Homework #1

Shivankur Kapoor

{kapoors@usc.edu}

9154524479

1. Density Estimation

A)

- ***Beta Distribution Parameter Estimation***

PDF = $\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$, where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(n)$ is the gamma function defined as $\Gamma(n) = (n-1)!$

To calculate the maximum likelihood, we will define the log likelihood function as follows

$$\begin{aligned} l(\alpha,\beta|x) &= \ln \mathcal{L}(\alpha,\beta|x) = \sum_{i=1}^{N} \ln\left(\mathcal{L}_i(\alpha,\beta|x_i)\right) \\ &= \sum_{i=1}^{N} \ln\left(f(x_i;\alpha,\beta)\right) \\ &= \sum_{i=1}^{N} \ln\left(\frac{x_i^{\alpha-1}(1-x_i)^{\beta-1}}{B(\alpha,\beta)}\right) \\ &= (\alpha-1)\sum_{i=1}^{N}\ln(x_i) + (\beta-1)\sum_{i=1}^{N}\ln(1-x_i) - NlnB(\alpha,\beta) \end{aligned}$$

Given $\beta = 1$

$l(\alpha,1|x) = (\alpha-1)\sum_{i=1}^{N}\ln(x_i) - NlnB(\alpha,1)$

Now $B(\alpha,1) = \frac{\Gamma(\alpha)\Gamma(1)}{\Gamma(\alpha+1)} = \frac{(\alpha-1)!}{\alpha!} = \frac{1}{\alpha}$

$$l(\alpha,1|x) = (\alpha-1)\sum_{i=1}^{N}\ln(x_i) + Nln\alpha$$

Taking the derivative of the above equation with respect to α and setting it to 0 to find the maxima–

$$\frac{\delta l(\alpha,1)}{\delta \alpha} = \sum_{i=1}^{N} lnx_i + \frac{N}{\alpha} = 0$$

$$\boldsymbol{\alpha = -\frac{N}{\sum_{i=1}^{N} lnx_i}}$$

- ***Gaussian Distribution Parameter Estimation***

$$\text{PDF} = \frac{1}{(\sqrt{2\pi})\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Given $\mu = \sigma = \theta$

$$f(x|\theta,\theta) = \frac{1}{(\sqrt{2\pi})\theta} e^{-\frac{(x-\theta)^2}{2\theta^2}}$$

To calculate the parameter θ, we define the log likelihood function as follows –

$$l(\theta,\theta|x) = \ln\mathcal{L}(\theta,\theta|x) = \sum_{i=1}^{N} ln\, f(x_i|\theta,\theta)$$

$$= \sum_{i=1}^{N} \left(-\ln\left(\sqrt{2\pi}\theta\right) - \frac{(x_i-\theta)^2}{2\theta^2}\right)$$

Taking the derivative of the above equation with respect to θ and setting it to zero to estimate θ –

$$\frac{\delta l(\theta,\theta|x)}{\delta\theta} = -\frac{N}{\theta} + \sum_{i=1}^{N} \frac{x_i(x_i - \theta)}{\theta^3} = 0$$

$$N\theta^2 + N\theta - \sum_{i=1}^{N} x_i^2 = 0$$

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

Since θ is also the variance of the distribution, hence cannot be negative

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

B) ***Kernel Density Estimation***

Given $X_1, X_2, X_3, X_4 \ldots\ldots X_n$ are i.i.d sampled according to density function $f(x)$

Kernel density estimator - $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{N} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$

To calculating bias –

$$E\left[\hat{f}(x)\right] - f(x) = ?$$

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

Because the random variables are independent -

$$= \frac{1}{n} \sum_{i=1}^{N} E\left[\frac{1}{h} K\left(\frac{x - X_i}{h}\right)\right]$$

Because the random variables are identically distributed –

$$= \frac{n}{n} E\left[\frac{1}{h} K\left(\frac{x - X_i}{h}\right)\right]$$

By definition of Expectation of continuous function of random variable –

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

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

Substituting $z = \frac{x-t}{h}$

$$zh = x - t$$

$$t = x - zh$$

$$dt = -h dz$$

Hence the equation becomes (inverting the limits to get rid of –ve sign)–

$$= \frac{h}{h} \int K(z) f(x - zh) dz$$

Using Taylor series to expand $f(x - zh)$ –

$$= \int K(z) dz \left[f(x) - zhf'(x) + \frac{z^2 h^2}{2} f''(x) - \frac{z^3 h^3}{3!} f'''(x) \ldots\right]$$

By definition of Kernel density estimator –

$$\int dz K(z) = 1$$

$$\int dz K(z) z = 0$$

$$\int dz K(z) z^2 = \sigma_k^2$$

Hence the equation becomes –

$$= \int K(z) dz f(x) - \int zK(z) dz h f'(x) + \int \frac{z^2 K(z) dz h^2 f''(x)}{2} \ldots\ldots$$

$$= f(x) - 0 + \frac{h^2 f''(x)}{2} \int z^2 K(z) dz + \frac{h^3 f'''(x)}{3!} \int dz K(z) z^3 \ldots.$$

So bias becomes –

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

where $O(h^2)$ represents the higher order terms in the equation (>2)

2. Naïve Bayes

A) Given $X = \{X_1, X_2, X_3 \ldots\ldots X_D\}$ is continuous random vector in $\mathbb{R}^D$

Y = {0,1} and follows a Bernoulli distribution with parameter $\pi = P(Y = 1)$

For each feature X_j, $P(X_j|Y = y_k)$follows Gaussian distribution

According to Bayes theorem –

$$P(Y = 1\,|X) = \frac{P(X|Y = 1)P(Y = 1)}{P(X)}$$

$$P(Y = 1\,|X) = \frac{P(X|Y = 1)P(Y=1)}{P(X\,|\,Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}$$

Using the naïve assumption on above equation for X

$$= \frac{\prod_{j=1}^{D} P(X_j|Y = 1)P(Y = 1)}{\prod_{j=1}^{D} P(X_j|Y = 1)P(Y = 1) + \prod_{j=1}^{D} P(X_j|Y = 0)P(Y = 0)}$$

$$= \frac{1}{1 + \frac{\prod_{j=1}^{D} P(X_j|Y = 0)P(Y = 0)}{\prod_{j=1}^{D} P(X_j|Y = 1)P(Y = 1)}}$$

$$= \frac{1}{1 + \frac{\prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_j-\mu_{j0})^2}{2\sigma_j^2}} (1-\pi)}{\prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_j-\mu_{j1})^2}{2\sigma_j^2}} (\pi)}}$$

$$= \frac{1}{1 + \frac{e^{\sum_{j=1}^{D} -\frac{(x_j-\mu_{j0})^2}{2\sigma_j^2}} (1-\pi)}{e^{\sum_{j=1}^{D} -\frac{(x_j-\mu_{j1})^2}{2\sigma_j^2}} (\pi)}}$$

$$= \frac{1}{1 + e^{\sum_{j=1}^{D} \frac{(x_j-\mu_{j1})^2-(x_j-\mu_{j0})^2}{2\sigma_j^2}} e^{\ln\left(\frac{1-\pi}{\pi}\right)}}$$

$$= \frac{1}{1 + e^{\sum_{j=1}^{D} \frac{(x_j - \mu_{j1})^2 - (x_j - \mu_{j0})^2}{2\sigma_j^2} + \ln\left(\frac{1-\pi}{\pi}\right)}}$$

$$= \frac{1}{1 + e^{\sum_{j=1}^{D} \frac{\mu_{j0} - \mu_{j1}}{\sigma_j^2} x_j + \sum_{j=1}^{D} \frac{\mu_{j1}^2 - \mu_{j0}^2}{2\sigma_j^2} + \ln\left(\frac{1-\pi}{\pi}\right)}}$$

Comparing the above equation with –

$$= \frac{1}{1 + e^{(-w_0 + W^T X)}}$$

$$\boldsymbol{w_0 = -\left(\sum_{j=1}^{D} \frac{\mu_{j1}^2 - \mu_{j0}^2}{2\sigma_j^2} + \ln\left(\frac{1-\pi}{\pi}\right)\right)}$$

$$\boldsymbol{W^T X = \sum_{j=1}^{D} \frac{\mu_{j0} - \mu_{j1}}{\sigma_j^2} x_j}$$

B) Given N training examples – $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots . (x_N, y_N)$

where $x_i = (x_{i1}, x_{i2}, x_{i3} \ldots x_{iD})^T$ and $y_i = \{k\}, k = \{0,1\}$

According to Naïve Bayes classification -

$$P(Y = y_k \mid X = x_i) = \prod_{j=1}^{D} P(X = x_{ij} | Y = y_k) P(Y = y_k)$$

Likelihood function for the Naïve Bayes can be defined as follows –

$$L(\pi, \mu, \sigma) = \prod_{k=0}^{K} \prod_{i=1}^{N_k} \prod_{j=1}^{D} P(X = x_{ij} | Y = y_k) P(Y = y_k)$$

Since we are given that X follows a Gaussian distribution,

$$P(X = x_{ij} | Y = y_k) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_{ij} - \mu_{jk})^2}{2\sigma_j^2}}$$

and let's represent prior probability $P(Y = y_k) = \pi_k$

$$L(\pi, \mu, \sigma) = \prod_{k=0}^{K} \prod_{i=1}^{N_k} \prod_{j=1}^{D} \left(\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x_{ij} - \mu_{jk})^2}{2\sigma_j^2}} \right) \pi_k$$

Taking the log of the above equation, we find the log likelihood function as follows –

$$\ell(\pi, \mu, \sigma) = \sum_{k=0}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{D} \left(\log \frac{1}{\sqrt{2\pi}\sigma_j} - \frac{(x_{ij} - \mu_{jk})^2}{2\sigma_j^2} \right) + \log \pi_k$$

As it's given that K = {0,1} where $\pi_1 = \pi \ and \ \pi_0 = 1 - \pi$

$$l(\pi, \mu, \sigma) = \sum_{i=1}^{N_1} \sum_{j=1}^{D} (\log \frac{1}{\sqrt{2\pi}\sigma_j} - \frac{(x_{ij} - \mu_{j1})^2}{2\sigma_j^2}) + \log \pi$$
$$+ \sum_{i=1}^{N_0} \sum_{j=1}^{D} (\log \frac{1}{\sqrt{2\pi}\sigma_j} - \frac{(x_{ij} - \mu_{j0})^2}{2\sigma_j^2}) + \log 1 - \pi$$

where N_1 = total examples where Y = 1 and N_0 = total examples where Y = 0

Hence $N_1 + N_0$= N

Taking partial derivatives with respect to parameters and equating to zero –

$$1. \ \frac{\partial l(\pi,\mu,\sigma)}{\partial \pi} = \sum_{i=1}^{N_1} \frac{1}{\pi} - \sum_{i=1}^{N_0} \frac{1}{1-\pi} = 0$$

$$\frac{N_1}{\pi} - \frac{N_0}{1-\pi} = 0$$

$$\frac{(1-\pi)N_1 - \pi N_0}{\pi(1-\pi)} = 0$$

$$\pi(N_1 + N_0) = N_1$$

$$\pi = \frac{N_1}{(N_1 + N_0)} = \frac{Number\ of\ examples\ where\ Y = 1}{Total\ number\ of\ examples}$$

$$2. \ \frac{\partial l(\pi,\mu,\sigma)}{\partial \mu_{j1}} = \sum_{i=1}^{N_1} \frac{2(x_{ij} - \mu_{j1})}{2\sigma_j^2} = 0$$

$$\sum_{i=1}^{N_1} (x_{ij} - \mu_{j1}) = 0$$

$$\sum_{i=1}^{N_1} x_{ij} - \sum_{i=1}^{N_1} \mu_{j1} = 0$$

$$\sum_{i=1}^{N_1} x_{ij} - N_1 \mu_{j1} = 0$$

$$\mu_{j1} = \frac{\sum_{i=1}^{N_1} x_{ij}}{N_1}$$

Here $\sum_{i=1}^{N_1} x_{ij}$ represents the sum of $x_{ij}'s$ where Y = 1

Similarly we can calculate $\boldsymbol{\mu_{j0}}$

$$\mu_{j0} = \frac{\sum_{i=1}^{N_0} x_{ij}}{N_0}$$

Here $\sum_{i=1}^{N_0} x_{ij}$ represents the sum of $x_{ij}'s$ where Y = 0

So, μ_{jk} can be represented as-

$$\mu_{jk} = \frac{\sum_{i=1}^{N_k} x_{ij}}{N_k} = \frac{Sum\ of\ x_{ij}'s\ where\ Y = k}{Total\ number\ of\ examples\ where\ Y = k}$$

3 . $\frac{\partial l(\pi,\mu,\sigma)}{\partial \sigma_j} = \sum_{i=1}^{N_1} -\frac{1}{\sigma_j} + \frac{(x_{ij}-\mu_{j1})^2}{\sigma_j^3} + \sum_{i=1}^{N_0} -\frac{1}{\sigma_j} + \frac{(x_{ij}-\mu_{j0})^2}{\sigma_j^3} = 0$

$$-\frac{N_1}{\sigma_j} + \sum_{i=1}^{N1} \frac{(x_{ij}-\mu_{j1})^2}{\sigma_j^3} - \frac{N_0}{\sigma_j} + \sum_{i=1}^{N_0} \frac{(x_{ij}-\mu_{j0})^2}{\sigma_j^3} = 0$$

$$-N_1 + \sum_{i=1}^{N1} \frac{(x_{ij}-\mu_{j1})^2}{\sigma_j^2} - N_0 + \sum_{i=1}^{N_0} \frac{(x_{ij}-\mu_{j0})^2}{\sigma_j^2} = 0$$

$$-N_1 + \sum_{i=1}^{N1} \frac{(x_{ij}-\mu_{j1})^2}{\sigma_j^2} - N_0 + \sum_{i=1}^{N_0} \frac{(x_{ij}-\mu_{j0})^2}{\sigma_j^2} = 0$$

$$\sigma_j^2 = \frac{\left(\sum_{i=1}^{N1}(x_{ij}-\mu_{j1})^2 + \sum_{i=1}^{N_0}(x_{ij}-\mu_{j0})^2\right)}{N_0+N_1}$$

$$\sigma_j = \sqrt{\frac{\left(\sum_{i=1}^{N1}(x_{ij}-\mu_{j1})^2 + \sum_{i=1}^{N_0}(x_{ij}-\mu_{j0})^2\right)}{N_0+N_1}}$$

$$\boldsymbol{\sigma_j = \sqrt{\frac{\sum_{k=0}^{k=1}\sum_{i=1}^{Nk}(x_{ij}-\mu_{jk})^2}{N}}}$$

Here k represents the classes {0,1}

Here $\sum_{i=1}^{N_k}(x_{ij}-\mu_{jk})^2$ represents the sum of $(x_{ij}-\mu_{jk})^2$ where Y = k

3. Nearest Neighbor

A) Data Points –

	x	y	$\frac{x-\mu_x}{\sigma_x}$	$\frac{(y-\mu_y)}{\sigma_y}$	Major
1	0	49	-0.616	1.415	M
2	-7	32	-0.954	0.759	M
3	-9	47	-1.050	1.337	M
4	29	12	0.783	-0.011	EE
5	49	31	1.748	0.720	EE
6	37	38	1.169	0.990	EE
7	8	9	-0.230	-0.127	CS
8	13	-1	0.011	-0.513	CS
9	-6	-3	-0.905	-0.590	CS
10	-21	12	-1.630	-0.011	CS
11	27	-32	0.686	-1.708	E
12	19	-14	0.300	-1.014	E
13	27	-20	0.686	-1.245	E

M – Mathematics CS – Computer Science

EE – Electrical Engineering E - Economics

$$\mu_x = 12.76 \qquad \mu_y = 12.30$$

$$\sigma_x = 20.71 \qquad \sigma_y = 25.93$$

Given test data point – (20,7)

Standardizing the test data point - (0.349, -0.204)

- **Using L2 norm –**

  For K = 1

  Calculating the L2 norm for test data point from every training data point, we get ***Minimum L2 distance is from data point 8 - (13, -1), distance = 0.457. So***

class is CS

For K = 5

Calculating the L2 norm for test data point from every training data point, we get following 5 nearest neighbors -

8	**13**	**-1**	**CS**	**0.457**
4	29	12	EE	0.475
7	**8**	**9**	**CS**	**0.584**
12	**19**	**-14**	**E**	**0.811**
13	**27**	**-20**	**E**	**1.109**

Since there is tie between CS an E, but CS is closer to the test data point. So class is CS

- **Using L1 Norm**

For K = 1

Calculating L1 distance from test data point to every training data point, we get

Minimum L2 distance is from data point 4 - (29, 12), distance = 0.627. So class is EE

For K = 5

Calculating the L2 norm for test data point from every training data point, we get following 5 nearest neighbors -

4	29	12	EE	0.627
8	**13**	**-1**	**CS**	**0.646**
7	**8**	**9**	**CS**	**0.656**
12	**19**	**-14**	**E**	**0.858**
13	**27**	**-20**	**E**	**1.379**

Since there is tie between CS an E, but CS is closer to the test data point. So class is CS

- ***Comparison -***

K	L1	L2
1	**EE**	**CS**
5	**CS**	**CS**

The result for K = 1 is different for L1 and L2 distance metrics. L1 norm, which is the Manhattan distance, is smallest for an example in EE (data point 4). The rest of the order of 5 nearest neighbors is same for both L1 and L2

B) Following information is given -

$$\sum_c N_c = N$$

K_c data points are inside the sphere, where K are total labeled data points

$$\sum_c K_c = K$$

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

- To find – $P(x)$

$$P(x) = \sum_c P(x|Y = c)P(Y = c)$$

$$= \sum_c \frac{K_c N_c}{N_c VN}$$

$$= \sum_c \frac{K_c}{VN}$$

$$= \boldsymbol{\frac{K}{VN}}$$

- To find – $P(Y = c|x)$

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

$$= \frac{K_c N_c VN}{N_c VNK}$$

$$= \boldsymbol{\frac{K_c}{K}}$$

4. Decision Tree

A) Information Gain(Weather) = $I(\frac{73}{100}, \frac{27}{100}) - [\frac{28}{100} I\left(\frac{23}{28}, \frac{5}{28}\right) + \frac{72}{100} I(\frac{50}{72}, \frac{22}{72})]$

$$= \left(-\frac{73}{100}\log\left(\frac{73}{100}\right) - \frac{27}{100}\log(\frac{27}{100})\right) - [0.28(-\frac{23}{28}\log\left(\frac{23}{28}\right) - \frac{5}{28}\log(\frac{5}{28})) + 0.73(-\frac{50}{72}\log\left(\frac{50}{72}\right) - \frac{22}{72}\log\left(\frac{22}{722}\right))]$$

$$= 0.0125$$

Information Gain(Traffic) = $I(\frac{73}{100}, \frac{27}{100}) - [\frac{73}{100} I\left(\frac{73}{73}, \frac{0}{73}\right) + \frac{27}{100} I(\frac{0}{27}, \frac{27}{27})]$

$$= \left(-\frac{73}{100}\log\left(\frac{73}{100}\right) - \frac{27}{100}\log(\frac{27}{100})\right) - [0.73(-\frac{73}{73}\log\left(\frac{73}{73}\right) - \frac{0}{73}\log(\frac{0}{73})) + \frac{27}{100}(-\frac{0}{27}\log\left(\frac{0}{27}\right) - \frac{27}{27}\log\left(\frac{27}{27}\right))]$$

$$= 0.8414$$

Since the IG(Traffic) > IG(Weather), so we choose traffic to split the data in the first step.

B) **T1 and T2 will be same** since learning algorithm for decision tree is not affected by the scaling or normalization. A node of a tree partitions the data into 2 sets by comparing a feature (which splits data best). This decision is not affected if the data has been scaled, the partitioning will remain same.

C) Given -

Gini Index = $\sum_{k=1}^{K} p_k(1 - p_k)$

Cross Entropy = $-\sum_{k=1}^{K} p_k log p_k$

To prove –

Gini Index <= Cross Entropy

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

$$\sum_{k=1}^{K} p_k(1 - p_k) + \sum_{k=1}^{K} p_k log p_k \leq 0$$

$$\sum_{k=1}^{K} p_k(1 - p_k) + p_k log p_k \leq 0$$

$$\sum_{k=1}^{K} p_k(1 - p_k + log p_k) \leq 0 \qquad Eq.\ (1)$$

Since p_k is the probability density, therefore –

$$p_k \geq 0$$

So if we can prove that $(1 - p_k + log p_k) \leq 0, we$ can prove Eq. (1)

To prove the above inequality, we take the first and second derivatives of the above equation

Let $f(p_k) = 1 - p_k + logp_k$

$$f'(p_k) = -1 + \frac{1}{p_k}$$

$$f''(p_k) = -\frac{1}{p_k^2}$$

Since the second derivate is a negative quantity, so this means that solving the first derivate for 0 will give us the maxima for $f(p_k)$

From the first derivative, we can compute the maxima/minima –

$$f'(p_k) = -1 + \frac{1}{p_k} = 0$$

$$p_k = 1$$

Hence the maxima for $f(p_k) = 0$

This implies that $f(p_k) \leq 0$

From the above conclusion we can infer that Eq. (1) will always be <= 0

Therefore -

$$\underbrace{\sum_{k=1}^{K} p_k(1 - p_k + logp_k)}_{<=0} \leq 0$$

>=0

5. Programming Assignment

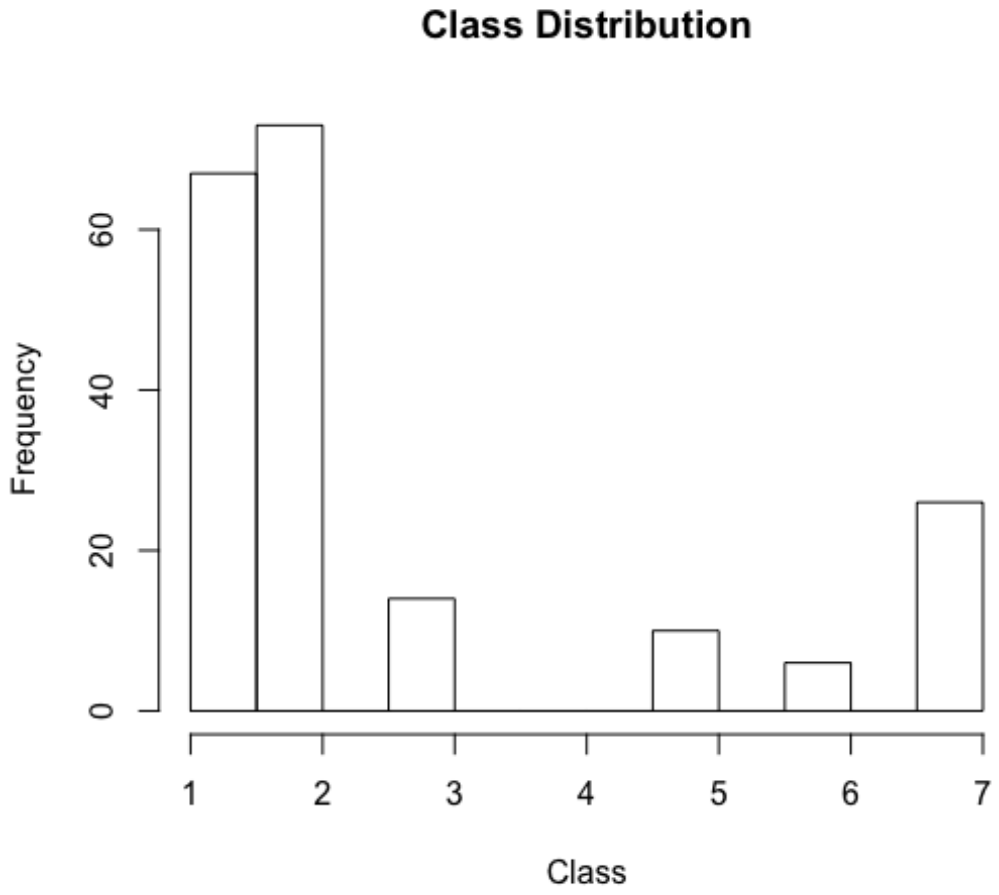
A.) Data Inspection

- Number of attributes – 10
- No, all the attributes are not important for classification. 1st attribute (id) is not required for classification as it holds no information, it is the count of the instances.
- There are a total of 6 classes that appears in the training set – 1,2,3,5,6,7
- Following table shows the class distribution –

Class	Frequency
1	67

2	73
3	14
5	10
6	6
7	26

So class 2 is the majority class.



No, the class distribution can't be considered a uniform distribution as evident from the above histogram plot.

B) Performance Comparison –

- **Naïve Bayes Classifier**

Test Accuracy	**33.33%**
Training Accuracy	**55.10%**

- **KNN Classifier**

Results for test data

K	L1	L2
1	66.67%	61.11%
3	61.11%	61.11%
5	50.56%	55.56%
7	50.00%	55.56%

Results for training data

K	L1	L2
1	75.00%	71.43%
3	73.47%	71.94%
5	67.86%	67.86%
7	68.88%	66.84%

- **Discussion** – Though neither of the classifiers performed very well, it is evident from the above results that KNN classifier performed better than Naïve Bayes classifier on both training and test tests, and by a large margin. The reason of the poor performance of Naïve Bayes is our assumption that the distribution of the features come from a Gaussian curve. This is not true for all the features, hence Naïve Bayes classifier computes incorrect probabilities for the test examples. Also, in Naïve Bayes we assume that all features are conditionally independent, which might not be true for all the features. On the other hand, KNN does not rely on the probability distribution but just calculates the distance of test examples from training examples scattered in the feature space. Hence it performs better compared to Naïve Bayes.