<u>CSCI567 – Machine Learning</u>

Assignment #1

By,

Suraj R

USC ID: 4979219097

Question 1:

1(a) Density Estimation:

We know that Beta Distribution is given by

$$f(x) = \left(\frac{x^{\alpha-1}\,(1-x)^{\beta-1}\,(\alpha+\beta-1)!}{(\alpha-1)(\beta-1)}\right) \tag{1}$$

And

$$\beta = 1$$

Substituting this in (1), we do get,

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

Likelihood function L is given by,

$$L = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \alpha x_i^{\alpha-1}$$

Then,

$$ln(L) = ln(\prod_{i=1}^{n} \alpha x_i^{\alpha-1}) = \sum_{i=1}^{n} (\alpha-1) ln(x_i)$$

$$nln(\alpha) + \sum_{i=1}^{n} (\alpha-1) ln(x_i)$$

$$\frac{\partial lnL}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} ln(x_i) = 0$$

Hence MLE of α is,

$$\hat{\alpha} = -\frac{n}{\sum_{i=1}^{n} ln(x_i)}$$

Normal Distribution

Given that,

$$f(x) = \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right]$$

Here $\mu = \theta$ and since the variance of a diagonal matrix is also θ it follows that$\sigma^2 = \theta$.

Substituting this in the above equation,

$$f(x) = \left[\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\theta)^2}{2\theta}}\right]$$

The likelihood function as discussed in the lecture is given by,

$$L = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\theta)^2}{2\theta}}\right]$$

The log likelihood is given by,

$$ln(L) = \ln\left(\prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\theta)^2}{2\theta}}\right]\right)$$

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

$$= -\frac{n}{2} ln(2\pi\theta) - \sum_{i=1}^{n}\frac{x_i^2}{2\theta} - \frac{n\theta}{2} + \sum_{i=1}^{n} x_i$$

Taking the partial derivative with respect to $\partial\theta$

$$\frac{\partial lnL}{\partial\theta} = -\frac{n}{2\theta} + \sum_{i=1}^{n}\frac{x_i^2}{2\theta} - \frac{n}{2} = 0$$

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

Since it's a quadratic equation, the roots of the equation is,

$$\hat{\theta} = -\frac{-n \pm \sqrt{n^2 - \sum_{i=1}^{n} 4nx_i^2}}{2n}$$

Hence the MLE of θ is as above.

1 (b)

Let x1,x2,…,xn are the training inputs and y1,y2…,yn be the outputs. There are N I.i.d samples which are from the normal distribution having mean and variance

$$L = \frac{1}{\sqrt{2\pi\sigma^{2}}} \prod_{i=1}^{N} e^{\frac{-(y_i - \omega - \omega^t x_i)}{2\sigma^{2}}} \qquad (1)$$

Taking the partial derivative with respect to ω_0 to find the estimate of parameter ω_0 and setting it to zero.

$$\sum_{i=1}^{N} \omega_0 = \sum_{i=1}^{N} y_i \sum_{i=1}^{N} \omega^t x_i$$

Therefore

$$\widehat{\omega_0} = \frac{1}{n} \sum_{i=1}^{N} \omega^t x_i$$

Hence $\widehat{\omega_0} = \bar{y} - \omega^t x'$. Substituting this in 1 and taking the partial derivative with respect to $\widehat{\omega_0}$ and setting it to zero,

$$0 + \frac{1}{\sigma^{2}} \sum_{i=1}^{N} (y_i - \bar{y}) - \omega^t (x_i - \bar{x}) = 0$$

$$\widehat{\omega} = [\sum_{i=1}^{N} (x_i - \bar{x}) - \omega^t (x_i - \bar{x})][\sum_{i=1}^{N} (y_i - \bar{y}) - \omega^t (y_i - \bar{y})]$$

Also, $x_i^c = x_i - \bar{x}, y_i^c = y_i - \bar{y}$

Substituting this in the above equation,

$$\widehat{\omega} = [\sum_{i=1}^{N} (x_i^c)(x_i^c)^{T}]^{-1} [\sum_{i=1}^{N} (y_i^c)(y_i^c)^{T}]^{-1}$$

$$\widehat{\omega} = (X_c^T X_c)^{-1} X_c^T Y_c$$

Hence proved.

1 (c)

We know that,

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

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

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

$$= \frac{1}{h}\int_{2} K\left(\frac{x-t}{h}\right) f(t)dt$$

$$E_{X_1,\cdots,X_n}\left[\hat{f}(x)\right] = \frac{1}{h}\int_{2} K\left(\frac{x-t}{h}\right) f(t)dt$$

Given that $z = \frac{x-t}{h} \Rightarrow t = -zh + x$

Applying this in the above equation and by using Taylor's theorem we get,

$$E\hat{f}(x) = \int K(z)f(x-hz)dz$$

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

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

As we know that,

$$\int K(z)dz = 1, \int zK(z) = 0 \; and \; \int z^2K(z) = \sigma_K^2$$

$$= f(x) + h^2\sigma_K^2\frac{f''(x)}{2} + o(h^2)$$

$$which\ is\ equal\ to, \qquad E[\widehat{f}(x)] = f(x) + h^2\sigma_K^2\frac{f''(x)}{2} + o(h^2) \qquad (2)$$

In order to compute the bias, we have from equation 2 and the bias is given as below.

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

1. d

The kernel density estimation is given by,

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

In MLE, there is only one global minimum which we find, but in kernel, there are many local minimum. So if we estimate h using MLE, MLE will give us only one value and many will be ignored. Hence MLE cannot be used to estimate h in $\hat{f}(x)$

Question 2.a

Based on the given data,

The Features are provided in x and y coordinates as below.

$$x = \{15, -7, -4, 29, 32, 37, 18, 40, -8, -11\}$$

$$y = \{49, 38, 47, 24, 36, 43, 9, -28, -19, 12\}$$

Using the given data, we compute the mean and variance of each feature:

For x:

$$\mu_x = \frac{1}{n}\sum_{n} x_n = \frac{15 - 7 - 4 + 29 + 32 + 37 + 18 + 40 - 8 - 11}{10} = 14.1$$

$$\sigma_x^2 = \frac{1}{(n-1)}\sum (x_n - \mu_x)^2$$

$$= \frac{1}{9}[(15 - 14.1)^2 + (-7 - 14.1)^2 + (-4 - 14.1)^2 + (29 - 14.1)^2 + (32 - 14.1)^2 + (37 - 14.1)^2 + (18 - 14.1)^2 + (40 - 14.1)^2 + (-8 - 14.1)^2 + (11 - 14.1)^2]$$

$$\sigma_x^2 = 404.98$$

$$\sigma_x = 20.12$$

Similarly for y,

$$\mu_y = \frac{1}{n}\sum_{n} y_n = \frac{49 + 38 + 47 + 24 + 36 + 43 + 9 - 28 - 19 + 12}{10} = 21.1$$

$$\sigma_y^2 = \frac{1}{(n-1)}\sum (y_n - \mu_y)^2 = 743.65$$

$$\sigma_y = 27.26$$

Using the above values, we calculate the normalized and scaled co-ordinates.

$$x' = \left(\frac{x - \mu_x}{\sigma_x}\right) \; and \; y' = \left(\frac{y - \mu_y}{\sigma_y}\right)$$

Mathematics	0.04	1.01
Mathematics	-1.04	0.61
Mathematics	-0.89	-0.92
Electrical Engineering	0.74	0.10
Electrical Engineering	0.88	0.54
Electrical Engineering	1.13	0.80
Computer Science	0.19	-0.44
Computer Science	1.28	-1.80
Computer Science	-1.09	-1.47
Computer Science	-1.24	-0.33

Given the coordinates of the student: $x = 9 \; and \; y = 18$

Using the above data, the normalized coordinates are: $x = -0.25 \; and \; y = -0.11$

We need to classify the given student based on the above information.

We need to calculate L_1 and L_2 using the formula which is mentioned as below.

$$L_1 = |x - x_n| + |y - y_n|$$

$$L_2 = \sqrt{|x - x_n|^2 + |y - y_n|^2}$$

ID	Class	X_n	Y_n	L_1 Distance	L_2 Distance
1	Mathematics	0.04	1.01	1.41	1.25
2	Mathematics	-1.04	0.61	1.51	0.52
3	Mathematics	-0.89	-0.92	1.45	0.66
4	Electrical Engineering	0.74	0.1	1.2	0.04
5	Electrical Engineering	0.88	0.54	1.78	0.42
6	Electrical Engineering	1.13	0.8	2.29	0.83
7	Computer Science	0.19	-0.44	0.77	0.11
8	Computer Science	1.28	-1.8	3.22	2.86
9	Computer Science	-1.09	-1.47	2.2	1.85
10	Computer Science	-1.24	-0.33	1.21	0.05

Using the above information, we now calculate

For *K=1* and L_1

$nn(x) \; with \; L1 \; is \; 0.19$

Hence the given student belongs to Computer Science.

For $K=3$ and L_1

$nn(x)$ $with$ L_1 and $K = 3$ Is Electrical Engineering = 1.2 Computer Science = {0.77, 1.21}

Hence the given student belongs to Computer Science.

For $K=1$ and L_2

$nn(x)$ $with$ $L2$ and $K = 1$ is Electrical Engineering = 0.04

Hence the given student belongs to Electrical Engineering.

For $K=3$ and L_2

$nn(x)$ $with$ $L2$ and $K = 3$ is Computer Science {0.05, 0.11} Electrical Engineering {0.04}

Hence the given student belongs to Computer Science.

We can see that when k=3, we got the same results as Computer Science irrespective of L_1 *and* L_2. But when k=1, the neighbors changed based on L_1 *and* L_2 metrics.

Hence we can see the difference in the values.

Question 2.b

From the definition of Total Probability theorem, we have,

$\sum_n K_c = K$

$$p(x|y = c_i) = \frac{K_c}{N_c V}$$

$$p(y = c_i) = \frac{N_c}{N}$$

$$p(x) = \sum_{i=1}^{n} p(x|y = c_i) \times p(y = c_i)$$

$$\frac{K_{c_1}}{N_{c_1}V} \times \frac{N_{c_1}}{N} + \frac{K_{c_2}}{N_{c_2}V} \times \frac{N_{c_1}}{N} + \cdots + \frac{K_{c_n}}{N_{c_n}V} \times \frac{N_{c_n}}{N}$$

$$= \frac{K_{c_1} + K_{c_2} + K_{c_3} + \cdots + K_{c_n}}{NV} = \frac{K}{NV}$$

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

Now calculate $p(Y = c|x)$

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

$$= \left[\frac{\frac{K_c}{N_c V} \times \frac{N_c}{N}}{\frac{K}{N_c V}} \right] = \frac{K_c}{K}$$

$$\Leftrightarrow \boldsymbol{p(y = c|x) = \frac{K_c}{K}}$$

Question 3:

Consider linear regression of the form,

$$y(x, \omega) = \omega_0 + \omega^T x$$

And the sum of squares of error function of the form,

$$E(\omega) = \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{y(\widetilde{x_n^m}, \omega) - t_n\}^2$$

Upon substituting,

$$= \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{(\omega_0 + \omega^T (x_n + \epsilon_m))\}^2$$

$$= \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{\omega_0 + \omega^T \epsilon_m + \omega^T x_n - t_n\}^2$$

$$= \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{(y(x_n, \omega) - t_n)\}^2 + \omega^T \epsilon_m$$

$$= \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{(y(x_n, \omega) - t_n)\}^2 + \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \omega^{T^2} \epsilon_m^2$$

Using the $(a+b)^2$ formula and expanding and equating the noise term to zero, we get

$$\boldsymbol{E(\omega) = \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{(y(x_n, \omega) - t_n)\}^2}$$

3b the dropout noise corresponds to setting of $\widetilde{x_{n,d}}$ to 0 with the probability δ and $\frac{x_{n,d}}{(1-\delta)}$ with the probability $1-\delta$. Hence from the definition of variance and expected value, we know that

$$E(\widetilde{x_n}) = 0 * (\delta) + \frac{x_n}{1-\delta} * (1-\delta)$$

$$\boldsymbol{E(\widetilde{x_n}) = x_n}$$

$$Var[\omega^T \widetilde{x_n}] = E[(\omega^T \widetilde{x_n})^{\,2}] - E(\omega^T \widetilde{x_n})^{\,2}$$

$$Var[\omega^T \widetilde{x_n}] = E\big[\big((\omega^T \widetilde{x_n})^{\,2}\big] - x_n^2$$

$$= \sum_{d=1}^{D} \left[\omega_d^2 \frac{x_{n,d}^2}{(1-\delta)^{\,2}}(1-\delta)\right] - x_n^2$$

$$= \sum_{d=1}^{D} \left[\omega_d^2 \frac{x_{n,d}^2}{(1-\delta)^{\,2}} - x_{n,d}^2\right]$$

Hence

$$\boldsymbol{Var[\omega^T \widetilde{x_n}] = \frac{\delta}{1-\delta} \sum_{d=1}^{D} \omega_d^2\, x_{n,d}^2}$$

3.c By using the above results and the given E averaged over the noise distribution,

$$E(\omega) = \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \{y(\widetilde{x_n^m}, \omega) - t_n\}^{\,2}$$

$$E(\omega) = \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} [(\omega_0 - t_n) + (\omega^T \widetilde{x_n})]^{\,2}$$

$$Here\ let\ a = (\omega_0 - t_n)\ and\ b = (\omega^T \widetilde{x_n})^{\,2}\ and\ using\ (a+b)2\ formula$$

$$= \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} (\omega_0 - t_n)^{\,2} + (\omega^T \widetilde{x_n})^{\,2} + 2(\omega_0 - t_n)(\omega^T \widetilde{x_n})$$

On simplifying,

$$= \frac{1}{2} \sum_{n=1}^{N} (M\,(\omega_0 - t_n)^{\,2} + 2(\omega_0 - t_n)\omega^T \sum_{m=1}^{M} \widetilde{x_n^m} + \sum_{m=1}^{M} (\omega^T \widetilde{x_n})^{\,2}$$

Here,

$$\sum_{m=1}^{M} \widetilde{x_n^m} = M x_n\ and\ \sum_{m=1}^{M} (\omega^T \widetilde{x_n})^{\,2} = M(\omega^T \widetilde{x_n})^{\,2}$$

$$= \frac{1}{2} \sum_{n=1}^{N} (M\,(\omega_0 - t_n)^{\,2} + 2(\omega_0 - t_n)\omega^T M x_n + M(\omega^T \widetilde{x_n})^{\,2}$$

On simplifying using the results of mean and variance, we get,

$$= \frac{M}{2}\left(\sum_{n=1}^{N}(y(x_n,\omega) - t_n)^2) + \sum_{n=1}^{N}\frac{\delta}{1-\delta}\sum_{d=1}^{D}\omega_d^2\, x_{n,d}^2\right)$$

From the above proof, it has been proved that minimizing E averaging over the noise distribution is equivalent to minimizing the sum of squares error for noise free input variables with the addition of weight decay regularization term.

Question 4:

4 a. We can use 3 entropy formula,

$$H[X] = -\sum_{k=1}^{K} P(x_k) log P(x_k)$$

$$H[Y|X] = \sum_{k} P(x_k) H[Y|x_k]$$

$$IG = H[Y] - H[Y|X]$$

Entropy: $H(T) = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right) = 0.94$

$$H(temp, T) = \frac{4}{14}H(T_{hot}) + \frac{6}{14}H(T_{mild}) + \frac{4}{14}H(T_{cool})$$

$$= \frac{4}{14}\left(-\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right)\right) + \frac{6}{14}\left(-\frac{4}{6}\log\left(\frac{4}{6}\right) - \frac{2}{6}\log\left(\frac{2}{6}\right)\right) + \frac{4}{14}\left(-\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)\right)$$

$$= \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.81)$$

$$= 0.911$$

$$Gain(temperature, T) = 0.940 - 0.911 = 0.029$$

$$H(outlook, T) = \frac{5}{14}H\left(T_{sunny}\right) + \frac{4}{14}H(T_{overcast}) + \frac{5}{14}H(T_{rain})$$

$$= \frac{5}{14}\left(-\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) + \frac{4}{14}\left(-\frac{4}{4}\log\left(\frac{4}{64}\right)\right) + \frac{5}{14}\left(-\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right)\right)\right)$$

$$= 0.347 + 0 + 0.347$$

$$= 0.694$$

$$Gain(overcast, T) = 0.940 - 0.694 = 0.246$$

$$H(Humidity, T) = \frac{7}{14}H(T_{normal}) + \frac{7}{14}H(T_{high})$$

$$= \frac{7}{14}\left(\left(-\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right)\right) + \frac{7}{14}\left(-\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right)\right)\right)$$

$$= 0.493 + 0.296 = 0.789$$

$$Gain(humidity, T) = 0.151$$

$$H(wind, T) = \frac{8}{14}H(T_{weak}) + \frac{6}{14}H(T_{strong})$$

$$= \frac{8}{14}\left(-\frac{2}{8}\log\left(\frac{2}{8}\right) - \frac{6}{8}\log\left(\frac{6}{8}\right)\right) + \frac{6}{14}\left(-\frac{3}{6}\log\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right)\right)$$

$$= 0.464 + 0.429 = 0.893$$

$$Gain(wind, T) = 0.047$$

$$H(Temperature, Sunny) = \frac{2}{5}H(T_{hot}) + \frac{2}{5}H(T_{mild}) + \frac{1}{5}H(T_{cool})$$

$$= 0.4$$

$$Gain(temperature, T) = 0.571$$

$$H(humidity, Sunny) = \frac{3}{5}H(T_{high}) + \frac{2}{5}H(T_{normal}) = 0$$

$$Gain(humidity, T) = 0.971$$

$$H(temperature, rain) = \frac{3}{5}H(T_{mild}) + \frac{2}{5}H(T_{cool}) = 0.951$$

$$Gain(temperature, T) = 0.02$$

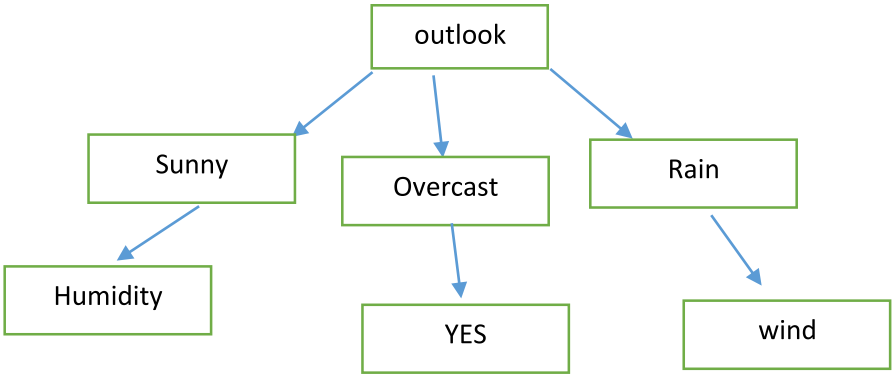
$$H(humidity, rain) = \frac{2}{5}H(T_{high}) + \frac{3}{5}H(T_{normal}) = 0.951$$

$$Gain(humidity, T) = 0.02$$

$$H(wind, rain) = \frac{3}{5}H(T_{weak}) + \frac{2}{5}H(T_{strong}) = 0$$

$$Gain(wind, T) = 0.951$$

Since Humidity and wind are having highest values at depth=2, we consider these attributes as the factors.



4 b

We know that,

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

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

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

$$\sum_{k=1}^{k} [p_k(\,(1 - p_k + lo\,g(p_k)] \leq 0$$

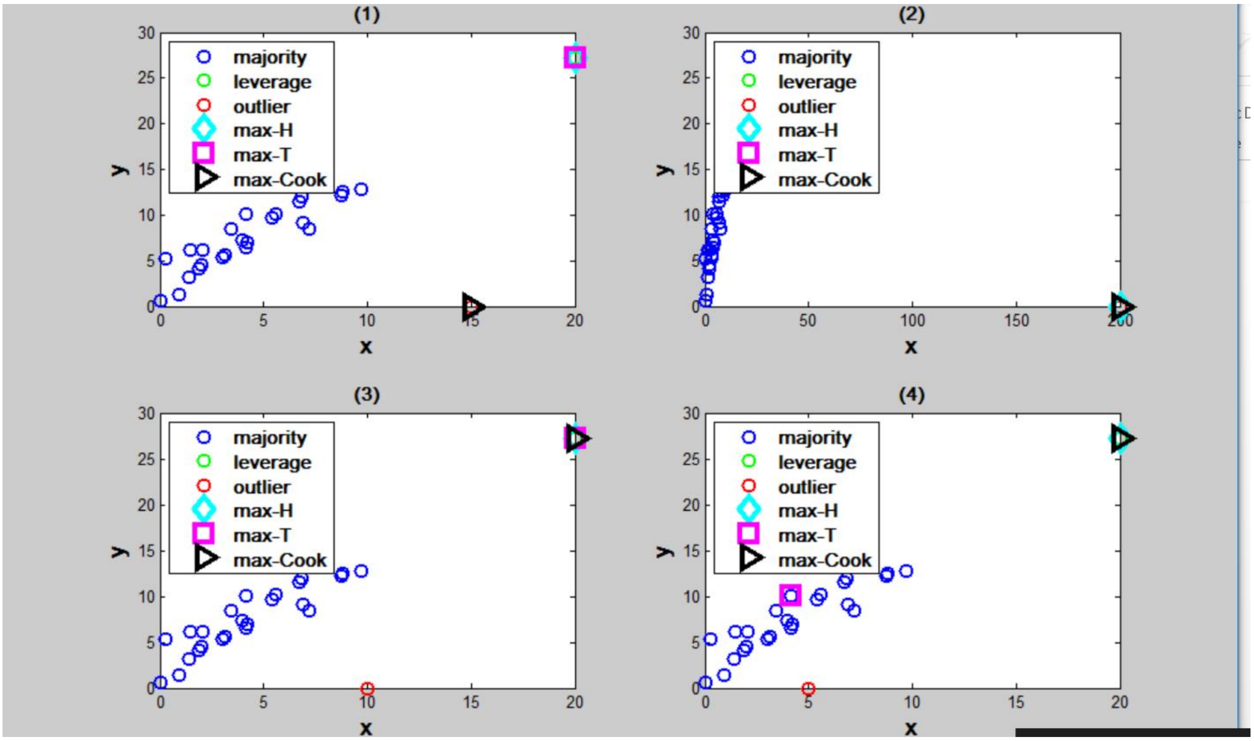
$$for\ eack\ k\ \ \in K,$$

$$p_k(1 - p_k + log(p_k)) \leq 0\ because\ 0 \leq p_k \leq 1$$

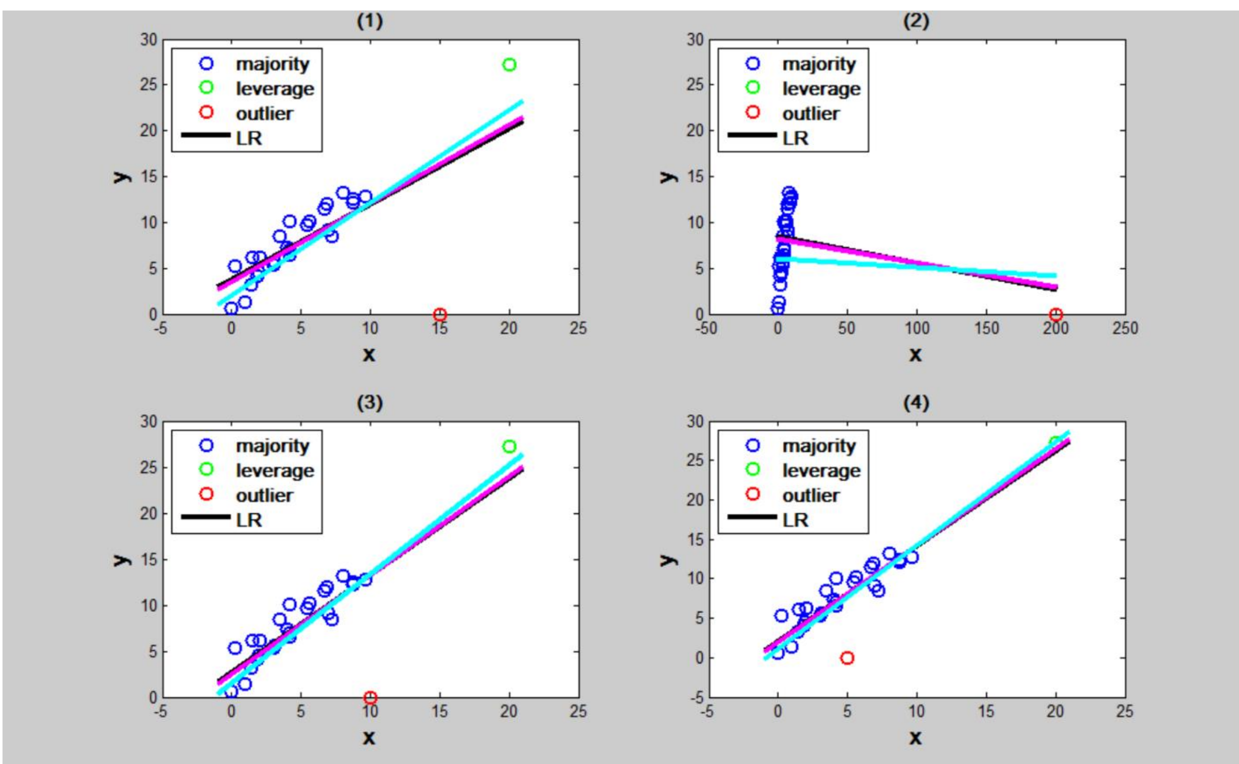
This term is less than $log(p_k)$ because$1 - p_k \geq 0$. Therefore Gini index is less than or equal to cross entropy. $[log(p_k) \leq 0 \forall\ 0 \leq p_k \leq 1$

Question 5.1

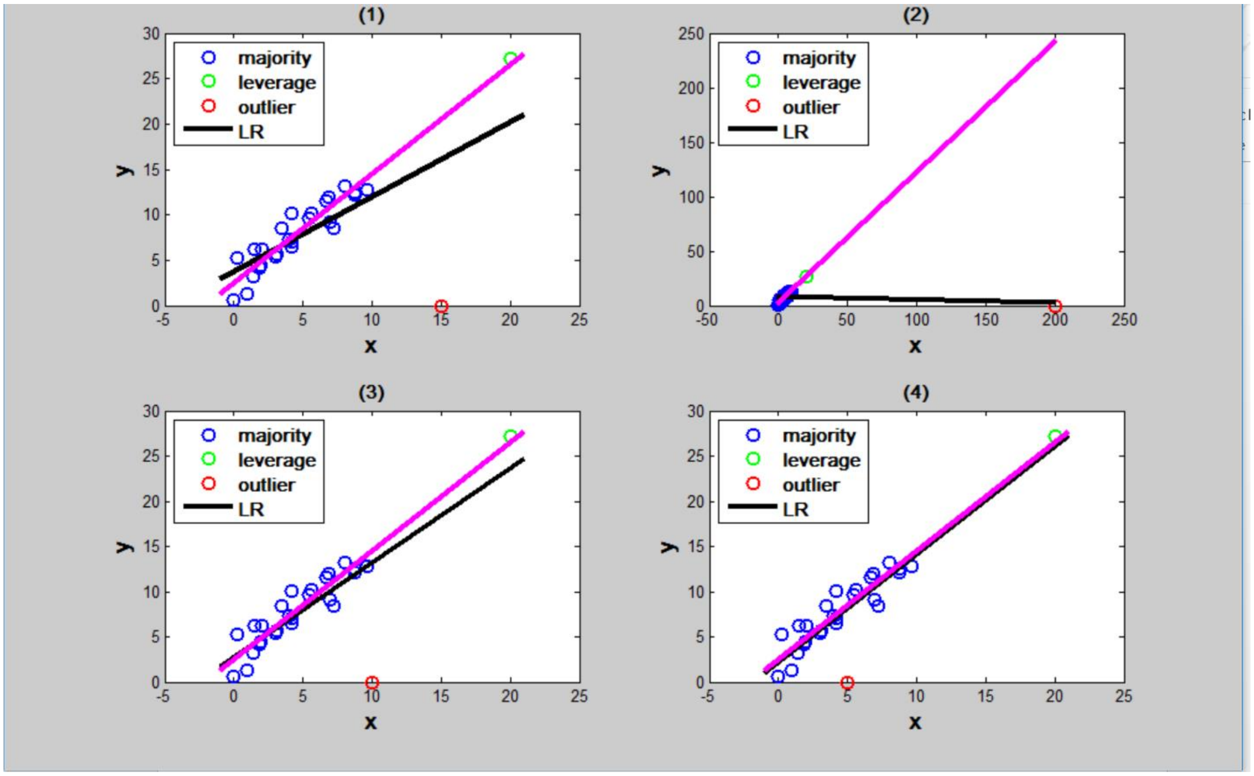
Linear regression model for the dataset data1



Linear regression model for the dataset data2



Linear regression model for the dataset data3 and data4



- The weight decay parameter doesn't significantly reduce the influence of outlier sample. Larger the value of lambda greater will be the penetration hence variance will decrease but bias increases.

- We could significantly reduce if we use Gaussian distribution because Laplace distribution has sharp peak when compared to Gaussian. Also Laplace distribution is quadratic and it grows faster than Gaussian.

Question 5 a.

Probability density function for Laplace is,

$$f(x|\mu, b) = \frac{1}{2b} e^{\left(-\frac{(x-\mu)}{b}\right)}$$

If Laplace distribution is used to model the noise,

$$f(noise|\mu, b) = \frac{1}{2b} e^{\left(-\frac{noise)}{b}\right)}$$

On taking log likelihood,

$$Lf(noise|\mu, b) = ln\, \pi_{i=1}^{n} \frac{1}{2b} e^{\left(-\frac{noise)}{b}\right)}$$

$$= ln \sum_{i=1}^{n} \frac{1}{2b} e^{\left(-\frac{noise)}{b}\right)}$$

$$= -n\, ln\, 2b - \sum_{i=1}^{n} \frac{|y_n - \omega^T x|}{b}$$

Thus inorder to maximize the left hand side, the right hand side should be minimized. Hence the objective function is $\sum_{i=1}^{n} \frac{|y_n - \omega^T x|}{b}$

Question 5.2

Leave one out strategy

- Test and Validation datasets using Decision Tree and *gini index* split criterion.

MinLeaf	Training	Test	Validation
1	94.88679	98.53488	97.87234
2	95.88679	98.06977	96.34043
3	95.88679	98.13953	96.74468
4	95.96226	97.13953	96.74468
5	95.81132	97.27907	96.74468
6	95.73585	97.27907	95.21277
7	95.73585	96.27907	95.21277
8	95.28302	96.27907	94.21277
9	96.92453	96.81395	93.55319
10	96.84906	96.81395	93.55319

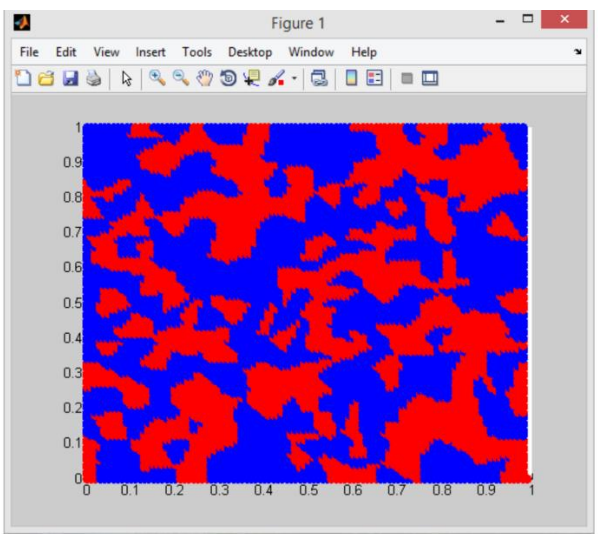
Test and Validation datasets using Decision Tree and *cross – entropy* split criterion.

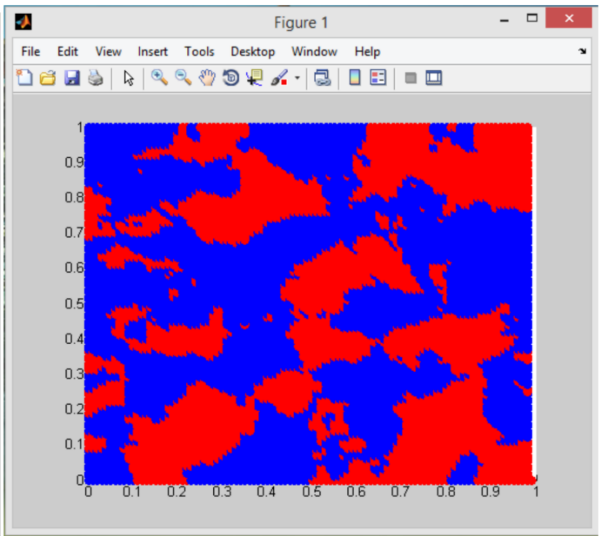
MinLeaf	Training	Test	Validation
1	97.26415	98.06977	96.340426
2	97.26415	98.67442	95.808511
3	97.26415	98.67442	95.744681
4	97.26415	96.67442	94.744681

5	98.81132	96.2093	96.744681
6	98.73585	96.81395	96.212766
7	98.73585	96.81395	97.212766
8	96.73585	94.81395	93.212766
9	97.83019	94.81395	92.553191
10	97.75472	96.81395	94.553191

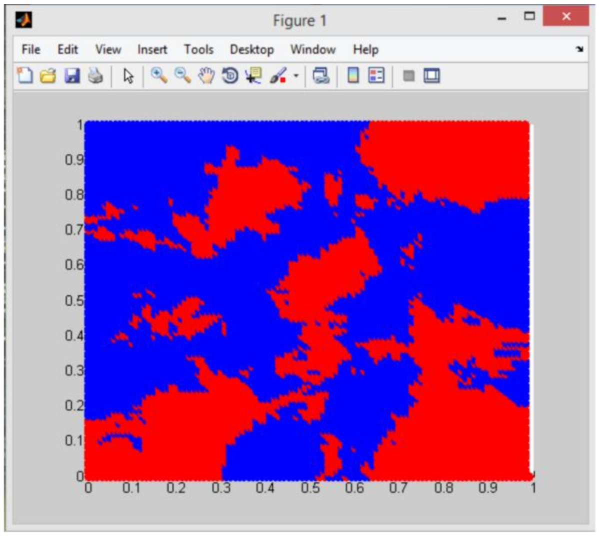
- Test and Validation datasets using k nearest neighbor approach and values of k as: 1,3,5, …… 15

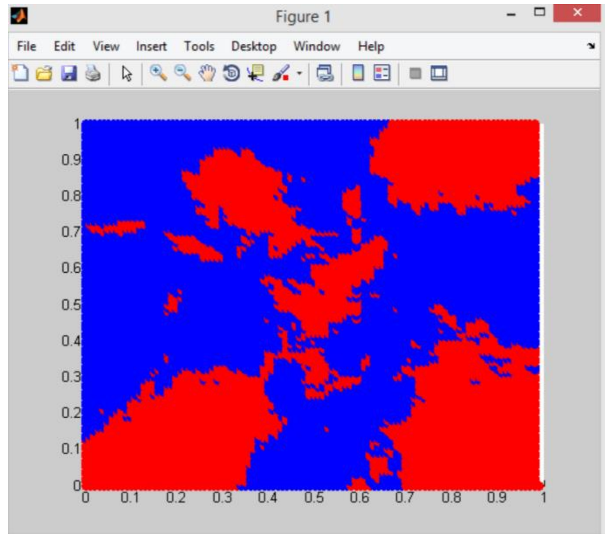
K	Training	Test	Validation
1	54.33962	55.74419	63.70213
3	91.81132	63.93023	63.70213
5	93.09434	59.13953	63.70213
7	92.26415	56.53488	63.70213
9	94.09434	60.74419	63.70213
11	91.39623	60.74419	63.70213
13	93.4717	60.74419	63.70213
15	94.01887	61.74419	63.70213





K=1 K=5





K=15 k=25

Collaboration: Neel shah