Homework Assignment 2

Problem 1:

For Iris data set from the UCI Machine Learning repository, we split the data set into three sets for each pair of class labels. We generated three classification models using the training data for each pair and calculated the mean accuracy for each model using the test data. Below is the comparison of the model accuracy for each pair.

Model	Mean Accuracy
For class labels 1 and 2 (Iris-setosa, Iris-versicolor)	100%
For class labels 1 and 3 (Iris-setosa, Iris-virginica)	100%
For class labels 2 and 3 (Iris-versicolor, Iris-virginica)	90%

As we can see from the results above, the model for class labels 1 and 2 (Iris-setosa, Iris-versicolor) and model for class labels 1 and 3 (Iris-setosa, Iris-virginica) have a mean accuracy of 100% while the model for class labels 2 and 3 (Iris-versicolor, Iris-virginica) has a mean accuracy of 90%.

Homework Assignment 2

Problem 2:

$$P(C_1|x) = \sigma(w^T x + w_0) = f(x)$$

The likelihood function for logistic regression:

$$L(w) = \prod_{n=1}^{N} p(C_1|x_n)^{y_n}(1-p(C_1|x_n))^{1-y_n}$$

$$= \prod_{n=1}^{N} f(x_n)^{y_n}(1-f(x_n))^{1-y_n}$$

The negative logarithm of the likelihood (Cross Entropy)

$$\mathcal{E}(w) = -\log(L(w)) = -\log\left(\prod_{n=1}^{N} p(C_1|x_n)^{y_n}(1-p(C_1|x_n))^{1-y_n}\right)$$

$$= -\sum_{n=1}^{N}\left[\log\left(p(C_1|x_n)^{y_n}(1-p(C_1|x_n))^{1-y_n}\right)\right]$$

$$= -\sum_{n=1}^{N}\left[\log(p(C_1|x_n))^{y_n} + \log(1-p(C_1|x_n))^{1-y_n}\right]$$

$$= -\sum_{n=1}^{N}\left[y_n \log(p(C_1|x_n)) + (1-y_n)\log(1-p(C_1|x_n))\right]$$

$$= -\sum_{n=1}^{N}\left[y_n \log f(x_n) + (1-y_n)\log(1-f(x_n))\right]$$

$$= -\sum_{n=1}^{N}\left[y_n \log(\sigma(w^T x + w_0)) + (1-y_n)\log(1-\sigma(w^T x + w_0))\right]$$

$$= -\sum_{n=1}^{N}\left[y_n \log\sigma(a) + (1-y_n)\log(1-\sigma(a))\right]$$

where $a = w^T x + w_0$

The derivative of the negative logarithm of likelihood function w.r.t w.

$$\frac{\partial}{\partial w} \mathcal{E}(w) = \frac{\partial}{\partial w} \left[-\log \prod_{n=1}^{N} p(C_1|x_n)^{y_n} (1-p(C_1|x_n))^{1-y_n}\right]$$

The derivative of the logistic sigmoid function:

$$\frac{\partial}{\partial w} \sigma(a) = \frac{\partial}{\partial a} \sigma(a) \cdot \frac{\partial(a)}{\partial w}$$

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

$$= \frac{e^{-a}}{(1+e^{-a})^2} \qquad (\because \text{Using the quotient rule})$$

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

$$= \frac{1}{1+e^{-a}} \left(1 - \frac{1}{1+e^{-a}}\right)$$

$$= \sigma(a)(1-\sigma(a))$$

$$\frac{\partial(a)}{\partial w} = \frac{\partial}{\partial w} (w^T x + w_0)$$

$$= \frac{\partial}{\partial w} w^T x + \frac{\partial w_0}{\partial w}$$

$$= x$$

$$\frac{\partial}{\partial w}\sigma(a) = \sigma(a)(1-\sigma(a))(x)$$

Now,

$$\frac{\partial}{\partial w}\mathcal{E}(w) = \frac{\partial}{\partial w}\left[-\sum_{n=1}^{N}\left[y_n \log\sigma(a) + (1-y_n)\log(1-\sigma(a))\right]\right]$$

$$= -\sum_{n=1}^{N}\left[\frac{\partial}{\partial w} y_n \log\sigma(a) + \frac{\partial}{\partial w}(1-y_n)\log(1-\sigma(a))\right]$$

$$= -\sum_{n=1}^{N}\left[\frac{\partial}{\partial w}(y_n).\log\sigma(a) + y_n.\frac{\partial}{\partial w}\log\sigma(a) + \frac{\partial}{\partial w}(1-y_n).\log(1-\sigma(a)) + (1-y_n)\frac{\partial}{\partial w}\log(1-\sigma(a))\right]$$

$$= -\sum_{n=1}^{N}\left[\frac{y_n}{\sigma(a)}\frac{\partial}{\partial w}\sigma(a) + \frac{(1-y_n)}{1-\sigma(a)}\frac{\partial}{\partial w}(1-\sigma(a))\right]$$

$$= -\sum_{n=1}^{N}\left[\frac{y_n}{\sigma(a)}\left[\sigma(a)(1-\sigma(a))(x)\right] + \frac{1-y_n}{1-\sigma(a)}\left[-\sigma(a)(1-\sigma(a))(x)\right]\right]$$

$$= -\sum_{n=1}^{N}\left[y_n(1-\sigma(a))x + (1-y_n)(-\sigma(a)x)\right]$$

$$= -\sum_{n=1}^{N}\left[y_n - y_n\sigma(a) + (-\sigma(a)) + (y_n\sigma(a))\right]x_n$$

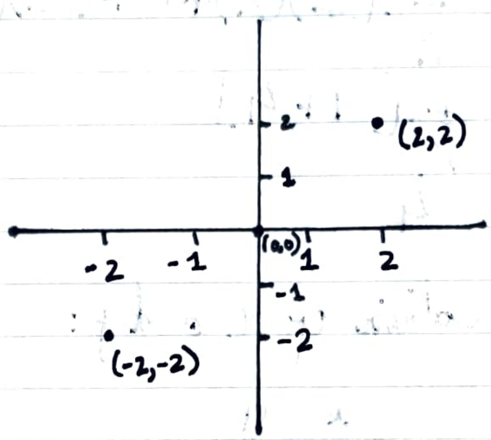
$$= \sum_{n=1}^{N}\left[\sigma(a) - y_n\right]x_n$$

$$= \sum_{n=1}^{N}(f(x_n) - y_n)x_n$$

Problem 3:

2) a)

$$A = \begin{array}{cc} x & y \\ \end{array}$$
$$\begin{bmatrix} 2 & 2 \\ 0 & 0 \\ -2 & -2 \end{bmatrix}$$



To find the covariance matrix, we calculate the mean of data points matrix A

$$A = \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} 2 & 2 \\ 0 & 0 \\ -2 & -2 \end{bmatrix} \end{array}$$

$n = 3$, Mean $= \bar{X} = 0$ $\bar{Y} = 0$

$$Var(X) = \frac{1}{N-1} \sum_{n=1}^{N} (X_n - \bar{X})^2$$

$$= \frac{1}{2} \left[(2)^2 + 0 + (-2)^2 \right]$$

$$= \frac{1}{2} [8] = 4$$

$Var(Y) = Var(X) = 4$

$$Cov(X,Y) = \frac{1}{N-1} \sum_{n=1}^{N} (X_n - \bar{X})(Y_n - \bar{Y})$$

$$= \frac{1}{2}[(2)(2) + 0 + (-2)(-2)]$$

$$= \frac{1}{2}[4+4]$$

$$= 4$$

The covariance matrix of x and y:

$$\Sigma = Cov(X,Y) = \begin{array}{cc} & \begin{array}{cc} x & y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \end{array}$$

The solution w is the eigenvector of Σ corresponding to the largest eigenvalue λ:

$$\Sigma w = \lambda w$$

$$\det(\Sigma - \lambda I) = \begin{vmatrix} 4-\lambda & 4 \\ 4 & 4-\lambda \end{vmatrix}$$

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

$$= 16 - 4\lambda - 4\lambda + \lambda^2 - 16 = 0$$

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

$$= \lambda(\lambda - 8) = 0$$

$\lambda = 0$ and $\lambda = 8$

The eigen vector for the first eigenvalue $\lambda = 0$ is

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

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

$$4x + 4y = 0$$

$$x = -y$$

One solution for both equations is $x = 1$ and $y = -1$

w is an eigenvector with eigenvalue 0 if

$$\Sigma w = \lambda w$$

$$(\Sigma - \lambda I) w = 0$$

$$\left(\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} - 0 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

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

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

$$0 = 0$$

Hence, the eigen-vector for eigen-value 0 is

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

The eigen-vector for the first eigen-value $\lambda = 8$ is:

$$\begin{bmatrix} 4-8 & 4 \\ 4 & 4-8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

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

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

$$-4x+4y = 0$$
$$x = y$$

One solution for both equation is $x = 1$ and $y = 1$.

w is an eigen-vector with eigen-value 8 if

$$(\Sigma - \lambda I)\, w = 0$$

$$\left(\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} - 8 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

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

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

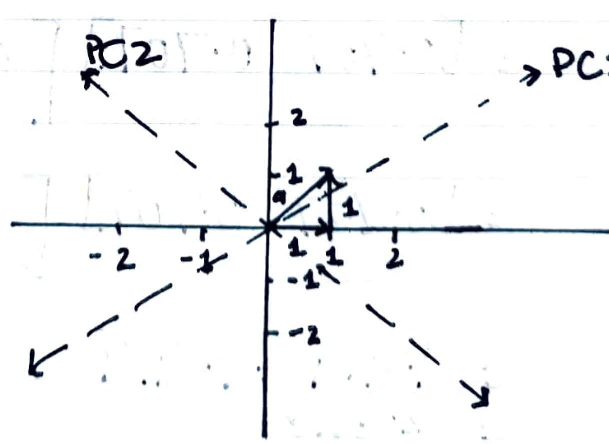
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Hence, the eigen vector for eigen value 8 is

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

The eigen vector associated with the largest eigenvalue corresponds to the first principal component and the eigen vector associated with the second largest eigenvalue corresponds to the second principle component.

Hence, the eigen vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ associated with the largest eigenvalue $\lambda = 8$ corresponds to the first principle component and the eigen vector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ ~~cor~~ associated with the 2nd largest eigen value $\lambda = 0$ corresponds to the second principle component.



The actual vector of the first principal component (with length = 1) can be found using the Pythagoras theorem

$$a^2 = b^2 + c^2$$
$$a^2 = 1^2 + 1^2$$
$$a^2 = 2$$
$$a = \sqrt{2}$$
$$a = 1.414$$

After normalizing the eigen-vector with the above value, we get the actual vector of the first principal component.

$$e_1 = \begin{bmatrix} 1/1.414 \\ 1/1.414 \end{bmatrix} = \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}$$

b) For projecting the three data points: (2,2), (0,0), (-2,-2) into 1D subspace, we need to multiply the data points with the first principle component using the below formula

$$P_{(2,2)} = e_1^T \begin{bmatrix} x_1 - \bar{X} \\ y_1 - \bar{Y} \end{bmatrix}$$

$$= \begin{bmatrix} 0.707 & 0.707 \end{bmatrix} \begin{bmatrix} 2 - 0 \\ 2 - 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.707 & 0.707 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$= 1.414 + 1.414$$
$$= 2.828$$

Similarly, $P_{(0,0)} = \begin{bmatrix} 0.707 & 0.707 \end{bmatrix} \begin{bmatrix} 0-0 \\ 0-0 \end{bmatrix}$

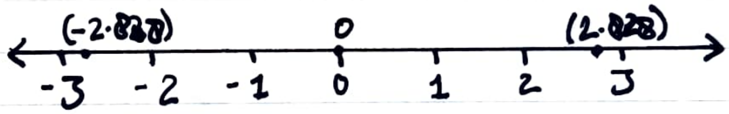
$$= 0$$

$$P_{(-2,-2)} = \begin{bmatrix} 0.707 & 0.707 \end{bmatrix} \begin{bmatrix} -2-0 \\ -2-0 \end{bmatrix}$$

$$= -1.414 - 1.414$$

$$= -2.828$$

Hence, the new data points into the 1D subspace by the first principle component are $(2.828, 0, -2.828)$, ~~(0,0)~~, ~~(-2.828~~



$$\text{Variance of the data} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2$$

$$= \frac{1}{2} \left[(2.828)^2 + 0 + (-2.828)^2 \right]$$

$$= \frac{1}{2} [8 + 8]$$

$$Var(X) = 8$$

c) The cumulative explained variance is used to get the ratio of variance

$$\text{Cumulative Explained Variance} = \frac{\text{Eigen-value}}{\text{Total eigenvalues}}$$

$$\text{Cumulative explained variance for the first principal component} = \frac{\lambda_1}{\lambda_1+\lambda_2}$$

$$= \frac{8}{8+0}$$

$$= 1$$

This shows that the first principal component captures the complete variance.

Problem 4:

1) The solution vector w is

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

for all x_i for which $\alpha_i > 0$.

The equation of SVM hyperplane $h(x)$ is given as

$$h(x) = w^T x + b$$

For w,
$$\begin{aligned} w_1 &= [(4)(1)(0.414) + (0.018)(-1)(2.5) \\ &\quad + (0.018)(1)(3.5) + (0.414)(2)(-1)] \\ &= 1.656 - 0.045 + 0.063 - 0.828 \\ &= 0.846 \end{aligned}$$

$$\begin{aligned} w_2 &= [(0.414)(1)(2.9) + (0.018)(-1)(1) \\ &\quad + (0.018)(1)(4) + (0.414)(-1)(2.1)] \\ &= 1.2006 - 0.018 + 0.072 - 0.8694 \\ &= 0.3852 \end{aligned}$$

To obtain the value of b, solve $\alpha_i[y_i(w^T x_i + b) - 1] = 0$ for any support vector.

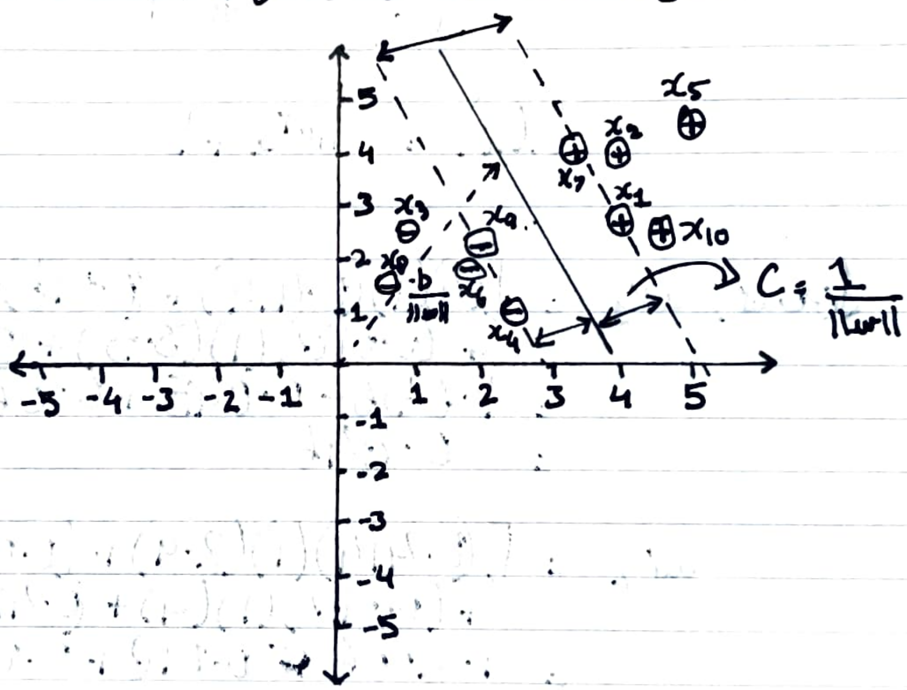
$$w = \begin{bmatrix} 0.846 \\ 0.3852 \end{bmatrix}$$

For x_1, $(0.414)\left[1\left(\begin{bmatrix}0.846 & 0.3852\end{bmatrix}\begin{bmatrix}4\\2.9\end{bmatrix}+b\right)-1\right]=0$

$$3.384 + 1.11708 + b = 1$$
$$b + 4.5 = 1$$
$$b = 1 - 4.5$$
$$b = -3.5$$

Hence the equation of the SVM hyperplane $h(x)$:

$$h(x) = \begin{bmatrix}0.846 & 0.3852\end{bmatrix} x + -3.5$$



Distance of hyperplane $w^T x + b = 0$ to origin:

$$\frac{-b}{\|w\|} = \frac{3.5}{\sqrt{(0.846)^2 + (0.3852)^2}}$$
$$= \frac{3.5}{\sqrt{0.864}} = \frac{3.5}{0.930}$$
$$= 3.76$$

2) The absolute distance of point x to hyperplane $w^Tx+b=0$:

$$\frac{|w^Tx+b|}{||w||}$$

For x_6,
$$d = \frac{\left|\begin{bmatrix}0.846 & 0.3852\end{bmatrix}\begin{bmatrix}1.9\\1.9\end{bmatrix} - 3.5\right|}{\sqrt{(0.846)^2+(0.3852)^2}}$$

$$= \frac{|1.607+0.7318-3.5|}{\sqrt{0.864}}$$

$$= \frac{1.16}{0.930}$$

$$= 1.24$$

$$C = \frac{1}{||w||}$$

$$= \frac{1}{0.930}$$

$$= 1.075$$

Since, the distance of point x_6 is greater than C, the point x_6 lies outside the margin of the classifier.

3)

$$h(x) = [0.846 \quad 0.3852] X - 3.5$$

For $z = (3,3)^T$,

$$h(x) = [0.846 \quad 0.3852] \begin{bmatrix} 3 \\ 3 \end{bmatrix} - 3.5$$
$$= 2.538 + 1.1556 - 3.5$$
$$= 0.1936$$

Since $h(x) > 0$, the point $z(3,3)$ lies on the positive side of the hyperplane and would have the label $y = 1$.

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.