Problem 2. Gradient Descent Algorithm and Logistic Regression (40 points):

1. In logistic regression method, please derive the derivative of the negative logarithm of the likelihood function with respect to parameter w. You need to show the detailed steps to obtain the following results.

$$\nabla w\ \varepsilon(w) = \Sigma\ (f(xn\ ) - yn)xn$$

(*) Answer : 1

⇒ General formula for logistic regression is,

$$p(Y = y \mid X = x) = \sigma(w^T x)^y \left[1 - \sigma(w^T x)\right]^{(1-y)}$$

⇒ Log-Likelihood function for the above equation,

$$LL(w) = \log \prod_{i=1}^{N} P(Y = y_i \mid X = x_i)$$

$$= \log \prod_{i=1}^{N} \sigma(w^T x_i)^{y_i} \left[1 - \sigma(w^T x_i)\right]^{(1-y_i)}$$

$$= \sum_{i=1}^{N} \left[y_i \log \sigma(w^T x_i) + (1 - y_i) \log\left[1 - \sigma(w^T x_i)\right] \right]$$

⇒ Now, calculating derivative of logistic sigmoid function,

$$\frac{\partial}{\partial a} \sigma(a) = \frac{\partial}{\partial a} \frac{1}{1+e^{-a}} = \frac{e^{-a}}{(1+e^{-a})^2}$$

$$= \frac{e^{-a}}{1+e^{-a}} \times \frac{1}{1+e^{-a}}$$

$$= \frac{1}{1+e^{-a}} \left(1 - \frac{1}{1+e^{-a}}\right) = \sigma(a)\left[1 - \sigma(a)\right] \quad \text{— (1)}$$

⇒ Now, the next step is to calculate derivative of log-likelihood function with respect to 'w'

$$\therefore \frac{\partial LL(w)}{\partial w} = \frac{\partial}{\partial w} \sum_{i=1}^{N} \left[y_i \log \sigma(w^T x_i) + (1-y_i)[1-\sigma(w^T x_i)] \right]$$

$$\boxed{= \sum_{i=1}^{N} \left[\frac{y_i}{\sigma(w^T x_i)} - \frac{1-y_i}{1-\sigma(w^T x_i)} \right] \frac{\partial}{\partial w} \sigma(w^T x_i)}$$

$$= \sum_{i=1}^{N} \left[\frac{y_i}{\sigma(w^T x_i)} - \frac{1-y_i}{1-\sigma(w^T x_i)} \right] \sigma(w^T x_i)[1-\sigma(w^T x_i)] x_i$$

(from ①)

$$= \sum_{i=1}^{N} \left[\frac{y_i - \sigma(w^T x_i)}{\sigma(w^T x_i)[1-\sigma(w^T x_i)]} \right] \sigma(w^T x_i)[1-\sigma(w^T x_i)] x_i$$

$$= \sum_{i=1}^{N} [y_i - \sigma(w^T x_i)] x_i$$

$$\therefore \frac{\partial LL(w)}{\partial w} = \sum_{i=1}^{N} [y_i - f(x)] x_i \quad \text{— ②}$$

⇒ Now,
The cost function for logistic regression can be derived by minimizing the log-likelihood function,

Hence,

$$\operatorname{argmax} \mathcal{E}(w) = -\operatorname{argmin} LL(w)$$

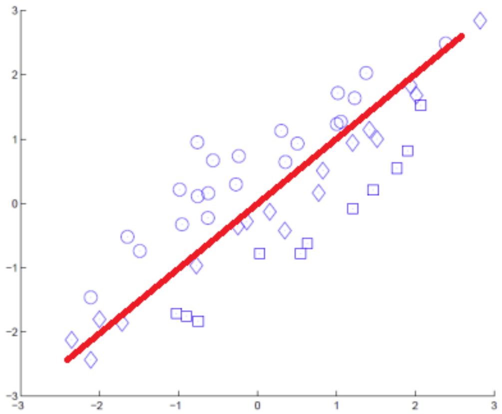
$$\therefore \frac{\partial}{\partial w} \mathcal{E}(w) = -\frac{\partial}{\partial w} LL(w)$$

$$\therefore \boxed{\frac{\partial}{\partial w} \mathcal{E}(w) = \sum_{i=1}^{N} [f(x_i) - y_i] x_i}$$ (from ②)

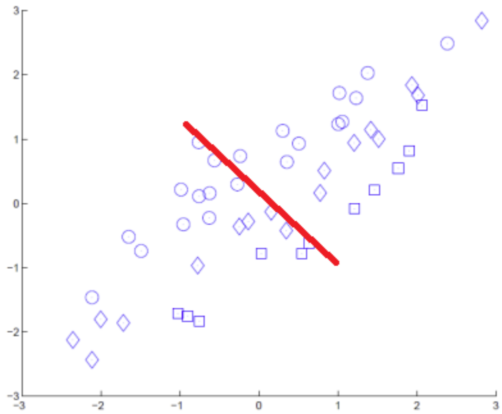
Problem 3: Principal Component Analysis (PCA) (20 points):

1. Given labels of the data, the goal of Fisher's Linear Discriminant is to find the projection direction that maximizes the ratio of between-class variance and the within-class variance. While PCA aims to reduce the dimension of the data by finding projection directions that maximizes the variance after projection. Note that PCA does not consider the label information. In the following figures, consider round points as positive class, and both diamond and square points as negative class. Please draw (a) the direction of the first principal component in the left figure by ignoring the label of the data points, and (b) the Fisher's linear discriminant direction in the right figure. Please draw a line to show the direction for each of them.

I have used Paint in order to plot the line in the graph.



1(a) First PCA component



1(b) First LDA component

2. Consider 3 data points in the 2D space: (2,2), (0,0), (-2,-2). Please answer the following questions.

   a) Calculate the first principal component by calculating the eigenvalue (non-zero) and eigenvector of the covariance matrix. You need to provide the actual vector of the first principal component (with length=1). You can use the unbiased estimation of the covariance

Answer 2 (a) :

x	y
2	2
0	0
-2	-2

mean for X, $\Rightarrow \frac{1}{3}(2+0-2) = 0$

mean for Y, $\Rightarrow \frac{1}{3}(2+0-2) = 0$

variance for X, $\Rightarrow \frac{1}{3-1}\left[(2-0)^2 + (0-0)^2 + (-2-0)^2\right]$

$\Rightarrow 4$

similarly, variance for Y, $\Rightarrow 4$

Now,

	x	y
μ	0	0
σ	2	2

Now, standardizing each point by doing $x_{new} = x - \mu$ we get,

x_{new}	y_{new}
2	2
0	0
-2	-2

	x_{new}	y_{new}
μ	0	0
var	4	4

⇒ Now, the covariance matrix is,

	x_{new}	y_{new}
x_{new}	var(x_{new})	cov(x_{new}, y_{new})
y_{new}	cov(x_{new}, y_{new})	var(y_{new})

$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$$

$$\text{cov}(x_{new}, y_{new}) = \frac{1}{3-1}\left[(2-0)(2-0) + (0-0)(0-0) + (-2-0)(-2-0)\right]$$

$$= \frac{1}{2}\left[4 + 0 + 4\right] = \boxed{4}$$

putting this value in covariance matrix, we get,

eigen value

covariance matrix

	xnew	ynew
xnew	4	4
ynew	4	4

⇒ Now the value of $A - \lambda I$ is,

$$A - \lambda I = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4-\lambda & 4 \\ 4 & 4-\lambda \end{bmatrix}$$

⇒ To find eigen values, $\det(A - \lambda I)$ will be equal to 0.

$\therefore \det(A - \lambda I) = 0$

$\therefore \begin{vmatrix} 4-\lambda & 4 \\ 4 & 4-\lambda \end{vmatrix} = 0$

$\therefore (4-\lambda)(4-\lambda) - 16 = 0$

$\therefore 16 - 8\lambda + \lambda^2 - 16 = 0$

$\therefore \lambda^2 - 8\lambda = 0$

$\therefore \boxed{\lambda = 0}$ or $\boxed{\lambda = 8}$

eigen value

⇒ The eigenvector for first eigenvalue $\lambda = 0$ is,

$$\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\therefore \begin{bmatrix} 4x + 4y \\ 4x + 4y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

One solution for above equation is,

$\boxed{x = 1}$ and $\boxed{y = -1}$ ⇒ eigen vector

⇒ The eigen vector for second eigen value $\lambda = 8$ is

$$\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$$

$$\therefore \begin{bmatrix} 4x + 4y \\ 4x + 4y \end{bmatrix} = \begin{bmatrix} 8 \\ 8 \end{bmatrix}$$

One solution for above equation is,

$\boxed{x = 1}$ and $\boxed{y = 1}$ ⇒ eigen vector

a) If we project the three data points into the 1D subspace by the principal component obtained in (a), what are the new coordinates of the three data points in the 1D subspace? What is the variance of the data after projection?

Answer 2(b):

⇒ The eigen matrix from ~~also~~ previous solution is,

$$\begin{array}{c} \lambda = 0 \\ \lambda = 8 \end{array} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

⇒ Now, the transformed dataset will be,

= (original matrix) * (top k values of eigen matrix)

$$= \begin{bmatrix} 2 & 2 \\ 0 & 0 \\ -2 & -2 \end{bmatrix}_{3\times 2} * \begin{bmatrix} 1 \\ 1 \end{bmatrix}_{2\times 1}$$

$$= \begin{bmatrix} 4 \\ 0 \\ -4 \end{bmatrix}$$

⇒ So, the new coordinates of the three data points in 1 - D space will be,

$$\boxed{4, \quad 0 \quad \text{and} \quad -4}$$

⇒ Now,

the total variance of above data is,

$$var(x) = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - (\bar{x}))^2$$

$$= \frac{1}{3-1} \left[(4-0)^2 + (0-0)^2 + (-4-0)^2 \right] \quad \left\{ \begin{array}{l} \bar{x} = \frac{1}{3}(4+0-4) \\ = 0 \end{array} \right.$$

$$\boxed{var(x) = 16}$$

c) What is the cumulative explained variance of the first principal component? Is there any variance that is not captured by it?

Answer 2(c) :

⇒ Eigen value corosponding to first principal component will be 8.

So, cumulative explained variance of the first principal component

$$= \% \text{ total variance}$$

$$= 8/8+0 = \boxed{100\ \%}$$

⇒ Since, we got 100 % of total variance captured by pricipal component first, there is nothing left to capture.

⇒ Hence, answer for "Is there any variance that is not captured by it ?" will be $\boxed{\text{NO}}$

Problem 4: Support Vector Machines (20 points):

Given 10 points in Table 1, along with their classes and their Lagranian multipliers (αi), answer the following questions:

Data	X_{i1}	X_{i1}	Y	α_i
X_1	4	2.9	1	0.414
X_2	4	4	1	0
X_3	1	2.5	-1	0
X_4	2.5	1	-1	0.018
X_5	4.9	4.5	1	0
X_6	1.9	1.9	-1	0
X_7	3.5	4	1	0.018
X_8	0.5	1.5	-1	0
X_9	2	2.1	-1	0.414
X_{10}	4.5	2.5	1	0

1) What is the equation of the SVM hyperplane $h(x)$? Draw the hyperplane with the 10 points.

Answer 4 (1) :

⇒ From the given table we can say that, there are four points counted as a support vectors (x_1 , x_4 , x_7 , x_9). because, every other vectors have value of $\alpha_i = 0$.

⇒ The weight factor of hyperplane is,

$$w = \sum_{i,\alpha>0} \alpha_i y_i$$

$$= 0.414\begin{pmatrix}4\\2.9\end{pmatrix} - 0.018\begin{pmatrix}2.5\\1\end{pmatrix} + 0.018\begin{pmatrix}3.5\\4\end{pmatrix} - 0.414\begin{pmatrix}2\\2.1\end{pmatrix}$$

$$= \begin{pmatrix}0.846\\0.385\end{pmatrix}$$

⇒ Now we can compute final bias as the average of the bias obtained from each support vector using,

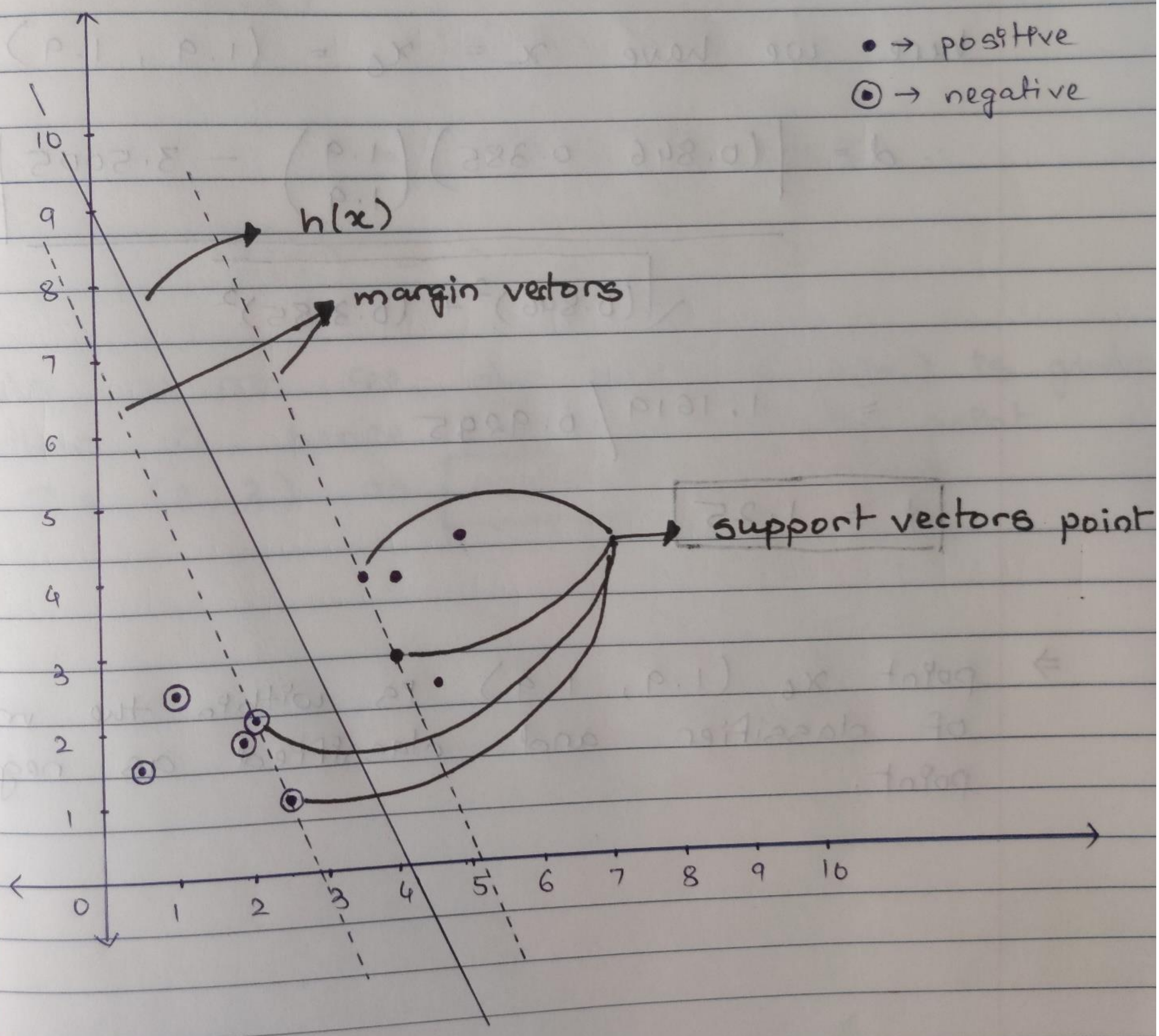
$$b_i = y_i - w^T x_i$$

x_i	$w^T x_i$	$b_i = y_i - w^T x_i$
x_1	4.5005	−3.5005
x_4	2.5	−3.5
x_7	4.501	−3.501
x_9	2.5005	−3.5005
		avg b_i = −3.5005

⇒ Thus, the equation for hyperplane is,

$$\boxed{h(x) = \begin{pmatrix} 0.846 \\ 0.385 \end{pmatrix} x - 3.5005 = 0}$$

⇒ Now, ploting the hyperplane with all points



2) What is the distance of x_6 from the hyperplane? Is it within the margin of the classifier?

Answer 4(2):

⇒ The distance between any point to the hyperplane is given by,

$$d = \frac{|\vec{w} \cdot \vec{x} + b|}{\|w\|}$$

here we have $x = x_6 = (1.9, 1.9)$

$$\therefore d = \frac{\left| \begin{pmatrix} 0.846 & 0.385 \end{pmatrix} \begin{pmatrix} 1.9 \\ 1.9 \end{pmatrix} - 3.5005 \right|}{\sqrt{(0.846)^2 + (0.385)^2}}$$

$$= 1.1619 / 0.9295$$

$$\boxed{d \approx 1.25}$$

⇒ point x_6 (1.9, 1.9) is within the margin of classifier and classified as negative point.

3) Classify the point $z = (3,3)^T$ using $h(x)$ from above.

(*) Answer 4(3):

⇒ To classify the point $z = (3, 3)$,

let's compute the value of $h(x)$ for the given input point,

$$\therefore \quad h(x) = \begin{pmatrix} 0.846 \\ 0.385 \end{pmatrix}^T \begin{pmatrix} 3 \\ 3 \end{pmatrix} - 3.5005$$

$$= 3.693 - 3.5005$$

$$h(x) = 0.1925 > 0.$$

⇒ As we can see, the value of $h(x)$ is greater than 0, hence we can classify point $z = (3, 3)$ as $\boxed{\text{positive}}$ with $y = 1$.