ID5001W: Machine learning and its applications
Midsem Exam

Note: Scanned copy of hand written text has been attached and mobile image is placed in zip folder. In case of any hazy or dark shade of scan copy image is found refer to mobile image.

(2) **(Multiclass Logistic Regression.)**

i. Let $X|Y = i$ be distributed as the multivariate normal given by $\mathcal{N}(\mu_i, \sigma^2 I)$ for all $i \in [K]$. Let π_i be equal to $P(Y = i)$. What is the posterior probability $P(Y = i|X = \mathrm{x})$?

ii. Consider the three class, 1-dimensional dataset, with 6 data points. With feature given by x and class label given by y.

x	3	2	5	5	7	8
y	1	1	2	2	3	3

The multinomial logistic loss is given as :

$$L = \sum_{i=1}^{6} -\log\left([\mathrm{SM}\,(w_1 x_i + b_1, w_2 x_i + b_2, w_3 x_i + b_3)]_{y_i}\right)$$

where SM is the softmax function from $\mathbb{R}^3 \to \mathbb{R}^3_+$ and the parameters are w_j, b_j for $j \in \{1, 2, 3\}$. Give a setting for the parameters so that $L < 0.1$. Argue that the loss can be made arbitrarily close to zero for some setting of the parameters.

iii. Consider the same dataset as above. The loss minimised in one-vs-all logistic regression is:

$$L = \sum_{i=1}^{6}\sum_{j=1}^{3} -\log\left(\sigma(y_{ij}(w_j x_i + b_j))\right)$$

where σ is the sigmoid function. $y_{ij} = +1$ if $y_i = j$ and -1 otherwise. Show that for any setting of $w_1, w_2, w_3, b_1, b_2, b_3$ the loss L is greater than $2\log(2)$.

iv. Repeat the two sub-problems above, with the 2-dimensional 4-class dataset with 8 points given below as well. Note that the parameters are $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4$ and b_1, b_2, b_3 and b_4, with $\mathbf{w}_j \in \mathbb{R}^2$ and $b_j \in \mathbb{R}$. The multinomial logistic and one-vs-all loss expressions also change appropriately.

x_1	1	2	3	4	3	4	7	7
x_2	1	0	4	3	6	6	2	3
y	1	1	2	2	3	3	4	4

(1+1+2+2 points)

Ans.

Q2-Screenshot.zip

2.

(i) considering, as per given in the question, $x|y=i$, be distributed as the multivariate normal given by $N(\mu_i, \sigma^2 I)$ for all $i \in [K]$

Also it is given that π_i is equal to $P(y=i)$

As per the question we have to find,

$$P(y=i \mid x=x)$$

$$= \frac{P(x=x \mid y=i)\, P(y=i)}{\sum_j P(x=x \mid y=j) * P(y=j)}$$

Now as the distribution is gaussian distribution, we are plotting the pdf equation,

$$\text{Numerator} = \left[(2\pi\sigma^2)^{-K/2} e^{\left[(x-\mu_i)^T (x-\mu_i) \right] / 2\sigma^2} \right] \pi_i$$

[considering $x|y=i \sim N(\mu_i, \sigma^2 I)$ $\forall i \in [K]$]

$$\text{Denominator} = \sum_j (2\pi\sigma^2)^{-K/2} e^{-\frac{\left[(x-\mu_i)^T (x-\mu_i) \right]}{2\sigma^2}} \pi_j$$

$\therefore$ posterior probability $(P(y=i \mid x=x))$

$$= \frac{\text{Numerator}}{\text{Denominator}}$$

$$= \frac{\left[(2\pi\sigma^2)^{-K/2} e^{\left[(x-\mu_i)^T (x-\mu_i) \right] / 2\sigma^2} \right] \pi_i}{\sum_j (2\pi\sigma^2)^{-K/2} e^{-\left[(x-\mu_i)^T (x-\mu_i) \right] / 2\sigma^2} \pi_j}$$

(ii) considering 3 class in 1 dimensional dataset with 6 data points.

Feature → x & label / target → (y)

x	3	2	5	5	7	8
y	1	1	2	2	3	3

multinomial logistic loss is given by,

$$L = \sum_{i=1}^{6} -\log\left(\left[SM(w_1 x_i + b_1,\ w_2 x_i + b_2,\ w_3 x_i + b_3)\right]_{y_i}\right)$$

where SM = softmax function $\mathbb{R}^3 \to \mathbb{R}^3$

The soft max function maps K-dimentional real valued vector Z to K dimentional vector of real values between 0 & 1 that add upto 1, which can be written as,

$$\text{SoftMax}_j = \frac{\exp(z_j)}{\sum_{i=1}^{K} \exp(z_i)}$$

for the given data, the goal is to find the parameter w_1, w_2, w_3 & bias b_1, b_2, b_3 such that it minimize the multinomial logistic loss. This loss can be arbitarily close to zero by choosing the parameter appropiately., This is because the soft max function continous in nature, so the exponential function in the soft max can make the output arbitarily close to one. This inturn, means that the negative logarithm of the soft max output can be made arbitrarily close to zero.

To find the appropiate setting of the parameter, one common approach is to used gradient-descent algorithm. The gradient of the loss w.r.t to parameter is calculated & then parameters are updated in the direction of negative gradient to minimize the loss. This process will repeat until convergence

Alternatively it is also possible to find an analytical solution by solving the system of eq. derived from the gradient of the loss.

It can be shown that loss is a convex wrt parameter, so there is a unique global minima. This means that for any settings of the parameter, it is possible to find another settings that gives a lower loss.

However it is not possible to determine any specific settings of the parameter that guarantees the loss to be less than 0.1 without actually finding the minimum of the loss function through optimization or analytical solution

iii) The loss minimized in one vs all logistic regression is,

$$L = \sum_{i=1}^{6} \sum_{j=1}^{3} -\log\left(\sigma\left(y_{ij}\left(w_j x_i + b_j\right)\right)\right)$$

where σ is sigmoid function, $y_{ij} = +1$ if $y_i = j$ & -1 otherwise

Now we need to calculate the negative log likelihood over all the classes & for all the data points.

Since the sigmoid function is always between 0 to 1, we have,

$$0 < \sigma(x) < 1 \quad \text{therefore } \log(\sigma(x)) > 0$$

& $\sigma(x) \to 1$ as $x \to \infty$ so $-\log(\sigma(x)) \to 0$ as $x \to \infty$

$\sigma(x) \to 0$ as $x \to -\infty$ $-\log(\sigma(x)) \to \infty$ as $x \to -\infty$

therefore,

$$\log\left(\sigma\left(y_{ij}\left(w_j x_i + b_j\right)\right)\right)$$

$$= -\log(-1/2) = -\log(2^{-1}) = \log 2$$

Since we have 6 data point & 3 classes, the minimum loss will be,

$$6 * 3 * \log(2) = 18 \log 2$$

That means for any settings of the parameters, the loss will always greater than $2\log(2)$.

(3) **(Kernel Regression)** Consider the following kernel regression problem. The data matrix containing 3 points with one dimension is given by $X^\top = [-1, 0, 2]$. The regression targets are given by $\mathbf{y}^\top = [1, 2, 0]$. Consider the feature vector regression problem given by the objective:

$$R(\mathbf{w}) = \sum_{i=1}^{3} (\mathbf{w}^\top \phi(x_i) - y_i)^2$$

where $\phi : \mathbb{R} \to \mathbb{R}^d$ is a feature vector corresponding to the kernel $k(u, v) = \sin(u)\sin(v) + \cos(u)\cos(v) + 1$.

i. Solve the kernel regression problem and give the solution $\alpha_1^*, \alpha_2^*, \alpha_3^*$. Does the problem have unique or multiple solutions?

ii. Use the above $\boldsymbol{\alpha}^*$ to make predictions at the 11 points ranging from $x = -5$ to $x = 5$ in steps of 1. Plot this as a curve.

iii. Give any feature mapping $\phi : \mathbb{R} \to \mathbb{R}^d$ such that $\phi(u)^\top \phi(v) = k(u, v)$

iv. Give the solution $\mathbf{w}^*$ to the feature vector regression problem assuming the feature function ϕ got above. **(2+2+1+1 points)**

Ans.

Q3-Screenshot.zip

3. Given, $x^T = [-1, 0, 2]$

$y^T = [1, 2, 0]$

consider the feature vector regression problem given by objective,

$$R(w) = \sum_{i=1}^{3} (w^T \phi(x_i) - y_i)^2$$

$\phi : \mathbb{R} \to \mathbb{R}^d$ is a feature vector corresponding to the kernel $k(u,v) = \sin(u)\sin(v) + \cos(u)\cos(v) + 1$ ... (i)

i) kernel matrix can be written as,

$$K = \begin{bmatrix} k(-1,-1) & k(-1,0) & k(-1,2) \\ k(0,-1) & k(0,0) & k(0,2) \\ k(2,-1) & k(2,0) & k(2,2) \end{bmatrix}$$

computing the kernel value with an assumption that given value are all in radian,

plotting the u & v value in kernel equation in (i),

$$K = \begin{bmatrix} 2 & 1.54 & 0.01 \\ 1.54 & 2 & 0.584 \\ 0.01 & 0.584 & 2 \end{bmatrix}$$

looking at the matrix it is evident that it is a symmetric matrix.

we know that,

$\alpha^* = (K + \lambda I)^{-1} y$, now putting $\lambda = 0$

we have got, $\alpha^* = K^{-1} y$

Now calculating α^*,

$$\alpha^* = \begin{bmatrix} 2 & 1.54 & 0.01 \\ 1.54 & 2 & 0.584 \\ 0.01 & 0.584 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

Calculating,

$$\Rightarrow \alpha^* = \begin{bmatrix} 1.4122258 & -1.118672 & 0.33936 \\ -1.18672 & 1.54378 & -0.4447 \\ 0.33936 & -0.4447 & 0.62813 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

$$\Rightarrow \alpha^{(*)} = \begin{bmatrix} -0.961 \\ 1.901 \\ -0.550 \end{bmatrix}$$

So, $\alpha_1^* = -0.961$, $\alpha_2^* = 1.901$ & $\alpha_3^* = -0.550$

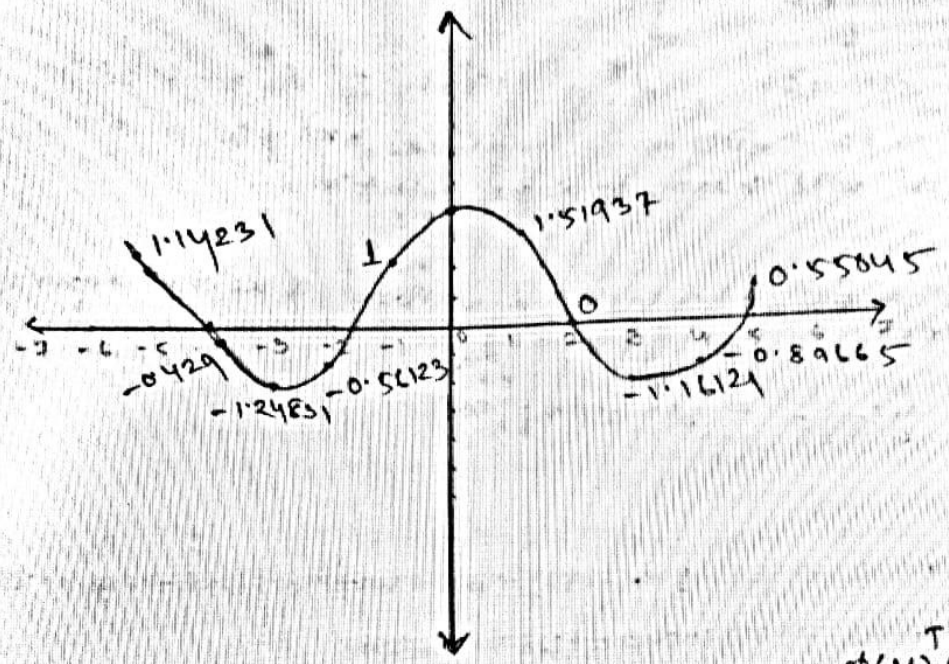
∴ since K is a full rank matrix, $|K| \neq 0$, that means equation will have unique solution.

(ii) Using the pre calculated α values & kernel equation we can compute y hat (prediction).

$$\hat{y}(x) = \sum_{i=1}^{m} \alpha_i^* K(x_i, x)$$

given X value	$K(x_1,x)*\alpha_1$	$K(x_2,x)*\alpha_2$	$K(x_3,x)*\alpha_3$	y prediction
-5	-0.33291	2.4400	-0.9648	1.14231
-4	-0.00962	0.6583	-1.0782	-0.42954
-3	-0.5612	0.01902	-0.70614	-1.24831
-2	-1.48052	1.109816	-0.19053	-0.56123
-1	-1.9223	2.92278	-0.0051	1
0	-1.48052	3.801647	-0.32118	2
1	-0.56119	2.92278	-0.8473	1.51937
2	-0.00962	1.10981	-1.1002	0
3	-0.3329	0.019023	-0.84732	-1.16121
4	-1.23384	0.65837	-0.32118	-0.89665
5	-1.86409	2.44084	0.00551	0.55045

Plotting the ypred value we can see



(iii) feature mapping $\phi: \mathbb{R} \to \mathbb{R}^d$ such that $\phi(u)^T \phi(v) = K(u,v)$

consider $\phi(x) = \begin{bmatrix} \sin(x) \\ \cos(x) \\ 1 \end{bmatrix}$

kernel equation $K(u,v) = \phi(u)^T \phi(v)$

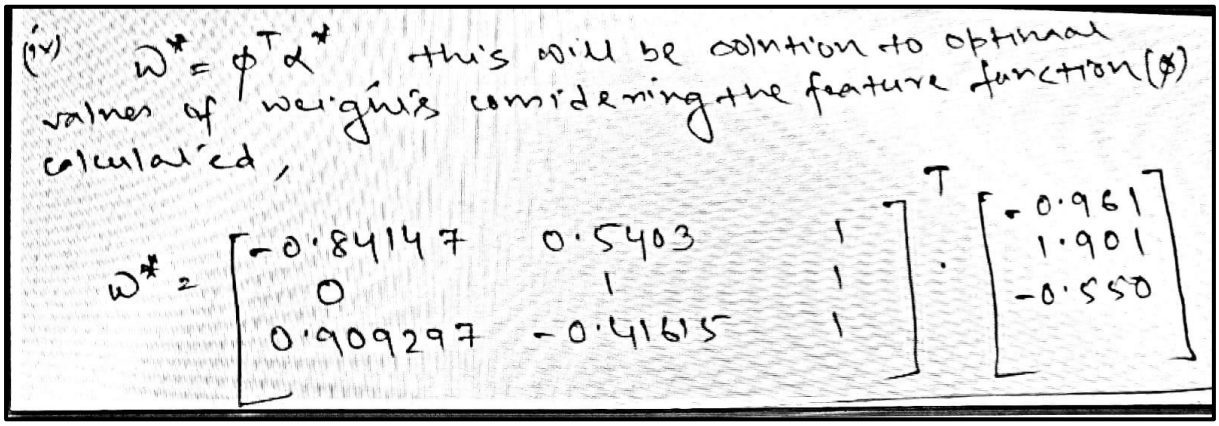
$$= \begin{bmatrix} \sin(u) & \cos(u) & 1 \end{bmatrix} \begin{bmatrix} \sin(v) \\ \cos(v) \\ 1 \end{bmatrix}$$

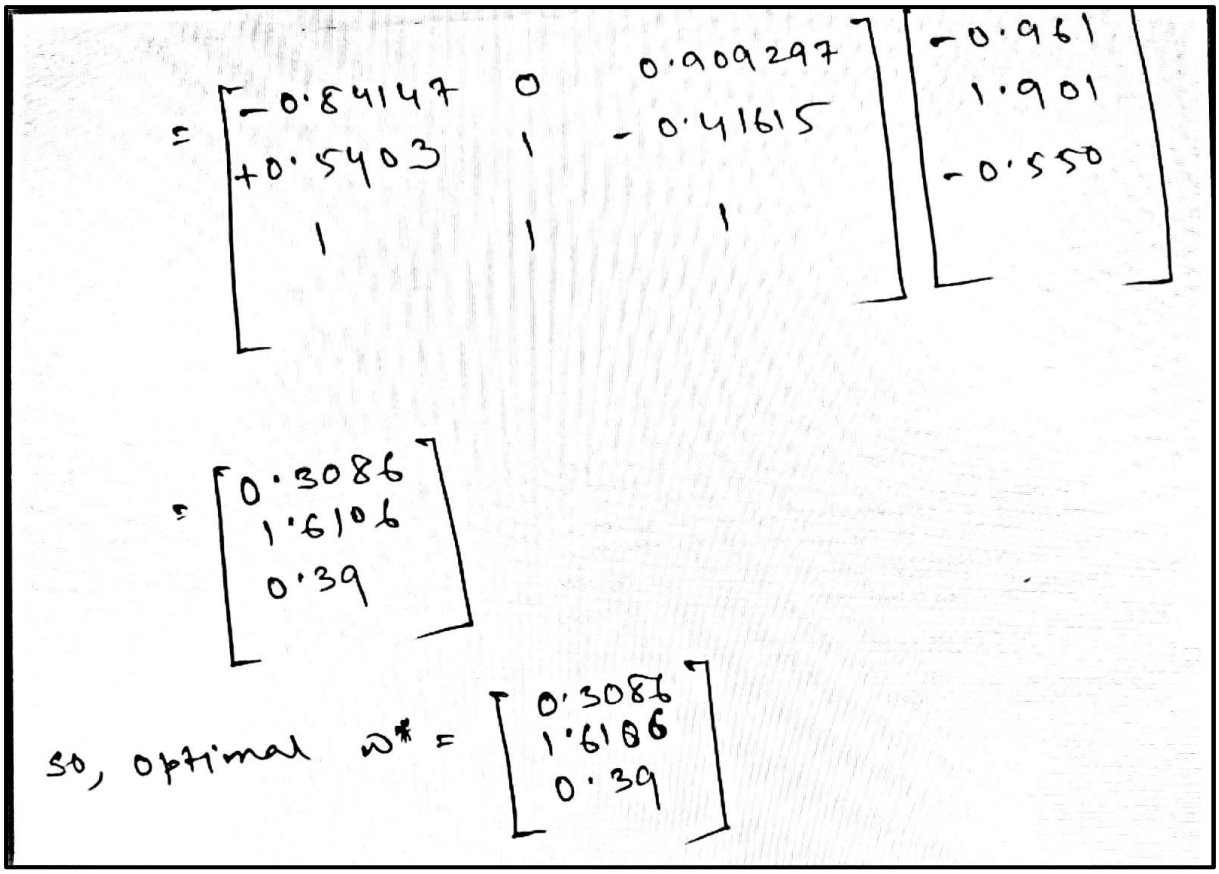
$$= \sin(u)\sin(v) + \cos(u)\cos(v) + 1$$

$$= \begin{bmatrix} \sin(-1) & \cos(-1) & 1 \\ \sin(0) & \cos(0) & 1 \\ \sin(2) & \cos(2) & 1 \end{bmatrix}$$

considering the given value are all in radian,

$$= \begin{bmatrix} -0.84147 & 0.5403 & 1 \\ 0 & 1 & 1 \\ 0.909297 & -0.41615 & 1 \end{bmatrix}$$



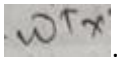


(4) **(Maximum Likelihhood.)** Consider the following parameter estimation problem. Let $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ be known constants. The d-dimensional instance vectors $X_1, X_2, \ldots, X_n$ are drawn from some distributions in an i.i.d. fashion. The real valued targets $Y_1, \ldots, Y_n$ are such that Y_i is drawn from a Normal distribution with mean $\mathbf{x}_i^\top \mathbf{w}^*$ and variance σ_i^2, for some fixed but unknown parameter $\mathbf{w}^* \in \mathbb{R}^d$. Derive the maximum likelihood estimate of $\mathbf{w}^*$.

Assume you have access to an equation solver sub-routine that takes in $A \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ and returns a solution to $A\mathbf{x} = \mathbf{b}$ (if a solution exists). How will you use this solver for this parameter estimation problem, for a given dataset with instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$, targets $y_1, \ldots, y_n$ and noise variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. **(4+2 points)**

Ans.

Q4 Screenshot.zip

Considering the dot product is symmetric in nature, $\mathbf{x}_i^\top \mathbf{w}^*$ will yield approximately same result as  . Only things to keep in mind is that the orientation of feature and datapoint will move from column to row and vice versa.

4)

d-dimentional instance vectors $x_1, x_2 \cdots x_n$ are drawn from some iid & real valued target $y_1 \cdots y_n$ are such that it follows Normal distribution with mean $x_i^T w^*$ & variance σ_i^2 which is fixed but unknown

So according the pdf of normal distribution we can say our y has pdf of,

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\sigma^2 (x - w^T x)^T (y - w^T x)}$$

According to the MLE estimates we need to maximize, $P(y|w)$, conditional probability (density)

$$\arg\max_w P(y|w)$$

We know conditional density $= \frac{\text{Joint density}}{\text{Marginal density}}$

So, $P(y|w) = \frac{P(y, w)}{P(w)}$ where,

$[P(y|w) = P(y)$ where $P(y) = P(y_1, y_2 \cdots y_n)$ & $y_1, y_2 \cdots y_n$ are drawn from the iid with normal distribution $N(w^T x_i, \sigma_i^2)]$

$$L(w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{(-1/2\sigma^2 (y_i - w^T x_i)^2)}$$

taking log on both side to simplified the equation,

$$L(L(w)) = \sum_{i=1}^{n} -\frac{1}{2\sigma^2} (y_i - w^T x_i)^2 + \text{constant term}$$

Now derivating both side & equating it to 0 to find optimal value,

$$\nabla L(L(w)) = 0$$

$$\Rightarrow \frac{1}{2\sigma^2} \cdot 2 \cdot (w^T x_i - y_i) x_i = 0$$

rewriting the equation in

where, $(x^T x) \rightarrow d \times d$ $\quad x^T \rightarrow d \times n$

$w \rightarrow d \times 1$ $\quad y \rightarrow n \times 1$

so LHS we have $d \times 1$ dimension,

RHS we have $d \times 1$ dimension

∴ dimentionality is matching, so finalizing the optimal (estimate) of w^* is,

$$\hat{w}^* = (x^T x)^{-1} x^T y \quad \text{..... (i)}$$

Now according to the 2nd part of the question, we have access to an equation solver subroutine that takes in $A \in \mathbb{R}^{d \times d}$ & $b \in \mathbb{R}^d$ & returns a solution to $Ax = b$ (if solution exists)

Now using dimentional analysis we know that for $Ax = b$, A is $(d \times d)$ & b is d dimention

if solution exists then,

$$Ax = b$$

$$\Rightarrow x = A^{-1} b$$

$$(d \times 1) \quad (d \times d) \cdot (d \times 1)$$

As per the optimal estimates of coefficients we get, $\hat{w}^* = (x^T x)^{-1} x^T y$ where,

$(x^T x) \rightarrow \sigma^2$ (noise variance)

$x \rightarrow$ Input &

$y \rightarrow$ output / target

solver has to check,

a) condition number should be very low for $(x^T x)$ term so that inverse exists. (In python compiler even for high condition number inverse is calculated but that will eventually yield bad estimates so must be avoided, hence adding this assumption now for solver).

b) b must be in a column space of A (to avoid no sol^n) & that is basis vector of column space of $(x^T x)$, must linearly combine to give $x^T y$.

c) finally solver has to compute the inverse of $(x^T x)$ term & pre multiplies it with $(x^T y)$ to yield the optimal estimates of the coefficients. $(\hat{w}^*)$.

(5) **(AdaBoost.)** Consider the following binary classification dataset. Run AdaBoost for 3 iterations on the dataset, with the weak learner returning a best decision stump (equivalently a decision tree with one node) (equivalently a horizontal or vertical separator). Ties can be broken arbitrarily. Give the objects asked for below. Highlight your answer by boxing it.

x_1	x_2	y
1	1	+1
1	2	−1
1	3	+1
2	1	−1
2	2	−1
2	3	−1
3	1	+1
3	2	+1
3	3	+1

i. Give the weak learners h_t for $t = 1, 2, 3$.

ii. Give the "edge over random" γ_t, and the multiplicative factor β_t for $t = 1, 2, 3$.

iii. Give the predictions of the final weighted classifier h on the training points.

(2+2+2 points)

Ans.

Q5 Screenshot.zip

Copying the Adaboost algorithm for reference of calculation,

Algorithm:

$\mathbf{w}^1 = \mathbf{1}$

For $t = 1$ to T:

$h_t = \text{WeakLearner}(S, \mathbf{w}^t)$

$\gamma_t = \frac{1}{2} - \frac{1}{\sum_i w_i^t} \sum_{i=1}^{m} w_i^t \frac{|h_t(x_i) - y_i|}{2}$

$\beta_t = (0.5 + \gamma_t)/(0.5 - \gamma_t)$

$l_{t,i} = |h_t(x_i) - y_i|/2$

$w_i^{t+1} = w_i^t \beta_t^{l_{t,i}}$

End For

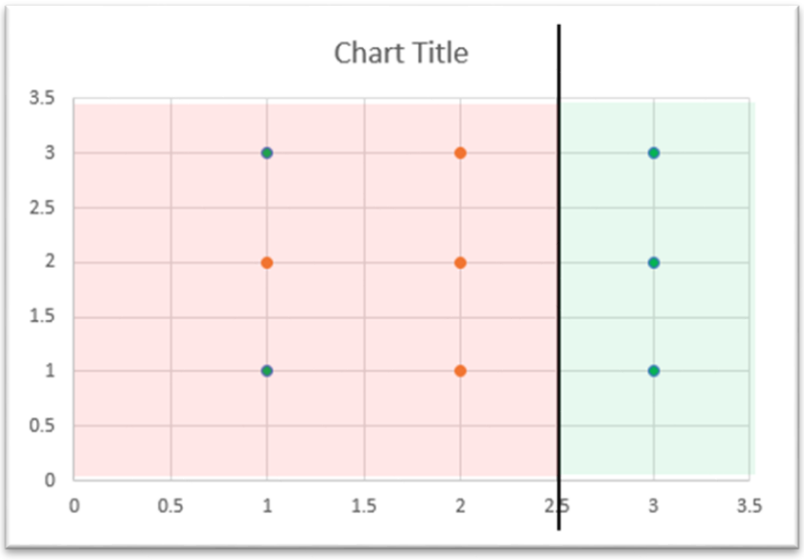
Output: $h(x) = \text{sign}\left(\sum_{t=1}^{T} (\log(\beta_t) h_t(x))\right)$

Begin the iteration for t =1, 2 and 3. Graphs are plotted, and corresponding table is shown in below,

Iteration 1:

CH22M503 – Aloy Banerjee
Mid-Sem Exam

ID5001W: Machine learning and its applications
Date – 12-02-2023



Accuracy score, When,

if $1.5 \geq x_1 > 1.5$ score = 5/9 = 0.555

if $2.5 > x_1 > 2.5$ score = 7/9 = 0.777

if $1.5 > x_2 > 1.5$ score = 5/9 = 0.555

if $2.5 > x_2 > 2.5$ score = 5/9 = 0.555

Now performing the iteration, **Iteration 1** & $r = 1$

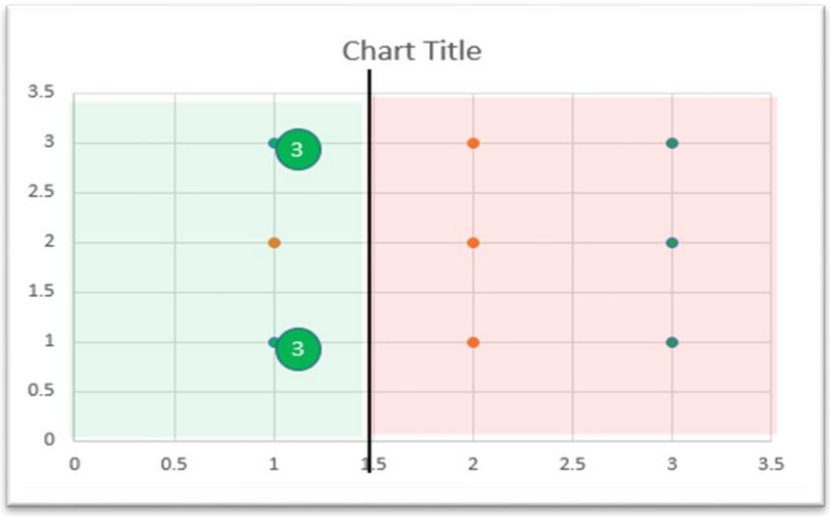
x_1	x_2	y	$\hat{y}_1$	w_1	g_1	z_1	w_2
1	1	+1	-1	1	1	1	3.5
1	2	+1	-1	1	0	0	1
1	3	+1	-1	1	1	1	3.5
2	1	-1	-1	1	0	0	1
2	2	-1	-1	1	0	0	1
2	3	-1	-1	1	0	0	1
3	1	+1	+1	1	0	0	1
3	2	+1	+1	1	0	0	1
3	3	+1	+1	1	0	0	1

$\gamma_1 = 0.5 - (2/9) = 0.27778$ & $\beta_1 = \frac{(0.5 + 0.27778)}{(0.5 - 0.27778)} = 3.5$

Iteration 2:

CH22M503 – Aloy Banerjee
Mid-Sem Exam

ID5001W: Machine learning and its applications
Date – 12-02-2023



Now after the iteration 1 we found updated weight for 1 & 3 rd data point.

Accuracy score will be,

if $1.5 > x_1 > 1.5$ score = $9/13 = 0.6923$

if $2.5 > x_1 > 2.5$ score = $9/13 = 0.6923$

if $1.5 > x_2 > 1.5$ score = $9/13 = 0.6923$

if $2.5 > x_2 > 2.5$ score = $9/13 = 0.1923$

Now performing the iteration, **Iteration 2 or i = 2**

x1	x2	y	h2	D2	g2	l2	w3
1	1	+1	1	3.5	0	0	3.5
1	2	−1	1	1	1	1	2.5
1	3	+1	1	3.5	0	0	3.5
2	1	−1	−1	1	0	0	1
2	2	−1	−1	1	0	0	1
2	3	−1	−1	1	0	0	1
3	1	+1	−1	1	1	1	2.5
3	2	+1	−1	1	1	1	2.5
3	3	+1	−1	1	1	1	2.5

$\gamma_2 = 0.5 - (4/14) = 0.214286$

& $\beta_2 = \dfrac{0.5 + 0.214286}{0.5 - 0.214286}$

$= 2.5$

Iteration 3:

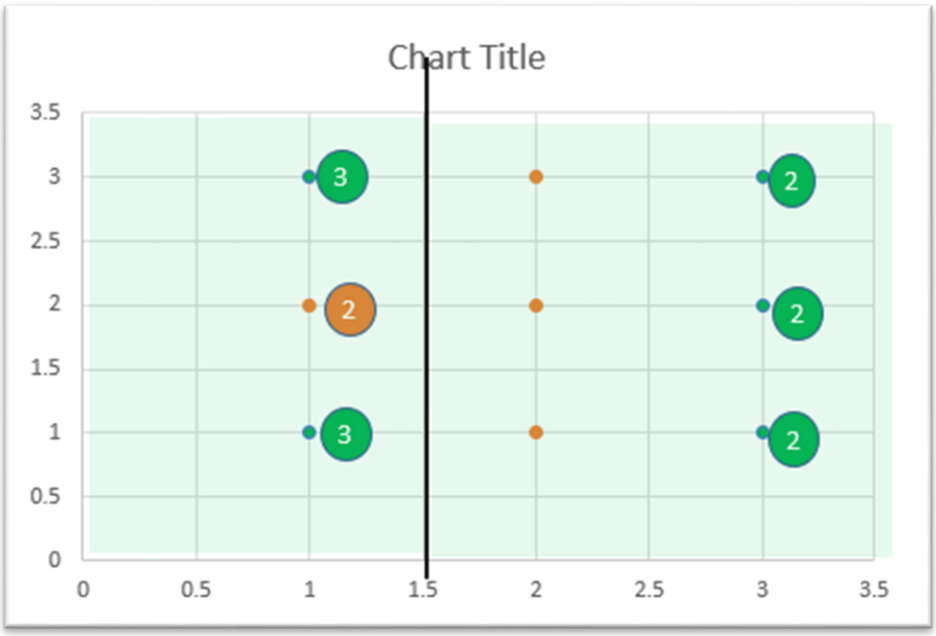


Chart Title
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
3.5
3
2
2
2
3
2

After the iteration2 we have seen that the weight of the 2,7,8 & 9 data point is modified & 1 & 3 rd remain same.

Accuracy score will be,

if $1.5 > x_1 > 1.5$ score = $12/17 = 0.705882$

if $2.5 > x_1 > 2.5$ score = $12/17 = 0.705882$

if $1.5 > x_2 > 1.5$ score = $12/17 = 0.705882$

if $2.5 > x_2 > 2.5$ score = $12/17 = 0.705882$

Now performing the 3rd iteration. **Iteration 3 or t=3**

x_1	x_2	y	h_3	w_3	g_3	l_3	w_4
1	1	+1	1	3.5	0	0	3.5
1	2	−1	1	2.5	2.5	1	6.5909
1	3	+1	1	3.5	0	0	3.5
2	1	−1	1	1	1	1	2.6363
2	2	−1	1	1	1	1	2.6363
2	3	−1	1	1	1	1	2.6363
3	1	+1	1	2.5	0	0	2.5
3	2	+1	1	2.5	0	0	2.5
3	3	+1	1	2.5	0	0	2.5

$\gamma_3 = 0.5 - (5.5/20) = 0.225$

& $\beta_3 = \dfrac{0.5 + 0.225}{0.5 - 0.225} = 2.6363$

(ii) for different iteration "edge over random" (γ) & "Multiplicative factor" (β) will be as below.

Iteration	γ	β
1	0.27778	3.5
2	0.214286	2.5
3	0.225	2.63

iii) As per the algorithm,

$$h(x) = \text{sign}\left(\sum_{t=1}^{T} (\log(\beta_t) \cdot h_t(x))\right)$$

$\log(\beta_t)\cdot h_t(x)$ $t=1$	$t=2$	$t=3$	Sum of $\sum_{t=1}^{T}(\log(\beta_t)\cdot h_t(x))$	Predicted $h(x)$	True value y
-0.54407	0.39794	0.421005	0.274877	+1	+1
-0.54407	0.39794	0.421005	0.274877	+1	-1
-0.54407	0.39794	0.421005	0.274877	+1	+1
-0.54407	0.39794	0.421005	-0.521	-1	-1
-0.54407	0.39794	0.421005	-0.521	-1	-1
-0.54407	0.39794	0.421005	-0.521	-1	-1
0.544068	0.39794	0.421005	0.567133	+1	+1
0.544068	0.39794	0.421005	0.567133	+1	+1
0.544068	0.39794	0.421005	0.567133	+1	+1

considering the above table we can see there is only one data point, is mis classified as -1 for data point 2.

(6) (**Naive Bayes Methods**) Consider a distribution over (X, Y) given by the following assumptions:
$Y \in \{-1, +1\}, X \in \{0, 1\}^3$.
$P(Y = +1) = a, \mathbf{P}(Y = -1) = 1 - a$,
$X|Y = -1 \sim \text{Bern}(\theta_1) \times \text{Bern}(\theta_2) \times \text{Bern}(\theta_3)$,
$X|Y = +1 \sim \text{Bern}(\tau_1) \times \text{Bern}(\tau_2) \times \text{Bern}(\tau_3)$.
We have 10 training points from the above distribution, given by the table below.

X_1	X_2	X_3	Y
1	0	0	+1
0	1	1	−1
0	1	0	+1
1	1	0	+1
1	1	1	−1
1	0	0	+1
1	0	1	+1
0	0	1	−1
0	1	1	+1
0	0	0	−1

i. Give the ML estimates for $a, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3$.

ii. For all the 8 points X in the instance space $\{0, 1\}^3$, give the estimate of the posterior probability $\mathbf{P}(Y = +1|X)$, and give the prediction that minimises the misclassification rate (or the Bayes classifier for the zero-one loss), in the form of a table with 8 rows. **(3+3 Points)**

Ans.

Q6 Screenshot.zip

6.

Considering a distribution over (x, y) given by the following assumptions,

$y \in \{-1, +1\}$, $x \in \{0, 1\}^3$

$P(Y = +1) = a$

$P(Y = -1) = (1 - a)$

$x | y = -1$ follow $\text{Bern}(\theta_1) \times \text{Bern}(\theta_2) \times \text{Bern}(\theta_3)$

$x | y = +1$ follow $\text{Bern}(\tau_1) \times \text{Bern}(\tau_2) \times \text{Bern}(\tau_3)$

i) As per the given dataset we have 10 given points, & a quick observation of the data reveals that there is two points are repeating,

x_1	x_2	x_3	y
1	0	0	+1
1	0	0	+1
0	1	1	-1
0	1	1	+1

→ Here {100} set always giving output as $y = +1$ but there is some contradiction is {011} set.

a is the probability of $P(Y = +1)$ so,

$$\hat{a} = 6/10 = 3/5$$

$P(Y = +1) = 3/5$ & $P(Y = -1) = (1 - 3/5) = 2/5$

considering the assumption it is clear that θ_1, θ_2 & θ_3 will be corresponding to each input feature in feature space considering $y = -1$, so,

$$\hat{\theta}_1 = 0+1+0+0/4 = 1/4$$

$$\hat{\theta}_2 = 1+1+0+0/4 = 2/4 = 1/2$$

$$\hat{\theta}_3 = 1+1+1+0/4 = 3/4$$

Similarly according the assumption, τ_1, τ_2 & τ_3 are corresponding to each input feature in feature space where $y = +1$, so,

$$\hat{\tau}_1 = 1+0+1+1+1+0/6 = 4/6 = 2/3$$

$$\hat{\tau}_2 = 0+1+1+0+0+1/6 = 3/6 = 1/2$$

$$\hat{\tau}_3 = 0+0+0+0+1+1/6 = 2/6 = 1/3$$

So, below is the ML estimates,

$\hat{a} = 3/5$, $\hat{\theta}_1 = 1/4$, $\hat{\theta}_2 = 1/2$, $\hat{\theta}_3 = 3/4$

$\hat{\tau}_1 = 2/3$, $\hat{\tau}_2 = 1/2$, $\hat{\tau}_3 = 1/3$

ii) According to bayes classification formulae, we write,

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

where $P(Y|X)$ = posterior probability
$P(X|Y)$ = class conditional
$P(Y)$ = Prior probability
$P(X)$ = marginal probability

we have to calculate

$$P(Y=+1|X) = \eta(x)$$

$$\Rightarrow \eta(x) = P(Y=+1|X) = \frac{P(X=x|Y=+1)\,P(Y=+1)}{P(X=x)}$$

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

Now according to bernoulli we know that the PMF is,

$$f(x) = \begin{cases} p^x \cdot (1-p)^{1-x} & \text{if } x = 0,1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

Now we know the parameter for $P(Y=+1|X)$ which is set of x.

Now plotting the x value in the $\eta(x)$ equation we can found the derived result.

$$\eta(x) = \frac{\left[\gamma_1^{x_1} \cdot (1-\gamma_1)^{(1-x_1)}\right]\left[\gamma_2^{x_2}(1-\gamma_2)^{(1-x_2)}\right]\left[\gamma_3^{x_3}(1-\gamma_3)^{(1-x_3)}\right] * 3/5}{P(X=x|Y=+1)\cdot 3/5 + P(X=x|Y=-1)\cdot 2/5}$$

Now calculating class conditional part.,

$$\left[(2/3)^{x_1}(1/3)^{(1-x_1)}\right]\cdot\left[(1/2)^{x_2}\cdot(1/2)^{(1-x_2)}\right]\cdot\left[(1/3)^{x_3}(2/3)^{(1-x_3)}\right]$$

$$= \left[(2/3)^{x_1}\cdot(1/3)\cdot(1/3)^{-x_1}\right]\left[(1/2)^{x_2}\cdot(1/2)\cdot(1/2)^{-x_2}\right]\cdot\left[(1/3)^{x_3}\cdot(2/3)^{(-x_3+1)}\right]$$

$$= \frac{1}{9}\cdot 2^{(x_1-x_3)} \quad \text{--- (i)}$$

$$\therefore P(X=x|Y=+1) = 1/9 \cdot 2^{(x_1-x_3)} \quad \text{--- (i)}$$

$$P(X=x|Y=-1) = \left[(1/4)^{x_1}(3/4)^{(1-x_1)}\right]*\left[(1/2)^{x_2}\cdot(1/2)^{(1-x_2)}\right]*\left[(3/4)^{x_3}\cdot(1/4)^{(1-x_3)}\right]$$

$$= \left[(1/4)^{x_1}(3/4)(3/4)^{-x_1}\right]\left[(1/2)^{x_2}\cdot(1/2)(1/2)^{-x_2}\right]\left[(3/4)^{x_3}\cdot(1/4)\cdot(1/4)^{-x_3}\right]$$

$= \left(\frac{3}{32}\right) \cdot \frac{1}{3}^{(x_4 - x_3)}$ ... (ii)

So the class conditionals are,

$P(X = x \mid Y = +1) = \frac{1}{9} \cdot 2^{(x_4 - x_3)}$ ... (i)

$P(X = x \mid Y = -1) = \frac{3}{32} \cdot \left(\frac{1}{3}\right)^{(x_4 - x_3)}$ ... (iii)

Now by plotting all the values in the $\eta(x)$ equation we have got,

$$\eta(x) = \frac{2^{(x_4 - x_3)} \times \frac{3}{8}}{\frac{2^{(x_4 - x_3)}}{3} \times \frac{3}{8} + \frac{3}{32} \cdot \left(\frac{1}{3}\right)^{(x_4 - x_3)} \cdot \frac{3}{4}}$$

$$= \frac{\frac{2}{3}^{(x_4 - x_3)}}{\frac{2^{(x_4 - x_3)}}{3} + \frac{3}{16} \cdot \left(\frac{1}{3}\right)^{(x_4 - x_3)}}$$

$$= \frac{1}{1 + \frac{9}{16} \cdot \frac{1}{6^{(x_4 - x_3)}}} \quad \text{... (iii)}$$

As this is bernoulli distribution & it can take only two possible value we can threshold at $1/2$ for finding our positive & negative result.

∴ So, we can write our,

$h^*(x) = +1$ if $\eta(x) > 1/2$

$= -1$ if $\eta(x) < 1/2$

As explained in first question, there are one set which is giving same output so we will remove it according to our calculated assumption & one point for $\{0, 1, 1\}$ set is mis classified which we have to correctly classified. So for that four data point we will take 2 data point & finally for given 10 data point we will calculate 8 data point.

(iii) $\eta(x)$ is calculated by plotting x_1, x_2 & x_3 value in eq nation

for example

$\{0, 0, 0\}$ point $\rightarrow \eta(x) = \frac{1}{1 + \frac{9}{16} \cdot \frac{1}{6^0}} = \frac{1}{1 + 9/16}$

$= 16/25$

$= 0.64$

Similar calculation has been used for all point

Now, doing prediction based on calculation of $\eta(x)$.

x_1	x_2	x_3	$\eta(x)$	$h^*(x)$	y	
0	0	0	0.64	+1	−1	→ Misclassified
0	0	1	0.228	−1	−1	
0	1	0	0.64	+1	+1	
0	1	1	0.228	−1	−1 & +1	
1	0	0	0.914	+1	+1	
1	0	1	0.64	+1	+1	
1	1	0	0.914	+1	+1	
1	1	1	0.64	+1	−1	→ Misclassified

As per the above table we can conclude that,

(i) {0,0,0} is classified as +1 but actual is y=−1

(ii) {1,1,1} is classified as +1 but actual is y=−1

→ Misclassified

(iii) {0,1,1} point is classified as −1 in data both +1 & −1 is given.

(iv) {1,0,0} point is classified as +1, as per given data

(v) Rest all data point's are correctly classified.

—x—