CSCI-567: Assignment #1

Due on Tuesday, September 23, 2015

Saket Choudhary
2170058637

Contents

Problem 1 **3**
Problem 1: (a) 1 3
Problem 1: (a) 2 4
Problem 1: (b) 1 4
Problem 1: (b) 2 5

Problem 2 **6**
Problem 2: (a) 6
Problem 2: (b) 7

Problem 3 **7**
Problem 3: (a) 7
Problem 3: (b) 8
Problem 3: (c) 10

Problem 4 **12**
Problem 4: (a) 12
Problem 4: (b) 14

Problem 5 **16**
Problem 5: (5.1) 16
Problem 5: (5.2d) 19
Problem 5: (5.2e) 20

Problem 1

Problem 1: (a) 1

Given: $X_i \sim Beta(\alpha, 1)$ MLE for α:
Consider $X = (X_1, X_2, \dots, X_n)$ Likelihood function: $L(\alpha|X)$ $L(\alpha|X) = \prod_{i=1}^{n} f(x_i)$ where

$$\begin{aligned} f(x_i) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(1)}x^{\alpha-1} \\ &= \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)}x^{\alpha-1} \\ &= \alpha x^{\alpha-1} \end{aligned}$$

$$L(\alpha|X) = (\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(1)})^n \prod_{i=1}^{n}(x_i)^{\alpha-1}$$

$$LL = \log(L(\alpha|X)) = n\log(\alpha) + (\alpha-1)\sum_{i=1}^{n} x_i$$

$$\frac{dLL}{d\alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n}\log(x_i)$$

$$\frac{dLL}{d\alpha} = 0 \implies \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} log(1/x_i)}$$

Minima at $\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} log(1/x_i)}$ is guaranteed due to log being a concave function.

$$\text{Thus, } \boxed{\alpha_{\hat{MLE}} = \frac{n}{\sum_{i=1}^{n} log(1/x_i)}}$$

Problem 1 continued on next page. . .

Problem 1: (a) 2

Given: $x_i \sim N(\theta, \theta)$ i.e $f(x_i) = (2\pi\theta)^{-\frac{1}{2}} e^{-\frac{(x_i-\theta)^2}{2\theta}}$ MLE estimate for θ:

$$L(\theta|X) = (2\pi\theta)^{-\frac{N}{2}} e^{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2\theta}}$$
$$LL = \log(L(\theta|X)) = -\frac{N}{2}\log((2\pi\theta)) - \sum_{i=1}^n \frac{(x_i-\theta)^2}{2\theta}$$
$$\frac{dLL}{d\theta} = -\frac{N}{2}(\frac{1}{\theta}) + \frac{\sum_{i=1}^n x_i^2}{2\theta^2} - \frac{N\theta}{2}$$
$$\frac{dLL}{d\theta} = 0 \implies N\theta^2 + N\theta - \sum_{i=1}^n x_i^2 = 0$$

The above equation is a quadratic and will have two solutions, Since, $\theta \geq 0$ (a constraint that comes from θ being the variance), the

$$\theta = \frac{-N \pm \sqrt{N^2 + 4N\sum_{i=1}^n x_i^2}}{2N}$$

$$\text{Since, } \hat{\theta} \geq 0 \implies \boxed{\hat{\theta_{MLE}} = \frac{-N + \sqrt{N^2 + 4N\sum_{i=1}^n x_i^2}}{2N}}$$

Problem 1: (b) 1

Given: $\hat{f(x)} = \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K(\frac{x-X_i}{h})$ To show: $E_{X_1,X_2,\dots X_n}[\hat{f(x)}] = \frac{1}{h}\int K(\frac{x-t}{h})f(t)dt$
Proof:

$$E[\hat{f(x)}] = E[\frac{1}{n}\sum_{i=1}^n \frac{1}{h}K(\frac{x-X_i}{h})]$$
$$= \frac{1}{nh}E[K(\frac{x-X_i}{h})]$$
$$= \frac{1}{h}E[K(\frac{x-X_1}{h})]$$
$$= \frac{1}{h}E[K(\frac{x-t}{h})]$$

where the penultimate equality comes from the fact that X_i are iid for all $i \in [1, n]$. and t X and hence.

$$E[\hat{f(x)}] = \frac{1}{h}E[K(\frac{x-X_1}{h})]$$
$$= \frac{1}{h}\int K(\frac{x-t}{h})f(t)dt$$
$$= RHS$$

Problem 1: (b) 2

Consider $z = \frac{x-t}{h} \implies t = x - hu$
Then,

$$E[\hat{f(x)}] = \frac{1}{h}\int K(z)f(x-hz)dz$$

$$f(x-hz) = f(x) - f'(x)hz + \frac{1}{2}f''(x)\frac{(hz)^2}{2} - \frac{1}{3}f'''(x)\frac{(hz)^3}{3!} + \cdots + (-1)^n\frac{1}{n!}f^{(n)}(x)(\frac{(hu)^n}{n!})$$

By definition, $\int k(z)dz = 1$.
Also define an auxillary variable $M_j = \int k(z)z^j dz$ for the j^{th} moment of the kernel function, and hence,

$$\int K(z)f(x-hz)dz = f(x) - hf'(x)M_1 + \frac{1}{2}(h^2)f^{''}(x)M_2 + \cdots + (-1)^n\frac{1}{n!}f^{(n)}M_n$$

Now,

$$Bias = E[\hat{f(x)}] - f(x) = -hf'(x)M_1 + \frac{1}{2}(h^2)f^{''}(x)M_2 + \cdots + (-1)^n\frac{1}{n!}f^{(n)}M_n$$

$$M_1 = \int xK(x)dx = 0$$

Hence,

$$\boxed{E[\hat{f(x)}] - f(x) = \frac{1}{2}(h^2)f^{''}(x)M_2 + \cdots + (-1)^n\frac{1}{n!}f^{(n)}M_n \text{ where } M_j = \int k(z)z^j dz}$$

And as $h \longrightarrow 0$, $Bias \longrightarrow 0$

Problem 2

Problem 2: (a)

Mean $\bar{x} = \frac{1}{N}\sum x = \frac{1}{10}\sum_{i=1}^{10} x_i = 8.6$
Mean $\bar{y} = \frac{1}{N}\sum y = \frac{1}{10}\sum_{i=1}^{10} y_i = 19.6$
Standard deviation $x_{sd} = \sqrt{\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{10-1}} = 21.3269$
Standard deviation $y_{sd} = \sqrt{\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{10-1}} = 25.1960$
Student with unknown major: $(9, 18)$:
Normalised to : $(0.0187, -0.0635)$

ID	x	y	x_n	y_n	L1	L2
M1	10	49	0.0623	1.107	1.2117	1.168
M2	-12	38	-0.9163	0.6928	1.6871	1.1998
M3	-9	47	-0.7829	1.0317	1.8926	1.354
EE1	29	19	0.9074	-0.0226	0.9272	0.8904
EE2	32	31	1.0409	0.4292	1.5125	1.1341
EE3	37	38	1.2633	0.6928	1.9985	1.4554
CS1	8	9	-0.0267	-0.3991	0.3834	0.3418
CS2	30	-28	0.9519	-1.7922	2.6661	1.9678
CS3	-18	-19	-1.1832	-1.4534	2.5942	1.8394
CS4	-21	12	-1.3167	-0.2862	1.5605	1.3535

Procedure: We first normalise the data point with unknown major using the mean and standard deviation of the known points, and then calculated the L1 and L2 distances. L1 distance between two points (x_1, y_1) and (x_0, y_0) is defined as : $L1 = |x_1 - x_0| + |y_1 - y_0|$
L2 distance is defined as $L2 = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$

For $L1$:
$K = 1$: For $K = 1$ the nearest neighbor is CS1 and hence the unknown sample 'could' be a computer science
$K = 3$: For $K = 3$ the nearest neighbors are M1, EE1, CS1 and hence there is a 'tie'. Choosing the label of the least distance would again result in CS1 as $CS1 < EE1 < M1$.

For $L2$:
$K = 1$: For $K = 1$ the nearest neighbor is CS1 and hence the unknown sample 'could' be a computer science
$K = 3$: For $K = 3$ the nearest neighbors are M1, EE1, EE2. Since two nearest neighbors are from EE, we assign it the unknown sample to be from Electrical engineering.

Comparison For $K = 1$, both $L1$ and $L2$ distance metric give the same results, however for $K = 3$, the $L1$ metric yields a tie, since the distances are similar but $L2$ metric being a square quantity of a number smaller than 1 further reduces the distances. The fact to realise is that $|x + y|$ is similar to $\sqrt{x^2 + y^2}$ when $x, y << 1$ that is when the points are close, but when x, y are large, the L2 metric is going to be higher, and hence L2 norm applies more 'penalty' to distant points in the sense that they are larger.
This also implies that in case of outliers, while $L1$ norm will not penalise, $L2$ norm will penalise in the sense that $\sqrt{x^2 + y^2}$ will be high. In case of outliers L2 is more robust. In this example, 'M1' is likely an outlier.

Problem 2: (b)

Total points: N
Total points with label class c: N_c
Given: $p(x|Y = c) = \frac{K_c}{N_c V}$ and $\sum K_c = K$ Class prior: $p(Y = c) = \frac{N_c}{N}$
Unconditional density $p(x) = \sum_c p(x|Y = c)p(Y = c) = \sum_c \frac{K_c}{N_c V} \times \frac{N_c}{N} = \sum_c \frac{K_c}{NV} = \frac{K}{NV}$
Posterior $\boxed{P(Y = c|x) = \frac{P(x|Y = c) \times P(Y = c)}{P(x)} = \frac{\frac{K_c}{N_c V} \times \frac{N_c}{N}}{\frac{K}{NV}} = \frac{K_c}{K}}$

Problem 3

Problem 3: (a)

Information gain $G = H[Y] - H[Y|X]$ where Y is the outcome variable and X is an attribute to be split. In our case Y= 'Rains or not' In order to maximise gain for a fixed Y we need to minimise the conditional entropy $H[Y|X]$ $p_{rain} = \frac{9+5+6+3+7+2+3+1}{80} = frac3680 = 0.9$ and hence $p_{no-rain} = 0.1$

$$\begin{aligned} H[Y] &= -p_{rain} \log(p_r ain) - p_{no-rain} \log(p_{no-rain}) \\ &= -(0.9 \log_2(0.9) + 0.1 \log_2(0.1) \\ &= 0.99277 \end{aligned}$$

Instead of maximising gains, it is sufficient to simply minimise the conditional entropy $H[Y|X]$ in this case.
For Temperature

$$\begin{aligned} \text{H[Rainy or Not rainy — Temperature]} &= -p_{hot} \times (p_{rain,hot} \log(p_{rain,hot}) + p_{norain,hot} \log(p_{norain,hot}) \\ &- p_{cold} \times (p_{rain,cold} \log(p_{rain,cold}) + p_{norain,cold} \log(p_{norain,cold})) \\ &= -\frac{1}{2} \times (\frac{23}{40} \log(\frac{23}{40}) + \frac{13}{40} \log(\frac{13}{40})) \\ &- \frac{1}{2} \times (\frac{13}{40} \log(\frac{13}{40}) + \frac{27}{40} \log(\frac{27}{40})) \\ &= 0.9467 \end{aligned}$$

Similarly,

$$\begin{aligned} \text{H[Rain or Not rainy — Sky]} &= \frac{1}{2}\text{H[Rainy or Not rainy — Sky=Cloudy]} \\ &+ \frac{1}{2}\text{H[Rainy or Not rainy — Sky=Clear]} \\ &= 0.9015 \end{aligned}$$

Similarly,

$$\begin{aligned}\text{H[Rain or Not rainy — Humidity]} &= \frac{1}{2}\text{H[Rainy or Not rainy — Humidity=High]}\\ &+ \frac{1}{2}\text{H[Rainy or Not rainy — Humidity=Low]}\\ &= 0.9467\end{aligned}$$

We choose Sky as the *root* because $H[RainorNotrainy|Sky]$ is minimum.
Further on,

$$\begin{aligned}\text{H[Rainy or not—Cloudy,Temperature]} &= -\frac{1}{2}(\frac{15}{20}\log(\frac{15}{20}) + \frac{5}{20}\log(\frac{5}{20}))\\ &- \frac{1}{2}(\frac{10}{20}\log(\frac{10}{20}) + \frac{10}{20}\log(\frac{10}{20}))\\ &= 0.9056\end{aligned}$$

$$\begin{aligned}\text{H[Rainy or not—Cloudy,Humidity]} &= -\frac{1}{2}(\frac{16}{20}\log(\frac{16}{20}) + \frac{4}{20}\log(\frac{4}{20}))\\ &- \frac{1}{2}(\frac{9}{20}\log(\frac{9}{20}) + \frac{11}{20}\log(\frac{11}{20}))\\ &= 0.8574\end{aligned}$$

Thus, we choose condition Humidity for Cloudy Sky.
Further more,

$$\begin{aligned}\text{H[Rainy or not—Clear,Temperature]} &= 0.7903\\ \text{H[Rainy or not—Clear,Humidity]} &= 0.8280\end{aligned}$$

So we choose $Temperature$ condition for Clear Sky. Now for $Temperature = Hot, Sky = Clear$ we have $p_{rain} = \frac{8}{20}$, and $p_{norain} = \frac{12}{20}$. So we choose 'Not Rainy' for $Temperature = Hot$ branch.
For, $Temperature = Cold, Sky = Clear$ we have $p_{rain} = \frac{3}{20}$ and hence once again we choose 'Not Rainy' for $Temperature = Cold$ branch.
For $Humidity = High, Sky = Cloudy$, $p_{rain} = \frac{16}{20}$ and hence we choose $Rain$ for $Humidity = High, Sky = Cloudy$
And, finally for $Humidity = Low, Sky = Cloudy$, $p_{rain} = \frac{9}{20}$ and hence we choose $Not\ Rainy$.
The decision and the pruned decision tree is shown in Figure 1,2.

Problem 3: (b)

Consider $f(p_k) = (1 - p_k) - (-\log p_k)$ We know that $0 \leq p_k \leq 1$ Then $f'(p_k) = -1 + \frac{1}{p_k} = -\frac{1-p_k}{p_k} \leq 0 \forall p_k \in [0, 1]$
And hence $f'(p_k)$ is a non-increasing function which $\implies f(p_k) \geq f(1) \forall p_k \in (0, 1]$ and hence, $(1 - p_k) - (-\log p_k) \geq 0 \implies p_k(1 - p_k) - (-p_k \log p_k) \geq 0 \implies p_k(1 - p_k) \geq -p_k \log p_k \implies \sum_{k=1}^{K} p_k(1 - p_k) \geq \sum_{k=1}^{K} -p_k \log p_k \implies$ Gini index is less than corresponding value of Cross Entropy

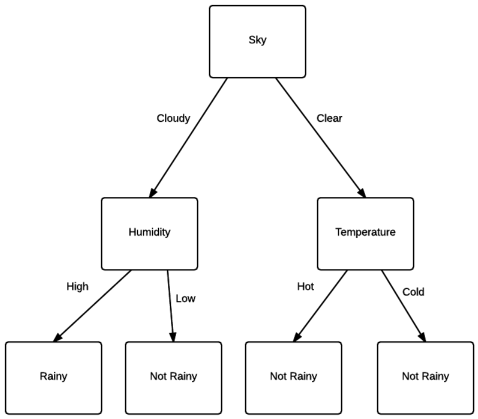


Figure 1: Problem 3 Decision Tree

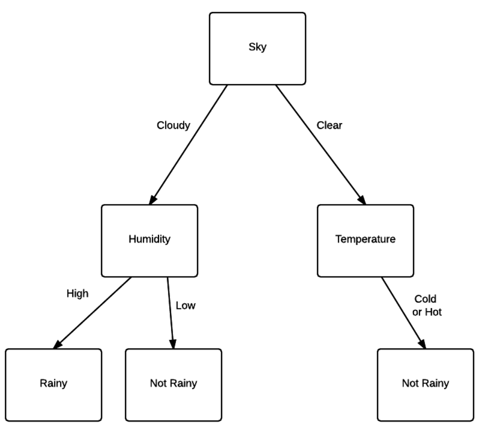


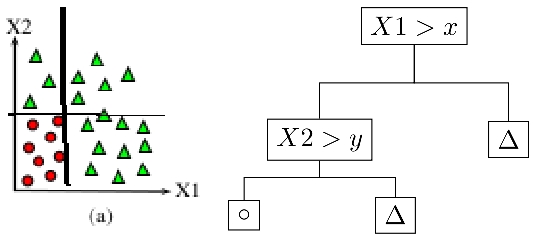
Figure 2: Problem 3 Pruned Decision Tree

Problem 3 continued on next page. . .

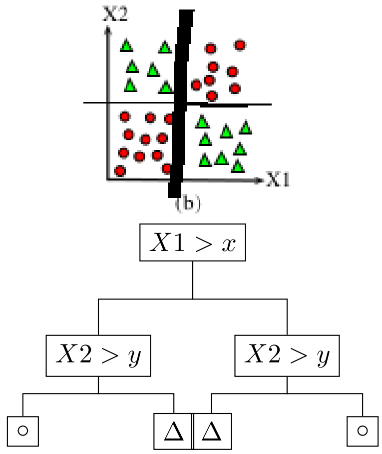
Problem 3: (c)

By default the rightmost branch corresponds to the parent condition being a YES[FIX ME]

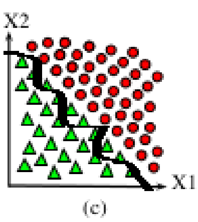
a:



b:

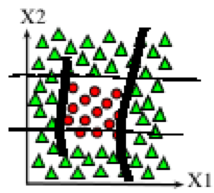


c:

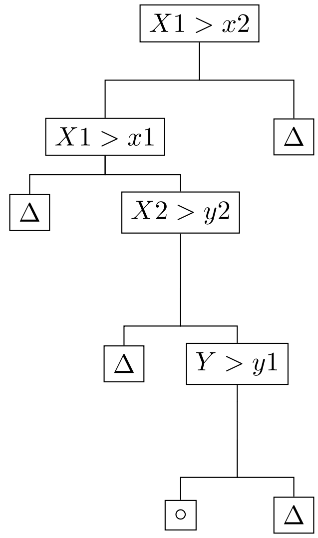


The only case where it is not possible to have a depth 6 decision tree is case (c). The decision boundary in this case is a 'zig-zag' ladder and hence the depth of decision tree is unbounded.

d:



(d)



Problem 4

Problem 4: (a)

Given a random variable $X \in \mathbf{R}^D$ and $Y \in [C]$ naive bayes defines the joint distribution:

$$\begin{aligned} P(X = x, Y = y) &= P(Y = y)P(X = x|Y = y) \\ &= P(Y = y)\prod_{d=1}^{D} P(X_d = x_d|Y = y) \end{aligned}$$

Y is a categorical variable with $P(Y = k) = p_k$ for $k \in [1, K]$
Given: $P(x_j|Y = y_k) \sim N(\mu_{jk}, \sigma_{jk}) \implies$

$$\log(P(x_j|Y = k)) = -\frac{\log(2\pi\sigma_{jk})}{2} - \frac{(x_j - x_{jk})^2}{2\sigma_{jk}} \tag{1}$$

For all $j \neq j'$ $x_j, X_{j'}$ are independent attributes.
Naive bayes:

$$\begin{aligned} P(X_i = \vec{x}, Y_i = y) &= P(Y_i = y)P(X_{i1} = x_1, X_{i2} = x_2 \dots X_{iD} = x_D|Y = y_k) \\ &= P(Y_i = y)\prod_{j=1}^{D} P(X_{ij} = x_j|Y = y_i) \text{ Assuming independence of attributes } x_i \end{aligned}$$

Each Y_i belongs to one of the K classes, thus $\sum_{k=1}^{K} p_k = 1$ for any Y_i
Let N_k represent the number of elements in class k for $k \in [1, K]$
Then, $\sum_{i=1}^{N} \log(Y_i = y_i) = \sum_{k=1}^{K} P(Y = k) \times N_k$
Consider the likelihood function

$$\begin{aligned} L(\mu, \sigma, p|(X, Y)) &= \prod_{1}^{N} P(Y_i = y_i) \times \prod_{j=1}^{D} P(X_i = x_{ij}|Y = y_i) \\ \log(L) &= \sum_{i=1}^{N} \log(P(Y_i = y_i)) + \sum_{i=1}^{N}\sum_{j=1}^{D} \log(P(X_i = x_{ij}|Y = y_i)) \\ &= \sum_{i=1}^{N} \log(P(Y_i = y_i)) + \sum_{i=1}^{N}\sum_{j=1}^{D} \log(P(X_{ij} = x_{ij}|Y = y_i)) \\ &= \sum_{k=1}^{K} P(Y = k) \times N_k + \sum_{k=1}^{K}\sum_{j=1}^{D} \log(P(X_{ij} = x_{ij}|Y = k)) \times N_k \end{aligned}$$

Now,

$$\frac{\partial LL}{\partial p_k} = 0 \tag{2}$$

$$\frac{\partial LL}{\partial \mu_{jk}} = 0 \tag{3}$$

$$\frac{\partial LL}{\partial \sigma jk} = 0 \tag{4}$$

$$\tag{5}$$

For for equation 2 and constraint $\sum_k p_k = 1$, we get: $p_k \frac{N_k}{N}$

For equation 3,

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

$$\frac{\sum_{i;Y_i=k}(x_{ij} - \mu_{jk})}{\sigma_{jk}} = 0$$

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

For equation 4,

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

$$\frac{\partial}{\partial \sigma_{jk}} \sum_{i;Y_i=k} \left(- \frac{\log(2\pi\sigma_{jk})}{2} - \frac{(x_{ij} - x_{jk})^2}{2\sigma_{jk}} \right) = 0$$

$$\frac{\partial}{\partial \sigma_{jk}} \sum_{i;Y_i=k} \left(- \frac{1}{\sigma_{jk}} + \frac{(x_{ij} - x_{jk})^2}{2\sigma_{jk}^2} \right) = 0$$

$$\hat{\sigma_{jk}} = \frac{\sum_{i;Y_i=k}(x_{ij} - \hat{\mu_{jk}})^2}{N_k}$$

Constraint $\boxed{\sum_k p_k = 1}$ Given K number of classes, the above constraint the MLE estimate of p_k is given by: $\boxed{\frac{N_k}{N}}$

Problem 4: (b)

Given: $P(Y=1)=\pi$; For X_j feature, $P(X_j=x_j|Y_k)=\theta_{jk}^{x_j}(1-\theta_{jk})^{1-x_j}$

$$\begin{aligned}
P(Y=1|X) &= \frac{P(X|Y=1)P(Y=1)}{P(X)}\\
&= \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}\\
&= \frac{1}{1+\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}}\\
&= \frac{1}{1+\exp(\log(\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}))}\\
&= \frac{1}{1+\exp(\log(P(X|Y=0)P(Y=0))-\log(P(X|Y=1)P(Y=1)))}\\
&= \frac{1}{1+\exp(-(\log(\frac{P(Y=1)}{P(Y=0)}))+\log(P(X|Y=0))-\log(P(X|Y=1)))}
\end{aligned}$$

Now assuming features satisfy the independence property, $P(X|Y=1)=\prod_{j=1}^{D}P(X_j|Y=1)=\prod_{j=1}^{D}\theta_{jk}^{x_j}(1-\theta_{jk})^{1-x_j}$
Alternatively,

$$\log(P(X_j|Y=1)) = \log(\theta_{j1}^{x_j}(1-\theta_{j1})^{1-x_j}) \tag{6}$$
$$= x_j\log(\theta_{j1})+(1-x_j)\log((1-\theta_{j1}) \tag{7}$$
$$= x_j\log(\frac{\theta_{j1}}{1-\theta_{j1}})+\log(1-\theta_{j1}) \tag{8}$$

and,

$$\log(P(X_j|Y=0)) = \log(\theta_{j0}^{x_j}(1-\theta_{j0})^{1-x_j}) \tag{9}$$
$$= x_j\log(\theta_{j0})+(1-x_j)\log((1-\theta_{j0}) \tag{10}$$
$$= x_j\log(\frac{\theta_{j0}}{1-\theta_{j0}})+\log(1-\theta_{j0}) \tag{11}$$

Hence,

$$\log(P(X_j|Y=0)-\log(P(X_j|Y=1))=x_j\log(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})})+\log\frac{(1-\theta_{j0})}{(1-\theta_{j1})} \tag{12}$$

$\implies$

$$\log(P(X|Y=0)-\log(P(X|Y=1))=\sum_{j=1}^{D}x_j\log(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})})+\sum_{j=1}^{D}\log\frac{(1-\theta_{j0})}{(1-\theta_{j1})} \tag{13}$$

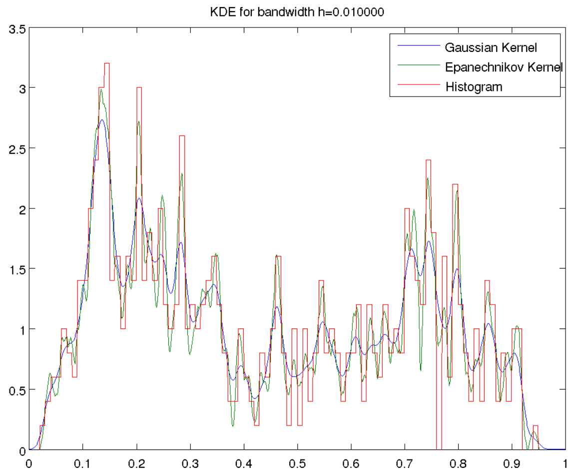
$$
\begin{aligned}
-(\log(\frac{P(Y=1)}{P(Y=0)})) + \log(P(X|Y=0) - \log(P(X|Y=1)) &= \sum_{j=1}^{D} x_j \log(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})}) + \\
&+ \big(\log(\frac{(1-\theta_{j0})}{(1-\theta_{j1})} + -(\log(\frac{P(Y=1)}{P(Y=0)}))\big) \\
&= \sum_{j=1}^{D} x_j \log(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})}) + \\
&+ \big(\log(\frac{(1-\theta_{j0})}{(1-\theta_{j1})} + (\log(\frac{P(Y=0)}{P(Y=1)}))\big) \\
&= \sum_{j=1}^{D} x_j \log(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})}) + \\
&+ \big(\log(\frac{(1-\theta_{j0})}{(1-\theta_{j1})} + (\log(\frac{(1-\pi)}{\pi}))\big)
\end{aligned}
$$

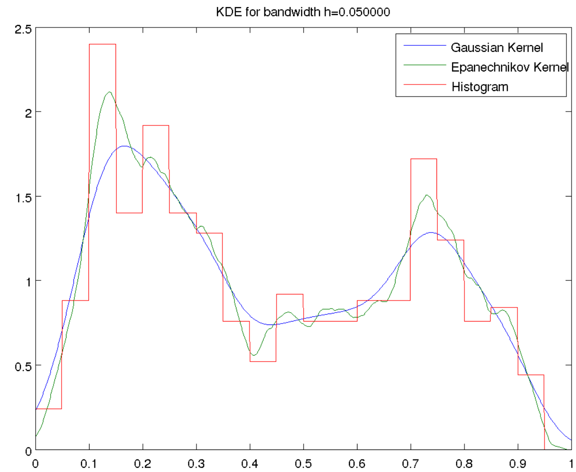
$$\text{And hence } \boxed{\vec{w}_j = \log(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})})}$$

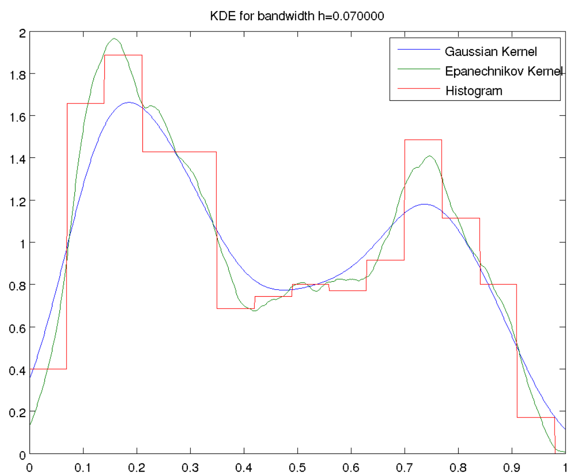
$$\boxed{w_0 = -\log(\frac{1-\pi}{\pi} \times (\frac{1-\theta_{j0}}{1-\theta_{j1}})^D)}$$

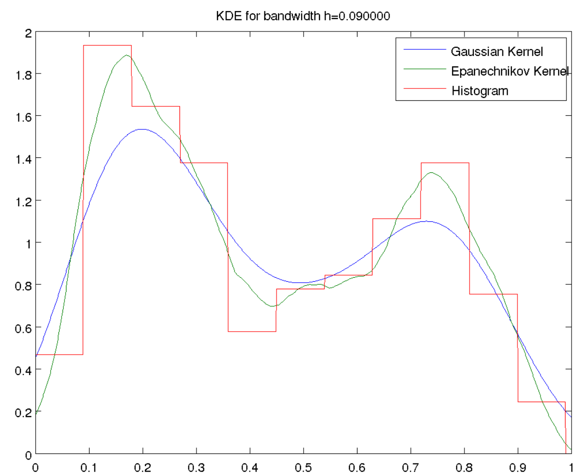
Problem 5

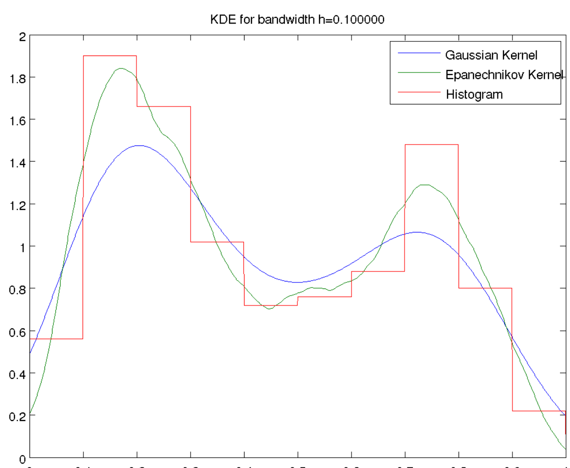
Problem 5: (5.1)

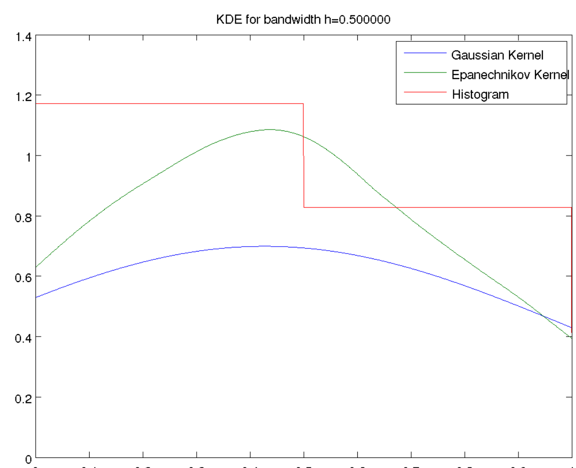












Problem 5 [Problem 5: (5.1)] continued on next page. . .

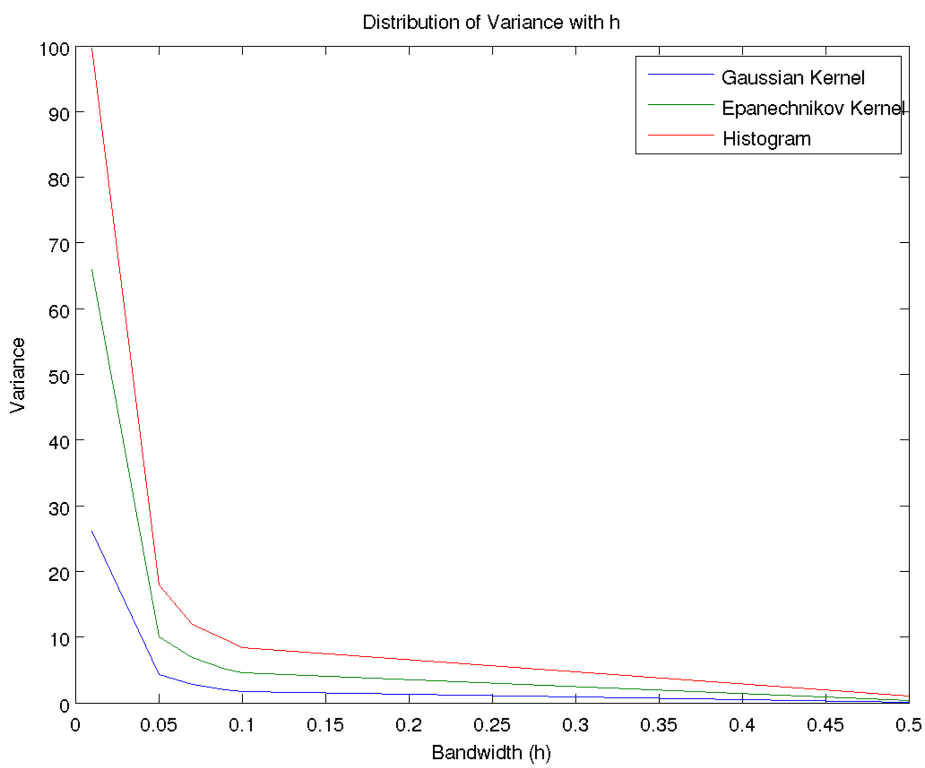


Figure 3: Integrated Variance distribution with respect to h for various kernels

From figure 3 we see that Gaussian kernel guarantees the minal variance for all values of h. Histogram consistently leads to higher variance and Epanechnikov kernels' variance is bounded between the former two. And hence we conclude Gaussian kernel outperforms Epanechnikov kernel which outperforms histogram for kernel density estimation(based on the criteria of minimising the variance)
An optimum value of h is found at the *knee* of the graph and is approximately $h = 0.1$. Selecting the *knee* guarantees minimal varaince and an optimum choice for a smaller h, because higher h are guaranteed to minimise the variance.

Problem 5 continued on next page. . .

Problem 5: (5.2d)

Accurcy table for k-nn

K	Training Accuracy	Validation Accuracy	Testing Accuracy
K=1	0.788973	0.759259	0.776744
K=3	0.768061	0.717593	0.744186
K=5	0.798479	0.782407	0.753488
K=7	0.861217	0.842593	0.809302
K=9	0.897338	0.902778	0.865116
K=11	0.908745	0.907407	0.883721
K=13	0.866920	0.861111	0.823256
K=15	0.817490	0.782407	0.762791

Based on the maximum training accuracy, we can choose K to be 11

Gini Index = GDI
Cross Entropy = DEV
Accuracy table for Decision Tree

MinLeaf	Training(GDI)	Training(DEV)	Validation(GDI)	Validation(DEV)	Testing(GDI)	Testing(DEV)
1	0.950664	0.950664	0.870370	0.842593	0.860465	0.865116
2	0.950664	0.950664	0.870370	0.842593	0.860465	0.865116
3	0.950664	0.950664	0.870370	0.842593	0.860465	0.865116
4	0.948767	0.948767	0.875000	0.847222	0.883721	0.883721
5	0.941176	0.939279	0.875000	0.856481	0.874419	0.869767
6	0.935484	0.933586	0.861111	0.842593	0.888372	0.883721
7	0.927894	0.925996	0.870370	0.851852	0.897674	0.893023
8	0.920304	0.918406	0.856481	0.837963	0.897674	0.893023
9	0.914611	0.882353	0.879630	0.824074	0.860465	0.823256
10	0.912713	0.880455	0.856481	0.800926	0.888372	0.851163

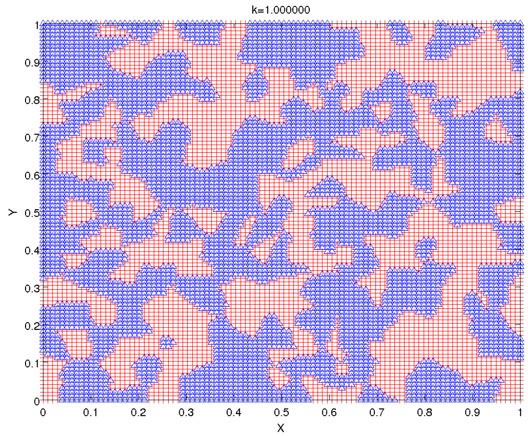
Naive Bayes Classifier

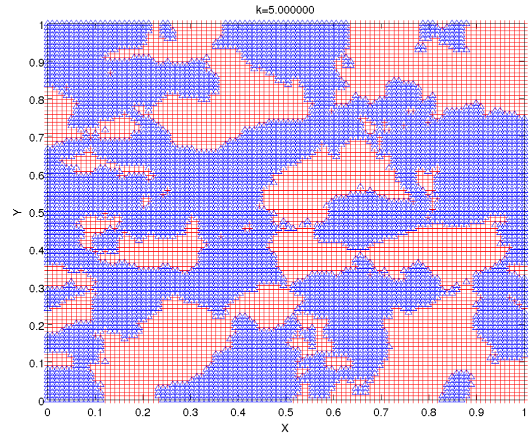
Dataset	Training Accuracy	Validation Accuracy	Testing Accuracy
Nursery	0.9065	0.9059	0.8970
TTT	0.7827	0.7130	0.7302

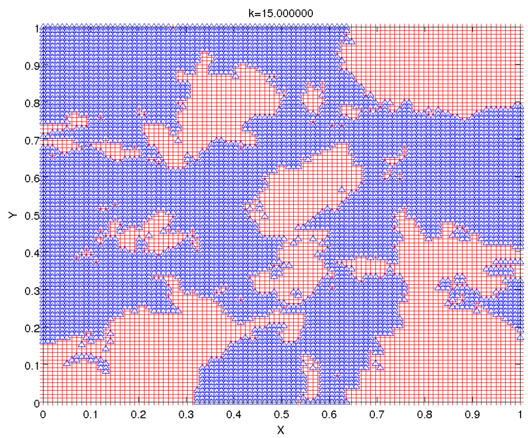
As evident from the results for Naive Bayes, it performs pretty good on Nursery dataset, since its attributes can be treated to be independent. On the other hand, tic tac toe dataset has an inherent dependence built in that decides the labels. For example, a XXX is a 'Positive' for sure and XXO can never be. This violates the inherent assumption of independence of features in Naive Bayes and hence the lower accuracy.

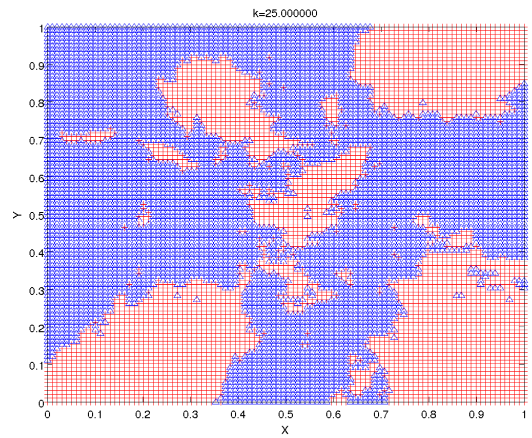
Problem 5 continued on next page. . .

Problem 5: (5.2e)









As evident from the plots above, increasing K results in more smooth boundaries. This is expected,because k increasing leads to more neihbors being weighted for deciding the final label, and this will often involve neighbors that are far apart, thus creating the soft boundaries