

# **CSCI-567: Assignment #1**

Due on Tuesday, September 23, 2015

**Saket Choudhary**  
**2170058637**

## Contents

<b>Problem # 1</b>	<b>3</b>
Problem # 1: (a) 1 . . . . .	3
Problem # 1: (a) 2 . . . . .	3
Problem # 1: (b) 1 . . . . .	4
Problem # 1: (b) 2 . . . . .	4
<b>Problem 2</b>	<b>5</b>
Problem 2: (a) . . . . .	5
Problem 2: (b) . . . . .	6
<b>Problem 3</b>	<b>6</b>
Problem 3: (a) . . . . .	6
Problem 3: (c) . . . . .	6
Problem 3: (c) . . . . .	7
<b>Problem 4</b>	<b>9</b>
Problem 4: (a) . . . . .	9
Problem 4: (b) . . . . .	11
<b>Problem 5</b>	<b>13</b>
Problem 5: (5.1) . . . . .	13
Problem 5: (5.2d) . . . . .	15
Problem 5: (5.2e) . . . . .	16

## Problem # 1

### Problem # 1: (a) 1

Given:  $X_i \sim \text{Beta}(\alpha, 1)$  MLE for  $\alpha$ :

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

$$f(x_i) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} = \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)\Gamma(1)} x^{\alpha-1} \quad (1)$$

$$= \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} x^{\alpha-1} = \alpha x^{\alpha-1} \quad (2)$$

$$L(\alpha|X) = \left( \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)\Gamma(1)} \right)^n \prod_{i=1}^n (x_i)^{\alpha-1} \quad (3)$$

$$LL = \log(L(\alpha|X)) = n \log(\alpha) + (\alpha - 1) \sum_{i=1}^n \log(x_i) \quad (4)$$

$$\frac{dLL}{d\alpha} = \frac{n}{\alpha} + \sum_{i=1}^n \log(x_i) \quad (5)$$

$$\frac{dLL}{d\alpha} = 0 \implies \hat{\alpha} = \frac{n}{\sum_{i=1}^n \log(1/x_i)} \quad (6)$$

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

### 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  $\theta$ :

$$L(\theta|X) = (2\pi\theta)^{-\frac{N}{2}} e^{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2\theta}} \quad (7)$$

$$LL = \log(L(\theta|X)) = -\frac{N}{2} \log((2\pi\theta)) - \sum_{i=1}^n \frac{(x_i - \theta)^2}{2\theta} \quad (8)$$

$$\frac{dLL}{d\theta} = -\frac{N}{2} \left( \frac{1}{\theta} \right) + \frac{\sum_{i=1}^n x_i^2}{2\theta^2} - \frac{N\theta}{2} \quad (9)$$

$$\frac{dLL}{d\theta} = 0 \implies N\theta^2 + N\theta - \sum_{i=1}^n x_i^2 = 0 \quad (10)$$

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

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

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

**Problem # 1: (b) 1**

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

Proof:

$$E[f(\hat{x})] = E[\frac{1}{n} \sum_{i=1}^n \frac{1}{h} K(\frac{x-X_i}{h})] \quad (11)$$

$$= \frac{1}{nh} E[K(\frac{x-X_i}{h})] \quad (12)$$

$$= \frac{1}{h} E[K(\frac{x-X_1}{h})] = \frac{1}{h} E[K(\frac{x-t}{h})] \quad (13)$$

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

$$E[f(\hat{x})] = \frac{1}{h} E[K(\frac{x-X_1}{h})] = \frac{1}{h} \int K(\frac{x-t}{h}) f(t) dt = RHS \quad (14)$$

**Problem # 1: (b) 2**

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

Then,

$$E[f(\hat{x})] = \frac{1}{h} \int K(z) f(x - hz) dz \quad (15)$$

$$(16)$$

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

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

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

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

## 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  $y_{sd} = \sqrt{\frac{\sum_{i=1}^{10} (y_i - \bar{y})^2}{10-1}} = 25.1960$   
 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	<u>1.2117</u>	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	<u>0.9272</u>	<u>0.8904</u>
EE2	32	31	1.0409	0.4292	1.5125	<u>1.1341</u>
EE3	37	38	1.2633	0.6928	1.9985	1.4554
CS1	8	9	-0.0267	-0.3991	<u>0.3834</u>	<u>0.3418</u>
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 and ..... TODO

**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  $P(Y=c|x) = \frac{P(x|Y=c) \times P(Y=c)}{P(x)} = \frac{\frac{K_c}{N_c V} \times \frac{N_c}{N}}{\frac{K}{NV}} = \frac{K_c}{K}$

**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  $H[Y|X]$   $p_{rain} = \frac{9+5+6+3+7+2+3+1}{80} = \frac{36}{80} = 0.45$  and hence  $p_{no-rain} = 0.55$

$$H[Y] = -p_{rain} \log(p_{rain}) - p_{no-rain} \log(p_{no-rain}) \quad (17)$$

$$= -(0.45 \log_2(0.45) + 0.55 \log_2(0.55)) \quad (18)$$

**Temperature**

**Problem 3: (c)**

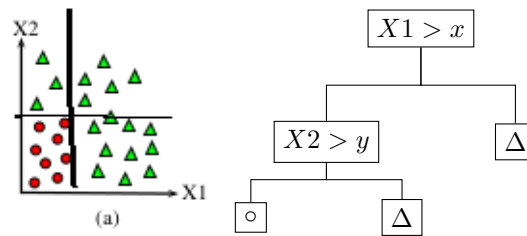
Consider  $f(p_k) = (1 - p_k) - (-\log p_k)$  We know that  $0 \leq p_k \leq 1$  Then  $f'(p_k) = -1 + \frac{1}{p_k} = -\frac{1-p_k}{p_k} \leq 0 \forall p_k \in [0, 1]$

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

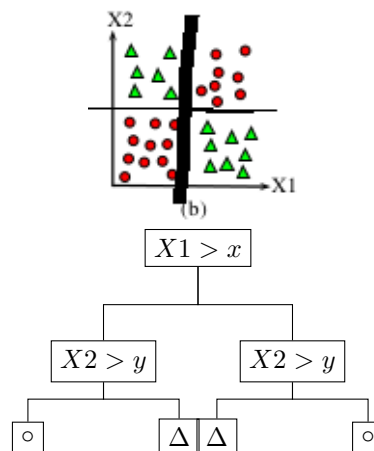
**Problem 3: (c)**

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

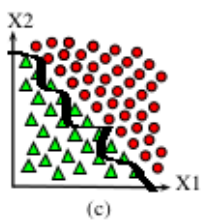
a:



b:

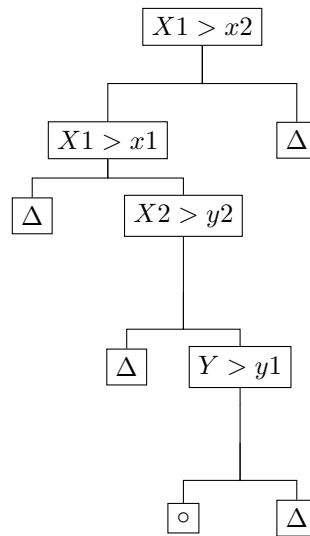
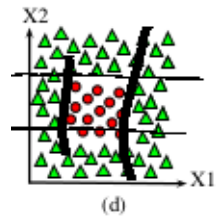


c:



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:





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

$$\log(P(x_j|Y = k)) = -\frac{\log(2\pi\sigma_{jk})}{2} - \frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}} \quad (19)$$

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

$$L(\mu, \sigma, p|(X, Y)) = \prod_{i=1}^N P(Y_i = y_i) \times \prod_{j=1}^D P(X_i = x_{ij}|Y = y_i) \quad (20)$$

$$\log(L) = \sum_{i=1}^N \log(P(Y_i = y_i)) + \sum_{i=1}^N \sum_{j=1}^D \log(P(X_i = x_{ij}|Y = y_i)) \quad (21)$$

$$= \sum_{i=1}^N \log(P(Y_i = y_i)) + \sum_{i=1}^N \sum_{j=1}^D \log(P(X_{ij} = x_{ij}|Y = y_i)) \quad (22)$$

$$= \sum_{k=1}^K P(Y = k) \times N_k + \sum_{k=1}^K \sum_{j=1}^D \log(P(X_{ij} = x_{ij}|Y = k)) \times N_k \quad (23)$$

Now,

$$\frac{\partial LL}{\partial p_k} = 0 \quad (24)$$

$$\frac{\partial LL}{\partial \mu_{jk}} = 0 \quad (25)$$

$$\frac{\partial LL}{\partial \sigma_{jk}} = 0 \quad (26)$$

$$(27)$$

For for equation 24 and constraint  $\sum_k p_k = 1$ , we get:  $p_k \frac{N_k}{N}$

For equation 25,

$$\begin{aligned} \frac{\partial \sum_{k=1}^K \sum_{j=1}^D \log(P(X_i = x_{ij}|Y = k)) \times N_k}{\partial \mu_{jk}} &= 0 \\ \frac{\sum_{i; Y_i=k} (x_{ij} - \mu_{jk})}{\sigma_{jk}} &= 0 \\ \hat{\mu}_{jk} &= \frac{\sum_{i; Y_i=k} x_{ij}}{N_k} \end{aligned}$$

For equation 26,

$$\begin{aligned} \frac{\partial \sum_{k=1}^K \sum_{j=1}^D \log(P(X_i = x_{ij}|Y = k)) \times N_k}{\partial \sigma_{jk}} &= 0 \\ \frac{\partial}{\partial \sigma_{jk}} \sum_{i; Y_i=k} \left( -\frac{\log(2\pi\sigma_{jk})}{2} - \frac{(x_{ij} - \mu_{jk})^2}{2\sigma_{jk}^2} \right) &= 0 \\ \frac{\partial}{\partial \sigma_{jk}} \sum_{i; Y_i=k} \left( -\frac{1}{\sigma_{jk}} + \frac{(x_{ij} - \mu_{jk})^2}{2\sigma_{jk}^3} \right) &= 0 \\ \hat{\sigma}_{jk} &= \frac{\sum_{i; Y_i=k} (x_{ij} - \hat{\mu}_{jk})^2}{N_k} \end{aligned}$$

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

**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,  $P(X|Y = 1) = \prod_{j=1}^D P(X_j|Y = 1) = \prod_{j=1}^D \theta_{j1}^{x_j} (1 - \theta_{j1})^{1-x_j}$

Alternatively,

$$\log(P(X_j|Y = 1)) = \log(\theta_{j1}^{x_j} (1 - \theta_{j1})^{1-x_j}) \quad (28)$$

$$= x_j \log(\theta_{j1}) + (1 - x_j) \log((1 - \theta_{j1})) \quad (29)$$

$$= x_j \log\left(\frac{\theta_{j1}}{1 - \theta_{j1}}\right) + \log(1 - \theta_{j1}) \quad (30)$$

and,

$$\log(P(X_j|Y = 0)) = \log(\theta_{j0}^{x_j} (1 - \theta_{j0})^{1-x_j}) \quad (31)$$

$$= x_j \log(\theta_{j0}) + (1 - x_j) \log((1 - \theta_{j0})) \quad (32)$$

$$= x_j \log\left(\frac{\theta_{j0}}{1 - \theta_{j0}}\right) + \log(1 - \theta_{j0}) \quad (33)$$

Hence,

$$\log(P(X_j|Y = 0) - \log(P(X_j|Y = 1))) = x_j \log\left(\frac{\theta_{j0}(1 - \theta_{j1})}{\theta_{j1}(1 - \theta_{j0})}\right) + \log\left(\frac{1 - \theta_{j0}}{1 - \theta_{j1}}\right) \quad (34)$$

$\Rightarrow$

$$\log(P(X|Y = 0) - \log(P(X|Y = 1))) = \sum_{j=1}^D x_j \log\left(\frac{\theta_{j0}(1 - \theta_{j1})}{\theta_{j1}(1 - \theta_{j0})}\right) + \sum_{j=1}^D \log\left(\frac{1 - \theta_{j0}}{1 - \theta_{j1}}\right) \quad (35)$$

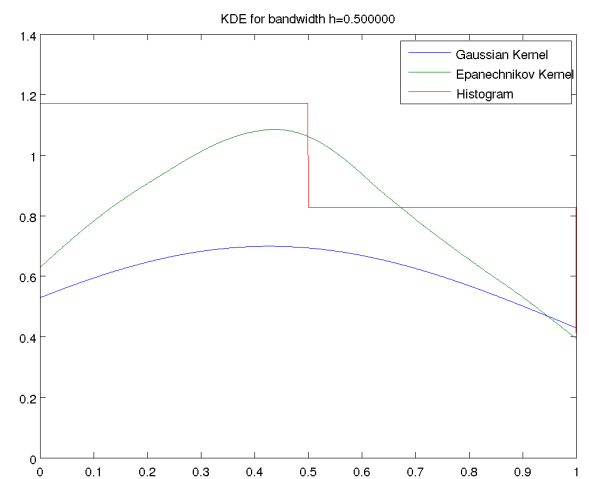
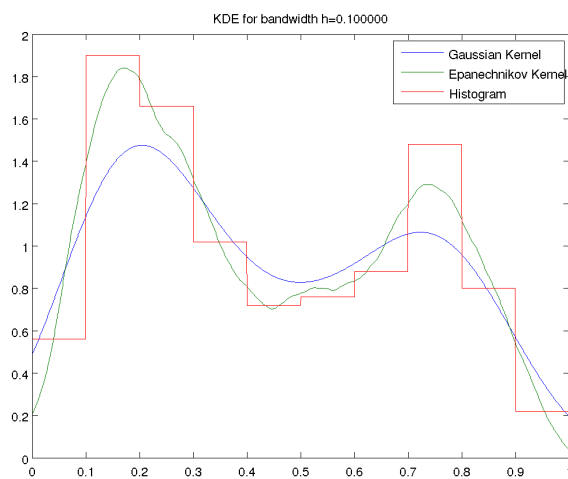
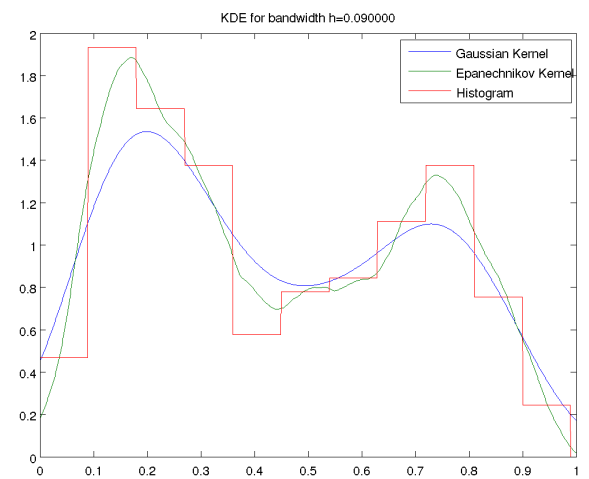
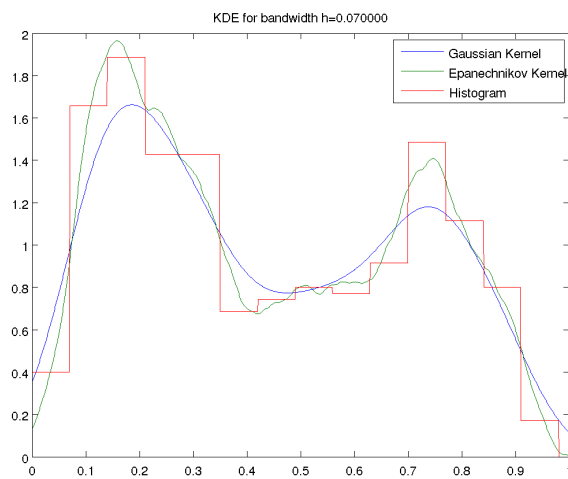
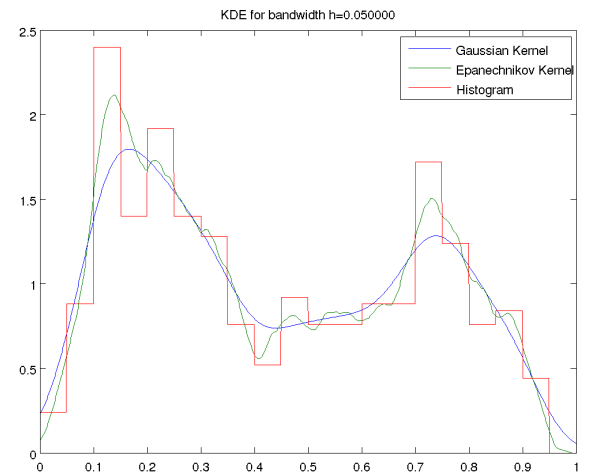
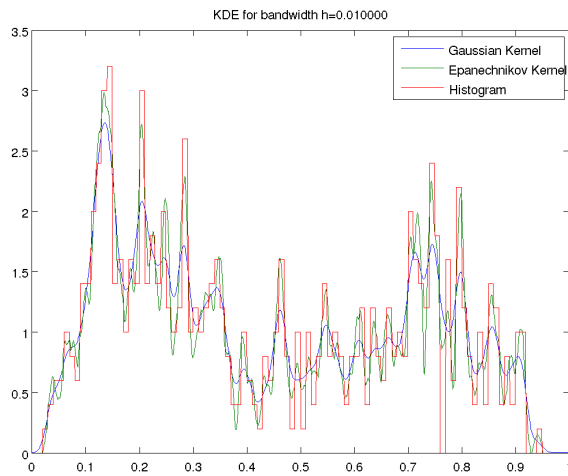
$$\begin{aligned}
-\left(\log\left(\frac{P(Y=1)}{P(Y=0)}\right)\right) + \log(P(X|Y=0) - \log(P(X|Y=1))) &= \sum_{j=1}^D x_j \log\left(\frac{\theta_{j0}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j0})}\right) + \\
&+ \left(\log\left(\frac{(1-\theta_{j0})}{(1-\theta_{j1})}\right) + -\left(\log\left(\frac{P(Y=1)}{P(Y=0)}\right)\right)\right)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

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

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

## Problem 5

### Problem 5: (5.1)



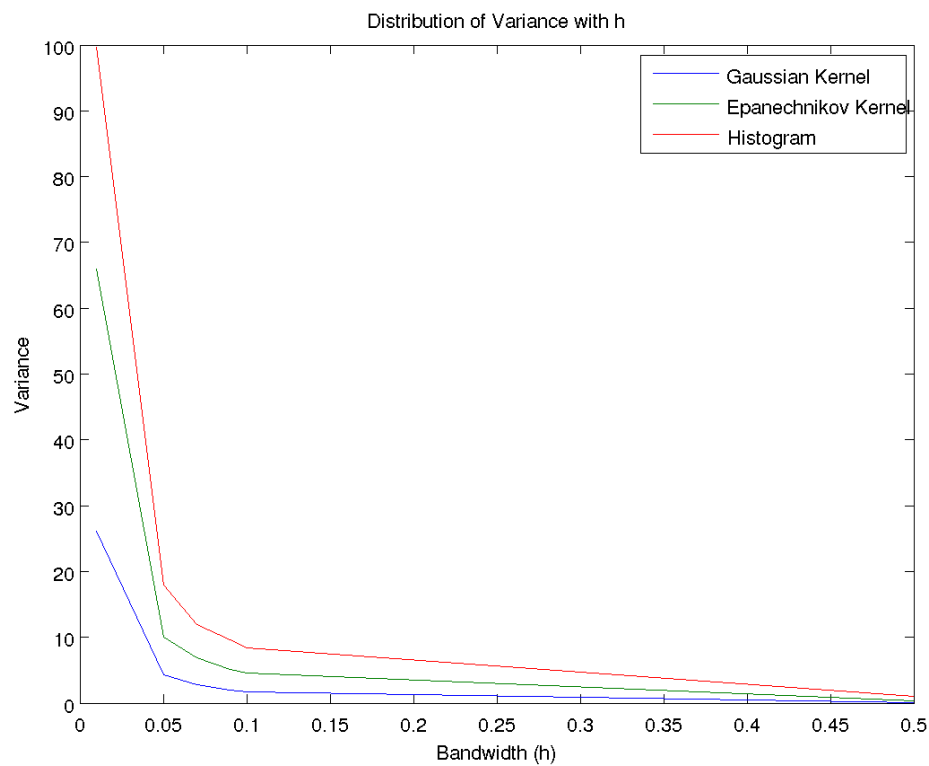


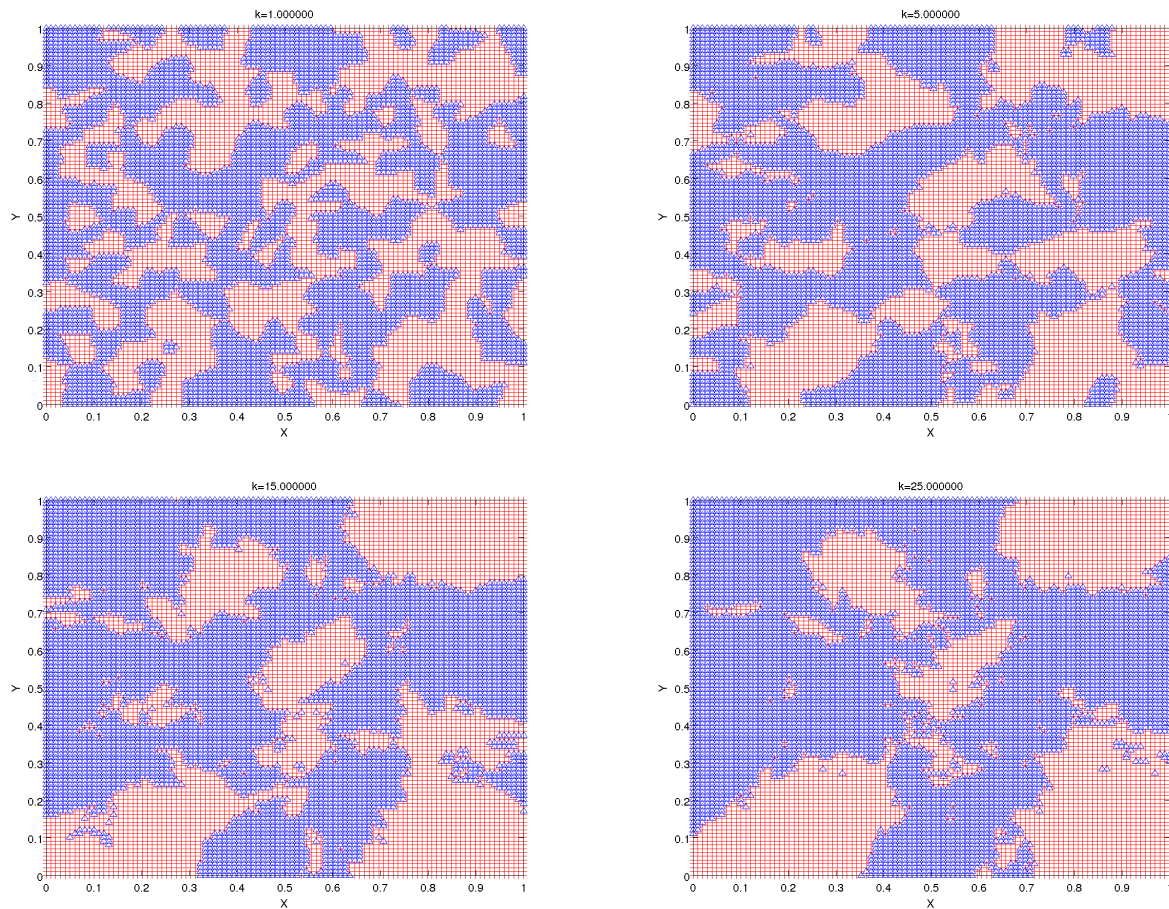
Figure 1: Variance distribution with respect to h for various kernels

From figure 1 we see that Gaussian kernel guarantees the minimal 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 variance and an optimum choice for a smaller  $h$ , because higher  $h$  are guaranteed to minimise the variance.

### Problem 5: (5.2d)

		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		
		MinLeaf	Training(GDI)	Training(deviance)	Validation(GDI)		
1.000000	0.950664	0.950664	0.870370	0.842593	0.860465	0.865116	
2.000000	0.950664	0.950664	0.870370	0.842593	0.860465	0.865116	
3.000000	0.950664	0.950664	0.870370	0.842593	0.860465	0.865116	
4.000000	0.948767	0.948767	0.875000	0.847222	0.883721	0.883721	
5.000000	0.941176	0.939279	0.875000	0.856481	0.874419	0.869700	
6.000000	0.935484	0.933586	0.861111	0.842593	0.888372	0.883721	
7.000000	0.927894	0.925996	0.870370	0.851852	0.897674	0.893020	
8.000000	0.920304	0.918406	0.856481	0.837963	0.897674	0.893020	
9.000000	0.914611	0.882353	0.879630	0.824074	0.860465	0.823256	
10.000000	0.912713	0.880455	0.856481	0.800926	0.888372	0.851116	

**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 neighbors being weighted for deciding the final label, and this will often involve neighbors that are far apart, thus creating the soft boundaries.