# Machine Learning Assignment 1

Anirudh Mukund Deshpande
USC ID: 2746365653
deshpana@usc.edu

September 2016

## 1 Density Estimation

(1.a.1) **Answer:**

(a) The formula for $\beta$ - distribution is given by -

$f(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, B(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}$

where $\theta$ takes the value in the range [0,1]. Lets call it *equation 1*
Since it is given that $\beta$=1, the constant $B(\alpha,\beta)$ which is dependent on $\alpha$ and $\beta$ could be transformed as follows -

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}dt = [\frac{t^\alpha}{\alpha}]_0^1 = \frac{1}{\alpha}$$

Substituting value of B$(\alpha,\beta)$ as $\frac{1}{\alpha}$ and $\beta$=1 in *equation 1* have,

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

Hence the equation of log-likelihood becomes -

$$l(\alpha,1) = \sum_{i=1}^{N} log(\alpha X_i^{\alpha-1}) = \sum_{i=1}^{N}(log\alpha + (\alpha-1)logX_i)$$

Lets call it *equation 2*
Differentiating this *equation 2* w. r. to $\alpha$ we have -

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{N}(\frac{1}{\alpha} + logX_i)$$

Equating it to zero, we have -

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{N}(\frac{1}{\alpha} + logX_i) = 0$$

Lets call this *equation 3* When *equation 3* is simplified, we get the following form -

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{N}(\frac{1}{\alpha} + logX_i) = 0\frac{N}{\alpha} + \sum_{i=1}^{N} logX_i = 0$$

Hence we have -

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

(1.a.2) **Answer:** We have formula of Gaussian distribution as -

$$\eta(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

By substituting $\mu$ and $\sigma$ with $\Theta$, we have -

$$\eta(\Theta, \Theta) = \frac{1}{\Theta\sqrt{2\pi}} e^{-(x-\Theta)^2/2\Theta}$$

The likelihood function for $\eta(\Theta, \Theta)$ can be defined as -

$$\begin{aligned} l(\theta|x) &= \sum_{i=1}^{N} \eta(\Theta, \Theta) \\ &= \sum_{i=1}^{N}[\frac{-1}{2}(2\pi\theta) - \frac{1}{2}\frac{(x-\theta)^2}{\theta}] \sum_{i=1}^{N}[\frac{-1}{2}(2\pi\theta) - \frac{1}{2}\frac{(x-\theta)^2}{\theta}] \end{aligned}$$

We have -

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= \sum_{i=1}^{N}[\frac{1}{2}\frac{1}{2\pi\theta}2\pi - \frac{1}{2}(-1)\theta^{-2}(x_i-\theta)^2 + \theta^{-1}2(x_i-\theta)(-1)] \\ &= \sum_{i=1}^{N}[\frac{-1}{2\theta} + \frac{(x-\theta)^2}{2\theta^2} + \frac{(x_i-\theta)}{\theta}] \\ &= \sum_{i=1}^{N}[\frac{-1}{2\theta} + \frac{x_i^2-\theta^2}{2\theta^2}] \\ &= \frac{1}{2\theta}\sum_{i=1}^{N}[\frac{x_i^2-\theta^2-\theta}{\theta}] \\ &= \frac{1}{2\theta^2}[\sum_{i=1}^{N} x_i^2 - \theta^2 N - \theta N] \end{aligned}$$

Equating the above equation to zero, we have -

$$\frac{1}{2\theta^2}[\sum_{i=1}^{N} x_i^2 - \theta^2 N - \theta N] = 0 => N\theta^2 + n\theta - \sum_{i=1}^{N} x_i^2 = 0$$

By solving the above equation, we have -

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

By simplifying the above equation, we have -

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

(1.b) According to the question, we have $\hat{f}(x)$ is -

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

This implies the expectation of variables $X_1, X_2, ..., X_N$ as -

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

Lets call the above *Equation 1* We already know that $X_1, X_2...X_n$ are the i. i. d.s which implies that they have identical probability density function. Hence we can write the above equation as -

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

Lets call the above *Equation 2* Where we know that -

$$E[\frac{1}{h}K(\frac{x-X_i}{h})] = \int_{-\infty}^{\infty} \frac{1}{h}K(\frac{x-u}{h})f(u)du$$

where u cumulatively represents all the random variables $X_1, ...X_n$. Lets call the above *Equation 3*

By substituting *Equation 2* and *Equation 3* in *Equation 1* we have -

$$\begin{aligned} E_{X_1,X_2...X_n}[\hat{f}(x)] &= \int_{-\infty}^{\infty} \frac{1}{h}K(\frac{x-u}{h})f(u)du \\ &= \frac{1}{h}\int_{-\infty}^{\infty} K(\frac{x-u}{h})f(u)du \end{aligned}$$

Now, let us substitute $\frac{x-u}{h} = z$ and take derivative with respect to z

$$\frac{dx}{dz} = \frac{d}{dz}(u + hz)$$

Since we know that $\frac{dx}{dz} = 0$, we get the following equation.

$$\begin{aligned} &= \frac{du}{dz} + h = 0 \\ &= du = -hdz \end{aligned}$$

By the equation $\frac{x-u}{h} = z$, we can identify that, dz tends to $-\infty$ when du tends to $\infty$ and dz tends to $\infty$ when du tends to $infty$,
This implies -

$$E[\hat{f}(x)] = \frac{1}{h}\int_{\infty}^{-\infty} K(z)f(x-zh)(-h)dz \quad = \frac{1}{h}(h)\int_{\infty}^{-\infty} -K(z)f(x-zh)dz = \int_{-\infty}^{\infty} K(z)f(x-zh)dz$$

We have Taylor series as,

$$f(x-hz) = f(x) + f'(x)(-hz) + \frac{f''(x)h^2z^2}{2} + \frac{f'''(x)(-hz)^3}{3!} + .... + \frac{f'''^n(x)(-hz)^3}{n!} + ...$$

Neglecting the higher order values, we can write the Taylors series as -

$$f(x-hz) = f(x) + f'(x)(-hz) + \frac{f''(x)h^2z^2}{2} + O(h^3)$$

Hence we have,

$$\int_{-\infty}^{\infty} K(z)f(x-zh)dz = \int_{-\infty}^{\infty} K(z)f(x)dz + \int_{-\infty}^{\infty} K(z)f'(x)(-hz)dz + \int_{-\infty}^{\infty} \frac{K(z)f''(x)(hz)^2}{2}dz + O(h^3)$$

Let us call the above equation as *Equation 4*

We already know that,

i. $\int_{-\infty}^{\infty} K(z)d(z) = 1$
ii. $\int_{-\infty}^{\infty} zK(z)dz = 0$

Hence *Equation 4* can be written as -

$$f(x) + \frac{f''(x)h^2}{2}\int_{-\infty}^{\infty} K(z)z^2du + O(h^3)$$

Hence the answer is -

$$\implies E[\hat{f}(x)] - f(x) = \frac{f''(x)h^2}{2}\int_{-\infty}^{\infty} K(z)z^2du + O(h^3)$$

## 2 Naive Bayes

(2.a) Using the Bayes rule and by replacing P(Y=1) = $\pi$ and P(Y=0) = (1- $\pi$),

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

Dividing the numerator and denominator by $\pi$ This can be written as follows -

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

Now, expanding the Random variable X = $X_1, X_2, ...X_N$, we know from the Naive Bayes assumption that,

(a) P(X—Y=1) = $\prod_{i=1}^{D}$

(b) P(X—Y=0) = $\prod_{i=0}^{D}$

Hence the equation can be transformed as -

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

Dividing the numerator and denominator by $\prod_{i=1}^{D} P(X_i|Y=1)$, we have -

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

From Gaussian Distribution, $\mathcal{N}(\mu_{jk}, \sigma_j) = \eta(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$

$$P(Y=1|X) = \frac{1}{1 + \frac{\prod_{i=1}^{D} \mathcal{N}(\mu_{i0}, \sigma_i)}{\prod_{i=1}^{D} \mathcal{N}(\mu_{i1}, \sigma_i)}\frac{(1-\pi)}{\pi}}$$

$$P(Y=1|X) = \frac{1}{1 + \frac{(1-\pi)}{\pi} exp[\frac{-1}{2}\sum_{i=0}^{D} \frac{(x_i-\mu_{i0})^2-(x_i-\mu_{i1})^2}{\sigma_i^2}]}$$

$$P(Y=1|X) = \frac{1}{1 + exp[\frac{-1}{2}\sum_{i=0}^{D} \frac{(x_i-\mu_{i0})^2-(x_i-\mu_{i1})^2}{\sigma_i^2} + log(\frac{(1-\pi)}{\pi})]}$$

Further upon the simplification,

$$P(Y=1|X) = \frac{1}{1 + exp[ln(\frac{(1-\pi)}{\pi}) + \frac{-1}{2}\sum_{i=0}^{D} \frac{\mu_{i0}^2-\mu_{i1}^2}{\sigma_i^2} - \sum_{i=0}^{D} \frac{x_i(\mu_{i1}-\mu_{i0})}{\sigma_i^2}]}$$

$$P(Y=1|X) = \frac{1}{1 + exp[-(\frac{1}{2}\sum_{i=0}^{D} \frac{\mu_{i0}^2-\mu_{i1}^2}{\sigma_i^2} + ln(\frac{\pi}{(1-\pi)})) - \sum_{i=0}^{D} \frac{(\mu_{i1}-\mu_{i0})x_i}{\sigma_i^2}]}$$

Where we have -

$$w_0 = \frac{1}{2}\sum_{i=0}^{D} \frac{\mu_{i0}^2-\mu_{i1}^2}{\sigma_i^2} + ln(\frac{\pi}{(1-\pi)}), \qquad w^T X = -\sum_{i=0}^{D} \frac{(\mu_{i1}-\mu_{i0})x_i}{\sigma_i^2}$$

(2.b) As per the definition of Gaussian distribution, we have

$$P(X_d|Y=y_k) = \frac{1}{\sigma_d\sqrt{2\pi}}e^{-\left(x_d-\mu_{dy_k}\right)^2/2\sigma^2}$$

As the random variables are conditionally independent,

$$P(X_d, Y=y_k) = \pi_{yk}\Pi_{d=1}^{D}\frac{1}{\sigma_d\sqrt{2\pi}}e^{-\left(x_d-\mu_{dy_k}\right)^2/2\sigma^2}$$

We have the log likelihood function as -

$$\begin{aligned}
L &= \Sigma_{i=1}^{N} log(P(X_i, Y_i)) \\
&= \Sigma_{i=1}^{N} [log\pi_k + \Sigma_{i=1}^{N} log(\pi_{yk} \Pi_{d=1}^{D} \frac{1}{\sigma_d \sqrt{2\pi}} e^{-\left(x_d - \mu_{dy_k}\right)^2 / 2\sigma^2})] \\
&= \Sigma_c \Sigma_{i:y=c} log\pi_{y_i} + \Sigma_{i=1}^{N} [log\pi_{yi} + \Sigma_c \Sigma_{i:y=c} \Sigma_{d=1}^{D} log(\sqrt{2\pi}\sigma_d) - \Sigma_c \Sigma_{i:yi=c} \Sigma_{d=1}^{D} (\frac{(x_d - \mu_{dc})^2}{2\sigma_d^2})] \\
&= \Sigma_c C_c log\pi_c + \Sigma_c \Sigma_{i:y=c} log\pi_{y_i} - \Sigma_c \Sigma_{i:y=c} \Sigma_{d=1}^{D} (\frac{(x_d - \mu dc)^2}{2\sigma_d^2})
\end{aligned}$$

We also have the constraint as below -

$$\Sigma_c \pi_c = 1$$

To satisfy the above constraint, we introduce the Lagrange multiplier $\lambda$ as below -

$$L = \Sigma_c C_c log\pi_c + \Sigma_c \Sigma_{i:y=c} log\pi_{y_i} - \Sigma_c \Sigma_{i:y=c} \Sigma_{d=1}^{D} (\frac{(x_d - \mu dc)^2}{2\sigma_d^2}) + \lambda(\Sigma_c \pi_c - 1)$$

To estimate $\pi_c$ we will need to equate the first order partial derivative to 0. Hence we have -

$$\begin{aligned}
\frac{\partial L}{\partial \pi_c} &= 0 \\
\frac{C_c}{\pi_c} + \lambda &= 0
\end{aligned}$$

Hence we have -

$$\pi_c = \frac{-C_c}{\lambda}$$

Now we can estimate $\lambda$ as

$$\begin{aligned}
\Sigma_c \pi_c &= \Sigma_c \frac{-C_c}{\lambda} \\
1 &= \Sigma_c \frac{-C_c}{\lambda} \\
\lambda &= \Sigma_c - C_c
\end{aligned}$$

Substituting $\lambda$ in equation 5 we can get an estimate for $\pi_c$

$$\pi_c = \frac{-C_c}{\Sigma_c - C_c} = \frac{C_c}{N}$$

To estimate $\mu_{dc}$ we will again equate the partial derivative to 0.

$$\begin{aligned}
\frac{\partial L}{\mu_{dc}} &= 0 \\
\Sigma_{i:y_i=c} \frac{-2(x_d - \mu_{dc})(-1)}{2\sigma_d^2} &= 0 \\
\Sigma_{i:y_i=c} x_d - \Sigma_{i:y_i=c}\ \mu_{dc} &= 0 \\
\Sigma_{i:y_i=c} x_d - C_c \mu_{dc} &= 0
\end{aligned}$$

Hence we have -

$$\mu_{dc} = \frac{\Sigma_c x_d}{C_c}$$

Where $C_c$ is the number of samples in C.

To estimate $\sigma_d$ we will begin again by equating partial derivative to 0.

$$\begin{aligned}\frac{\partial L}{\partial \sigma_d} &= 0\\ \Rightarrow 0 - \Sigma_c\Sigma_{i:y_i=c}\frac{1}{\sigma_d} + \Sigma_c\Sigma_{i:y_i=c}\frac{(x_d-\mu_{dc})^2}{\sigma^3} &= 0\\ \Rightarrow \Sigma_c\Sigma_{i:y_i=c}\frac{(x_d-\mu_{dc})^2}{\sigma_d^2} - \Sigma_c\Sigma_{i:y_i=c}1 &= 0\\ \Sigma_c\Sigma_{i:y_i=c}\frac{(x_d-\mu_{dc})^2}{\sigma_d^2} - N &= 0\end{aligned}$$

Hence we have,

$$\sigma_d = \sqrt{\Sigma_c\Sigma_{i:y_i=c}\frac{(x_d-\mu_{dc})^2}{N}}$$

Where N represents total number of samples.

# 3 Nearest Neighbour

(3.a) To normalize the data, we calculate mean and variance of 'x' and 'y' coordinates, using the formulae -

$$\mu_k = \frac{\sum_i k_i}{N}, \quad \sigma_k = \sqrt{\frac{\sum_{i=1}^{N}(k_i-\mu_k)^2}{N-1}}$$

We have mean and variance of 'x' and 'y' coordinates as -

$$\begin{aligned}\mu_x &= 12.7692, \quad \sigma_x = 20.7169\\ \mu_y &= 12.3077, \quad \sigma_y = 25.9306\end{aligned}$$

Formula for normalization of the 'x' and 'y' coordinates is as shown below -

$$\begin{aligned}x &= (\frac{x-\mu_x}{\sigma_x}),\\ y &= (\frac{y-\mu_y}{\sigma_y})\end{aligned}$$

Also, by normalizing the test data (20, 7) using $\mu_x$ and $\sigma_x$, we have normalized values as

$$\begin{aligned}x = 20 \quad normalized_x &= 0.3490\\ y = 7 \quad normalized_y &= -0.2046\end{aligned}$$

We normalize the training points and can compute the L1 and L2 distances of the test point with all the training points. After the computation, the the points and the corresponding distances obtained are shown as below.

**Mathematics**
$Distances : (2.5851, 1.8856), (2.2674, 1.6211), (2.9424, 2.0830)$
$Points : (-0.6164, 1.4150), (0.95430.7594), (-1.0508, 1.3378)$

**Electrical Engineering**
$Distances : (0.6272, 0.4752), (2.3253, 1.6781), (2.0161, 1.4500)$
$Points : (0.7834, -0.0118), (1.7488, 0.7208), (1.1696, 0.9908)$

**Computer Science**
$Distances : (0.6563, 0.5843), (0.6464, 0.4575), (1.6406, 1.3129), (2.1718, 1.9884)$
$Points : (-0.2302, -0.1275), (0.01113, -0.5132), (-0.9059, -0.5903), (-1.6301, -0.01186)$

**Economics**
$Distances : (1.8419, 1.5415), (0.8581, 0.8113), (1.3791, 1.0947)$
$Points : (0.6869, -1.7087), (0.30076, -1.0145), (0.6869, -1.2459)$

We can observe that for K=1, and for L1, nearest point is (0.7834, -0.0118).

When we have K=5, and for L1, The 5 nearest points are -

(0.7834, -0.0118) : 'Electrical Engineering',
(0.01113, -0.5132) : 'Computer Science',
(-0.2302, -0.1275) : 'Computer Science',
(0.30076,-1.0145) : 'Economics',
(0.6869,-1.2459) : 'Economics'

We can find that 2 classes 'Computer Science' and 'Economics' occur in the closer vicinity. The 'Computer Science' label is assigned since it is the closer one.'

For k = 1, 5 the classification results are as below -

**Results**

| K | Distance | Predicted major |
|---|---|---|
| 1 | Euclidean(L2)<br>Manhattan(L1) | Computer Science<br>Electrical Engineering |
| 5 | Euclidean(L2)<br>Manhattan(L1) | Computer Science<br>Computer Science |

(3.b) As per the question, we know that

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

ii. $p(Y = c) = \frac{N_c}{N}$

We know the formula for Marginal probability as below (Summation is over all the values of 'c')

$$p(x) = \sigma p(x|Y = c)p(y = c)$$

Substituting the above values in the equation, we have (Again, Summation is over all the values of 'c')-

$$p(x) = \sigma \frac{K_c}{N_c V} \frac{N_c}{N}$$
$$= \frac{\sigma K_c}{VN}$$

Since we know that $\sigma K_c = K$, we have -

$$p(x) = \frac{K}{NV}$$

We have the definition of Bayes theorem as below -

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

Using the Bayes theorem with the above found values we have -

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

Upon simplification, we get,

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

# 4 Decision Tree

(4.a) Let us calculate the entropy with respect to each of the features (Calling Accident rate as 'R', Traffic as 'T' and weather as 'W', Heavy as 'H', Low as 'L', Sunny as 'S' and Rainy as 'R')-

$$Entropy(R,T) = P(H) * E(73,0) + P(L) * E(0,27)$$

As we know that E(73,0) = 0 and E(0, 27) = 0, we have -

$$Entropy(R,T) = P(H) * 0 + P(Low) * 0$$

Also we have,

$$Entropy(R,W) = P(S) * E(23,5) + P(R) * E(50,22)$$

We can say that -

$$Entropy(R,W) > 0$$

From the comparison, It is implied that the Information gain is more for Traffic. Hence **Traffic** could be chosen as a feature for first split.

(4.b) The second student is eventually normalizing the data. T1 and T2 will remain same. The information gain of features in the case of student 1 and student 2 will remain the same as normalization is a linear operation on the data.

(4.c) The inequality we need to prove is -

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

Without the loss of generality, we can remove the summation and write the above equation as below. the summation can be removed because $p_k$ represent probability mass function on both the sides of the equation.

$$(1-p_k) <= -log p_k$$

Let us have a function, $f(p_k) = 1 - p_k + log p_k$ over [0,1], We need to prove that f(x) ¡ 0
By differentiating $f(p_k)$ with respect to $f(p_k)$, The first derivative is -

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

Taking the second derivative, we have -

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

The negative value of the second derivative indicates that the function is increasing. By substituting the values in the range $0 <= p_k <= 1$, It can be observed that f($p_k$) in the interval [0,1] is monotonically increasing and attains maximum value at $p_k = 1$. where the value of f($p_k$) becomes 0. Hence we have -

$$f(p_k) <= 0$$

As the maximum value of f($p_k$) is 0. This implies we have -

$$= 1 - p_k + log p_k <= 0$$
$$\Rightarrow 1 - p_k <= -log p_k$$

Hence we have,

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

# 5 Programming

## 5.1 Data Inspection

(a) We have a total of 11 columns in the data. As the 1st column indicates the *ID Number* and last column indicated *Type of glass*, we have a total of 9 features.

(b) For some of the classes, some attributes are not meaningful as they have zero variance. When the variance is zero, the attributes do not contribute anything prominent towards the feature set of the class which would help in classification.
Some of the examples are -

i. vehicle windows float processed has ’Ba’ variance zero.
ii. containers has ’Fe’ variance zero.
iii. tableware has ’Ba’ variance zero.
iv. tableware has ’Fe’ variance zero.
v. tableware has ’K’ variance zero.

(c) We have a total of 6 classes in the data set provided to us.

i. vehicle windows float processed
ii. vehicle windows non float processed
iii. containers
iv. tableware
v. building windows float processed
vi. building windows non float processed

(d) We observe that class 2 ($building_{w}indows_{n}on_{f}loat_{p}rocessed$) has 73 samples which is a majority among all the classes. This is followed by class 1 ($building_{w}indows_{f}loat_{p}rocessed$) which has 67 samples.
As the number of samples in class 1 and class 2 are higher (More than half of the total samples), the distribution is not uniform. The distribution is as given below -

| Class | Number of Samples |
|---|---|
| 1 | 67 |
| 2 | 73 |
| 3 | 14 |
| 5 | 10 |
| 6 | 6 |
| 7 | 26 |

## 5.2 Performance Comparison

KNN: The training and testing accuracy for $k = 1, 3, 5, 7$ is as depicted in the tables below.

| **Training Accuracy** | | |
|---|---|---|
| K | Distance | Accuracy |
| 1 | Euclidean(L2)<br>Manhattan(L1) | 71.7949<br>75.3846 |
| 3 | Euclidean(L2)<br>Manhattan(L1) | 72.3077<br>74.8718 |
| 5 | Euclidean(L2)<br>Manhattan(L1) | 72.8205<br>71.2821 |
| 7 | Euclidean(L2)<br>Manhattan(L1) | 69.2308<br>70.2564 |

| **Testing Accuracy** | | |
|---|---|---|
| K | Distance | Accuracy |
| 1 | Euclidean(L2)<br>Manhattan(L1) | 61.1111<br>66.6667 |
| 3 | Euclidean(L2)<br>Manhattan(L1) | 66.6667<br>72.2222 |
| 5 | Euclidean(L2)<br>Manhattan(L1) | 55.5556<br>61.1111 |
| 7 | Euclidean(L2)<br>Manhattan(L1) | 55.5556<br>55.5556 |

Naive Bayes: The training and testing accuracy of Naive Bayes classifier is as depicted below -

| **Naive Bayes Accuracy** | |
|---|---|
| Training | 55.1020 |
| Testing | 33.3333 |

### 5.3 Discussion

When we compare the results of both the classifiers, we find that KNN performs the classification task better than Naive Bayes. The reason being, Conditional independence assumption of Naive Bayes. Since Naive Bayes is a generative model, the values depend on the joint distribution of the random variables. Some of the attributes corresponding to some given classes which have the variance of 0 affects the accuracy of Naive Bayes. To improve the accuracy of Naive Bayes, we can employ different smoothing techniques.

**Collaborators:** Shitesh Saurav, Piyush Gupta