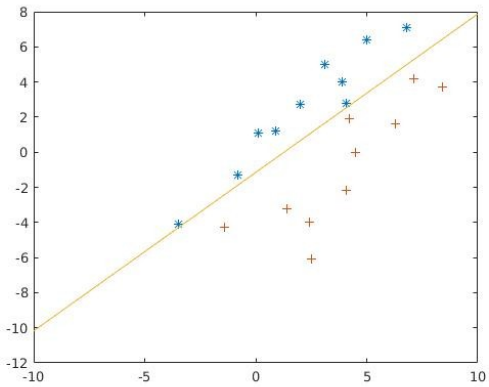
SMAI
ASSIGNMENT-2

Manan Bhandari
201431166

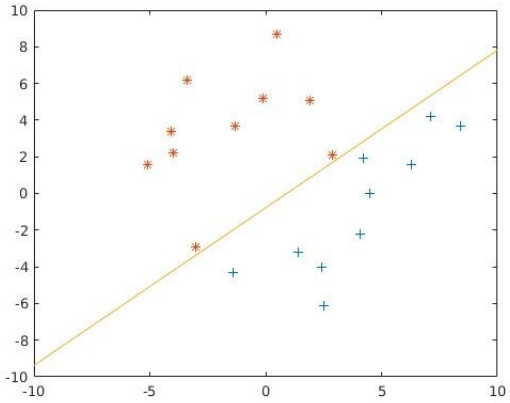
Q.1) Perceptron Algorithm:

1. ***Final Classifier with the data points of C1 and C2 being separated***



8 iterations for classifying C1 (labelled as +1) and C2 (labelled as -1). The above plot is the line with the classifier $w = [-10.2\ 11.3\ 13]^T$

2. ***Final Classifier with the data points of C2 and C3 being separated***

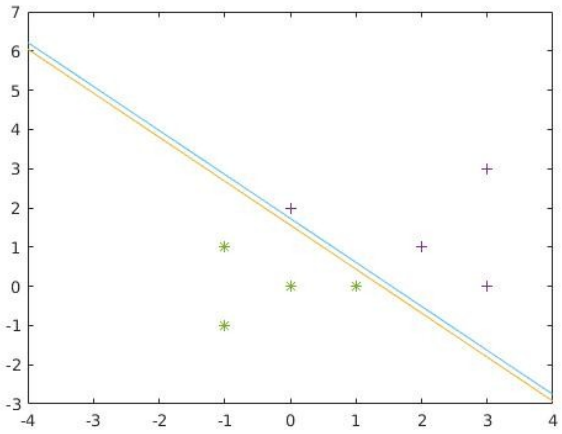


4 iterations for classifying C2 (labelled as +1) and C3 (labelled as -1). The above plot is the line with the classifier $w = [-5.50\ 6.40\ 5]^T$

3. The difference on the number of iteration required for convergence in the above two cases is that C3 data points are more widely spread and comparatively away from the line classifier and C2 data points whereas the C1 points are very close to each other, and to C2 data points and also closer to the classifier. So C1, C2 will require more number of iterations to arrive at an approximately correct classifier than C3, C2.

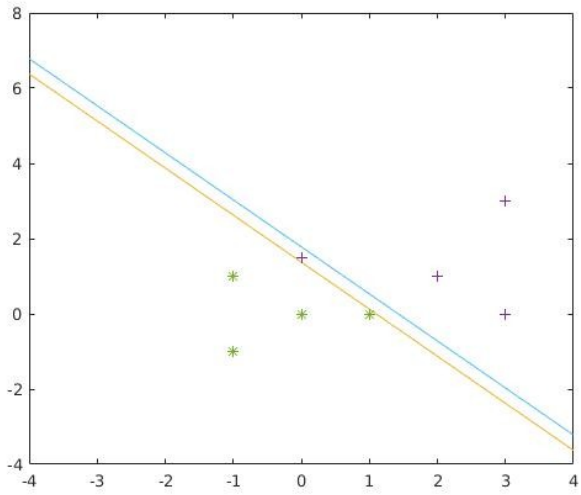
Q.3) Least Square Approach vs Fischer LDA

Classifier with the data points of C1 and C2



Linear Classifier from Least Square, w = [0.3750 0.3342 -0.5788]
Linear Classifier from Fischer LDA, w = [0.2863 0.2552 -0.3983]

Classifier with the data points of C2 and C3



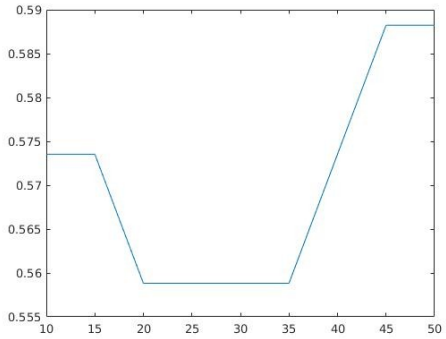
Linear Classifier from Least Square, w = [0.3778 0.3021 -0.5383]
Linear Classifier from Fischer LDA, w = [-0.2571 -0.2056 0.2828]

E.) Fisher linear discriminant is the projection that best separates the data in a least-squares sense. In fisher linear we got the classifier after finding direction of projection but in least square error we can directly get classifier.

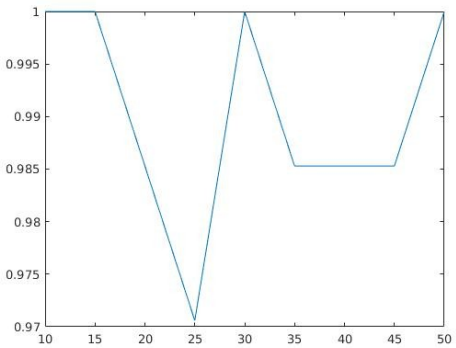
Q.2) 1. Breast Cancer Dataset (9 features)

10-fold Cross validation accuracy for each epoch value:

Voted Perceptron



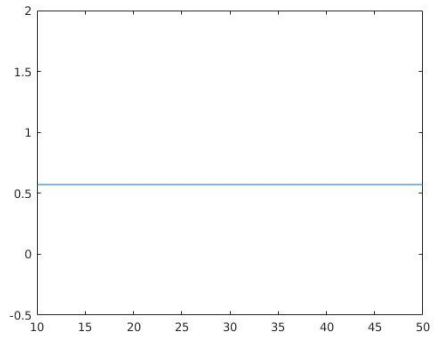
Vanilla Perceptron



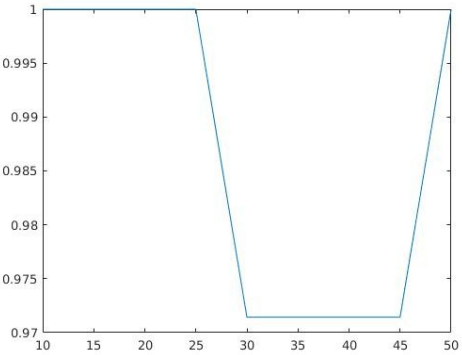
2. Ionosphere Dataset (34 features)

10-fold Cross validation accuracy for each epoch value:

Voted Perceptron



Vanilla Perceptron



d. Simply, first, for the voted perceptron, the classification computational cost is high and also that it requires greater storage space. In general, the voted perceptron perceptron is giving a better cross validation accuracy for these datasets when compared to vanilla perceptron but again there cannot be a strong conclusion.

Q.4)

4) Given: $S = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\}$

$$y_i = \begin{cases} m/m_1 & \text{if } x_i \in C_1 \\ -m/m_2 & \text{if } x_i \in C_2 \end{cases}$$

Show: linear classifier learnt using least square is same as the fischer's linear discriminant.

$$E = \sum_{i=1}^{m} (f(x) - y)^2 \qquad \text{(Least Square Method)}$$

1) $\frac{dE}{db} = 0 \Rightarrow \frac{d}{db}\left(\sum_{i=1}^{m} (w^T x_i + b - y_i)^2\right)$

$$\Rightarrow \sum (w^T x_i + b - y_i) = 0$$

$$w^T(x_1 + x_2 \ldots + x_m) + bm - (y_1 + y_2 \ldots + y_m) = 0$$

$$\frac{w^T(x_1 + x_2 \ldots + x_m)}{m} + b - \frac{[y_1 + y_2 + \ldots y_m)}{m} = 0$$

$$w^T \mu + b - \left(\frac{m_1 \times \frac{m}{m_1} + m_2 \times \frac{-m}{m_2}}{m}\right) = 0$$

$$w^T \mu + b - \frac{m - m}{m} = 0$$

$$\Rightarrow b = -w^T \mu$$

2) $\frac{dE}{dw} = 0 \Rightarrow \sum_{i=1}^{m} (w^T x_i + b - y_i) x_i = 0$

Substituting b from ① to splitting into different classes:

$$\Rightarrow \sum_{i=1}^{m_1} \left(w^T x_i - w^T \mu - \frac{m}{m_1}\right) x_i + \sum_{i=1}^{m_2} \left(w^T x_i - w^T \mu + \frac{m}{m_2}\right) x_i = 0$$

$$\Rightarrow \sum_{i=1}^{m_1} w^T (x_i - \mu) x_i - \sum_{i=1}^{m_1} \frac{m}{m_1} x_i + \sum_{i=1}^{m_2} w^T (x_i - \mu) x_i + \sum_{i=1}^{m_2} \frac{m}{m_2} x_i = 0$$

$$\Rightarrow \sum_{i=1}^{m_1} w^T (x_i - \mu) x_i + \sum_{i=1}^{m_2} w^T (x_i - \mu) x_i - \frac{m}{m_1} \sum_{i=1}^{m_1} x_i + \frac{m}{m_2} \sum_{i=1}^{m_2} x_i = 0$$

$$\Rightarrow \sum_{i=1}^{m} w^T (x_i - \mu) x_i = m(\mu_1 - \mu_2)$$

$$\Rightarrow \sum_{i=1}^{m} w^T (x_i - \mu)(x_i - \mu + \mu) = m(\mu_1 - \mu_2)$$

$$\Rightarrow w\Sigma(x_i-\mu)(x_i-\mu)^T + w^T\mu\Sigma(x_i-\mu) = m(\mu_1-\mu_2)$$

$$\Rightarrow w\Sigma(x_i-\mu)(x_i-\mu)^T + w^T\mu(m\mu-m\mu) = m(\mu_1-\mu_2)$$

$$\Rightarrow w\Sigma(x_i-\mu)(x_i-\mu)^T = m(\mu_1-\mu_2)$$

$$\Rightarrow w\left(\sum_{i=1}^{m_1}(x_i-\mu_1+\mu_1-\mu)(x_i-\mu_1+\mu_1-\mu)^T + \sum_{i=1}^{m_2}(x_i-\mu_2+\mu_2-\mu)(x_i-\mu_2+\mu_2-\mu)^T\right) = m(\mu_1-\mu_2)$$

$$\Rightarrow w\left(\sum_{i=1}^{m_1}(x_i-\mu_1)(x_i-\mu_1)^T + \sum_{i=1}^{m_2}(x_i-\mu_2)(x_i-\mu_2)^T + \sum_{i=1}^{m_1}(\mu_1-\mu)(\mu_1-\mu)^T + \sum_{i=1}^{m_2}(\mu_2-\mu)(\mu_2-\mu)^T\right) = m(\mu_1-\mu_2)$$

within class scatter matrix

Between-class matrix

$$\Rightarrow w(S_W + k_0 S_B) = m(\mu_1-\mu_2)$$

$$\Rightarrow S_W w = m(\mu_1-\mu_2)$$

$$w \propto S_W^{-1}(\mu_1-\mu_2)$$ (when we have ignored scale factors)

⇒ This is the same as the result we get from Fischer's data. Therefore, for the 2-class problem, the fischer criterian can be obtained as a special case of least square, if we adopt a different target coding scheme, like m/m_1 for C_1 and $-m/m_2$ for C_2.