}^{N} P(x|Y=c)P(Y=c)$$

$$= \frac{K_1}{N_1 V}.\frac{N_1}{N} + \frac{K_2}{N_2 V}.\frac{N_2}{N} + \frac{K_3}{N_3 V}.\frac{N_3}{N} + ... + \frac{K_c}{N_c V}.\frac{N_c}{N}$$

$$= \frac{1}{VN}[\sum_{c} K_c]$$

$$\therefore p(x) = \frac{K}{VN}$$

(ii)

$$P(Y=c|x) = \frac{P(x|Y=c)P(Y=c)}{P(x)}$$

$$\frac{\frac{K_c}{N_c V} \quad \frac{N_c}{N}}{\frac{K}{VN}}$$

$$\therefore P(Y=c|x) = \frac{K_c}{K}$$

4 DECISION TREE

(i)

$$H(x) = -\sum_{i=1}^{n} p_i log(p_i)$$

$$H(y|x) = -\sum_{i=1}^{n} p_{X=X_i} H(Y|x = x_i)$$

$$= \sum_{m,n} p_{X_i,y_i} log(p(y_i|x_i))$$

$$I(y;x) = H(y) - H(y|x)$$

$$H(AccRate) = -[\frac{73}{100} log(\frac{73}{100}) + \frac{27}{100} log(\frac{27}{100})] = 0.841465$$

$$I(Acc;Weather) = H(Acc) - H(Acc|Weather)$$

$$I(Acc;Traffic) = H(Acc) - H(Acc|Traffic)$$

$$H(Acc|Weather) = -[[\frac{23}{28} log(\frac{23}{28}) + \frac{5}{28} log(\frac{5}{28})] \frac{28}{100} + \frac{72}{100} [\frac{50}{72} log(\frac{50}{72}) + \frac{22}{72} log(\frac{22}{72})]]$$

$$= 0.82889$$

$$I(Acc;weather) = 0.0126$$

$$H(Acc;Traffic) = -[\frac{27}{100} (\frac{27}{27} log(\frac{27}{27}))] + [\frac{73}{100} (\frac{73}{73} log(\frac{73}{73}))]$$

$$= 0$$

$$\therefore I(Acc;Traffic) = 0.841$$

$$Hencesplitons Traffic as the information gain is max$$

(ii)

The decision tree with the normalized data will give the same out come as the one with raw data.This is because normalization does not affects the number of objects class/category it just scales the data to follow a normal curve. This does not affect the data and hence the splitting of normalized tree is same as that of the normal tree.

(iii)

$$GiniIndex = \sum_{k=1}^{k} p_k(1-p_k)$$

$$CrossEntropy = -\sum_{k=1}^{k} p_k log(p_k)$$

$$To\ \ prove, \sum_{k=1}^{k} p_k(1-p_k) <= -\sum_{k=1}^{k} p_k log(p_k)$$

$i.e\ \ p_k(1-p_k) <= -p_k log(p_k)$ for each term in the summation

$value\ \ of\ \ p_k\ \ in\ \ the\ \ range[0,1]$

consider a function p(x)=1-xlog(x) defined in the range [0,1]

to get maximum, p'x=0

$\therefore, x - frac1x = 0$

$\therefore, x = 1$ with max value=0

thus, p(x) <= 0

1-x+log(x) <=0

$\therefore$, 1-x <= -log(x)

i.e x(1-x) <= -xlog(x) ...which is what we want to prove

5 PROGRAMMING

i. There are total 10 attributes including row id which is the first column.(Class attribute is not being considered as attribute)
ii. No not all attributes are meaning full as the first record contains the row id which does not have any impact on classification as its just a number
iii. There are total of 7 classes but class 4 has no records
iv. The majority is for class 2 with 73 records. Its not a uniform distribution.

Table 1: Accuracy

Knn For K	Training L1	Training L2	Testing L1	Testing L2
1	75.0000	71.4286	66.6667	61.1111
3	74.4898	72.4490	72.2222	66.6667
5	70.9184	72.4490	61.1111	55.5556
7	69.8980	68.8776	55.5556	55.5556

* For KNN in case of a tie between classes, the point with minimum distance is considered for classification

Bayes	Training	Testing L2
1	54.5918	33.3333

As clearly seen by the outputs, KNN gives better accuracy than Naive Bayes as KNN is being non parametric makes no assumption about the data where as Naive Bayes assumes attributes are conditionally independent to each other given class values and are normally distributed. As a result Naive Bayes can have only linear boundary compared to flexible boundary KNN can has.

This assignment has been collaborated between me, Bhavya Gandhi, Dhaval Shah, Junaid Hundekar and Sagar Makwana