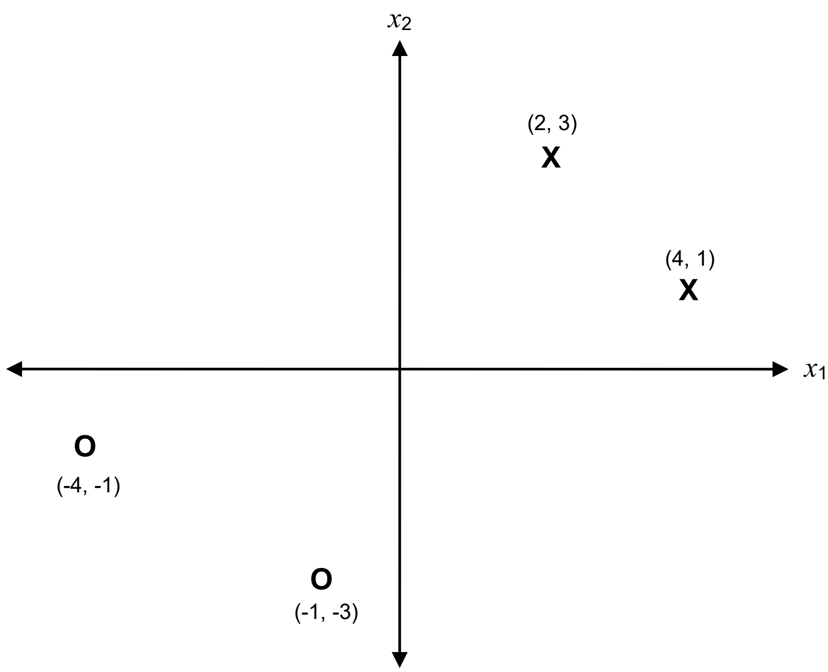
CSEN1083 – Data Mining

Problem Set #4

Problem 1

For the data given below, use the maximum likelihood estimate of Gaussian generative model to find the classifier parameters.



Solution:

- Let the target value for class C_1 (**X**) be $t = 1$ and for class C_2 (**O**) be $t = 0$.
- The prior probability of class C_1 is

$$p(C_1) = \pi = \frac{1}{4}\sum_{n=1}^{4} t_n = \frac{1}{4}(1+1+0+0) = 0.5$$

- The prior probability of class C_2 is thus

$$p(C_2) = 1 - \pi = 0.5$$

- The mean of the input vectors in class C_1 is

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{4} t_n \mathbf{x}_n = \frac{1}{2}\left(1\times\begin{bmatrix}2\\3\end{bmatrix} + 1\times\begin{bmatrix}4\\1\end{bmatrix} + 0\times\begin{bmatrix}-1\\-3\end{bmatrix} + 0\times\begin{bmatrix}-4\\-1\end{bmatrix}\right) = \begin{bmatrix}3\\2\end{bmatrix}$$

- The mean of the input vectors in class C_2 is

$$\boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{4} (1-t_n) \mathbf{x}_n = \frac{1}{2}\left(0\times\begin{bmatrix}2\\3\end{bmatrix} + 0\times\begin{bmatrix}4\\1\end{bmatrix} + 1\times\begin{bmatrix}-1\\-3\end{bmatrix} + 1\times\begin{bmatrix}-4\\-1\end{bmatrix}\right) = \begin{bmatrix}-2.5\\-2\end{bmatrix}$$

- The covariance matrix Σ is given by

$$\Sigma = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 = 0.5\mathbf{S}_1 + 0.5\mathbf{S}_2$$

$$\mathbf{S}_1 = \sum_{\mathbf{x}_n \in C_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T$$

$$= \left(\begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2 \end{bmatrix}\right)\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 2 \end{bmatrix}\right)^T + \left(\begin{bmatrix} 4 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 2 \end{bmatrix}\right)\left(\begin{bmatrix} 4 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 2 \end{bmatrix}\right)^T$$

$$= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}$$

$$\mathbf{S}_2 = \sum_{\mathbf{x}_n \in C_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T$$

$$= \left(\begin{bmatrix} -1 \\ -3 \end{bmatrix} - \begin{bmatrix} -2.5 \\ -2 \end{bmatrix}\right)\left(\begin{bmatrix} -1 \\ -3 \end{bmatrix} - \begin{bmatrix} -2.5 \\ -2 \end{bmatrix}\right)^T + \left(\begin{bmatrix} -4 \\ -1 \end{bmatrix} - \begin{bmatrix} -2.5 \\ -2 \end{bmatrix}\right)\left(\begin{bmatrix} -4 \\ -1 \end{bmatrix} - \begin{bmatrix} -2.5 \\ -2 \end{bmatrix}\right)^T$$

$$= \begin{bmatrix} 2.25 & -1.5 \\ -1.5 & 1 \end{bmatrix} + \begin{bmatrix} 2.25 & -1.5 \\ -1.5 & 1 \end{bmatrix} = \begin{bmatrix} 4.5 & -3 \\ -3 & 2 \end{bmatrix}$$

$$\Sigma = 0.5\mathbf{S}_1 + 0.5\mathbf{S}_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} 2.25 & -1.5 \\ -1.5 & 1 \end{bmatrix} = \begin{bmatrix} 3.25 & -2.5 \\ -2.5 & 2 \end{bmatrix}$$

CSEN1083 – Data Mining

Problem Set #4

Problem 2

Consider the data about conditions for playing tennis given below:

Day	Outlook	Temperature	Humidity	Play Tennis?
1	Sunny	Hot	High	No
2	Cloudy	Hot	High	**Yes**
3	Rain	Mild	High	No
4	Rain	Cool	Normal	No
5	Cloudy	Cool	Normal	**Yes**
6	Sunny	Mild	High	**Yes**
7	Sunny	Cool	Normal	**Yes**
8	Rain	Mild	Normal	No
9	Sunny	Mild	Normal	**Yes**
10	Cloudy	Mild	High	**Yes**

Using Naïve Bayes classifier, predict if one should play tennis on a Sunny, Hot with Normal humidity day. Note that since the data is discrete, you can use the frequentist statistics to compute the needed probabilities.

Solution:

Since the goal is to classify a sunny, hot with normal humidity day as good day to play tennis or not, we first define two classes C_1 and C_2, corresponding to Play = Yes and Play = No, respectively. To classify the given day with attributes $\mathbf{x}$, we need to compute $p(C_1 \mid \mathbf{x})$:

p(Play = Yes | Outlook = Sunny, Temperature = Hot, Humidity = Normal)

and $p(C_2 \mid \mathbf{x})$:

p(Play = No | Outlook = Sunny, Temperature = Hot, Humidity = Normal)

and find which conditional probability is larger. If the first one is larger, then our prediction is Play = Yes. If the second one is larger, then our prediction is Play = No. Note that $\mathbf{x}$ here is 3 dimensional corresponding to Outlook, Temperature and Humidity.

Since $p(C_1|\mathbf{x}) = \dfrac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x})}$

CSEN1083 – Data Mining

Problem Set #4

We need to compute $p(\mathbf{x} \mid C_1) = p(\text{Outlook} = \text{Sunny}, \text{Temperature} = \text{Hot}, \text{Humidity} = \text{Normal} \mid \text{Play} = \text{Yes})$. Using the Naïve Bayes assumption which assumes that the dimensions of the input data (the attributes of the day) are independent, we can re-write $p(\mathbf{x} \mid C_1)$ as $p(\mathbf{x}|C_1) = \prod_{i=1}^{D} p(x_i|C_1)$

$p(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{Yes})\ p(\text{Temperature} = \text{Hot} \mid \text{Play} = \text{Yes})\ p(\text{Humidity} = \text{Normal} \mid \text{Play} = \text{Yes})$

Similarly, $p(\mathbf{x} \mid C_2)$ can be re-written as $p(\mathbf{x}|C_2) = \prod_{i=1}^{D} p(x_i|C_2)$

$p(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{No})\ p(\text{Temperature} = \text{Hot} \mid \text{Play} = \text{No})\ p(\text{Humidity} = \text{Normal} \mid \text{Play} = \text{No})$

From the available data in the table and using frequentist statistics:

$p(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{Yes}) = 3/6$

$p(\text{Temperature} = \text{Hot} \mid \text{Play} = \text{Yes}) = 1/6$

$p(\text{Humidity} = \text{Normal} \mid \text{Play} = \text{Yes}) = 3/6$

$p(\text{Play} = \text{Yes}) = 6/10$

$p(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{No}) = 1/4$

$p(\text{Temperature} = \text{Hot} \mid \text{Play} = \text{No}) = 1/4$

$p(\text{Humidity} = \text{Normal} \mid \text{Play} = \text{No}) = 2/4$

$p(\text{Play} = \text{No}) = 4/10$

Therefore,

$p(\mathbf{x} \mid C_1) = p(\text{Outlook} = \text{Sunny}, \text{Temperature} = \text{Hot}, \text{Humidity} = \text{Normal} \mid \text{Play} = \text{Yes}) = (3/6) \text{ x } (1/6) \text{ x } (3/6)$

And

$p(\mathbf{x} \mid C_2) = p(\text{Outlook} = \text{Sunny}, \text{Temperature} = \text{Hot}, \text{Humidity} = \text{Normal} \mid \text{Play} = \text{No}) = (1/4) \text{ x } (1/4) \text{ x } (2/4)$

$$\frac{p(C_1|\mathbf{x})}{p(C_2|\mathbf{x})} = \frac{p(\mathbf{x}|Play = Yes)p(Play = Yes)}{p(\mathbf{x}|Play = No)p(Play = No)} = \frac{0.025}{0.0125} = 2$$

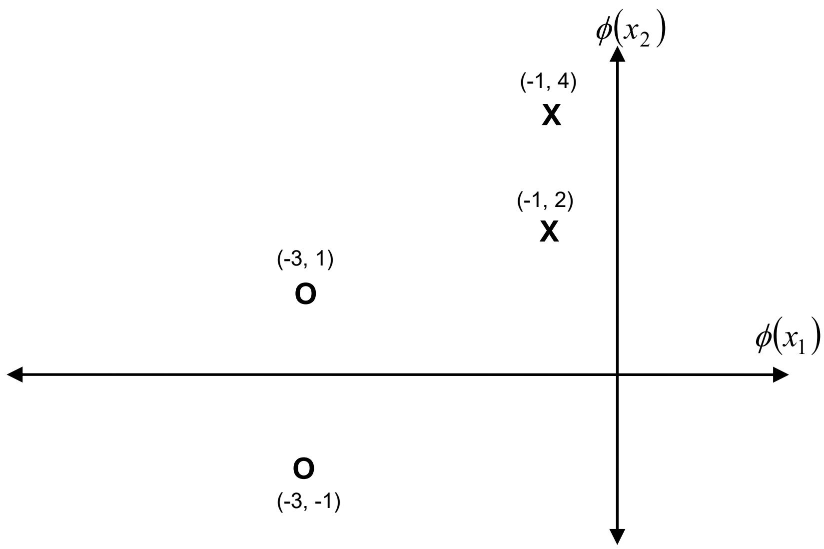
Since $\frac{p(C_1|\mathbf{x})}{p(C_2|\mathbf{x})} > 1$ then $\mathbf{x} \in C_1$ Therefore, one should play tennis in a sunny, hot with normal humidity day.

CSEN1083 – Data Mining

Problem Set #4

Problem 3

For the data given below, apply the logistic regression algorithm to find the weight vector **w** of the decision boundary. Show the output of each iteration till all the points are classified correctly. Assume that the weight vector is initialized $\mathbf{w}^{(0)} = [3\ 1]$. Use learning rate parameter $\eta = 0.5$.

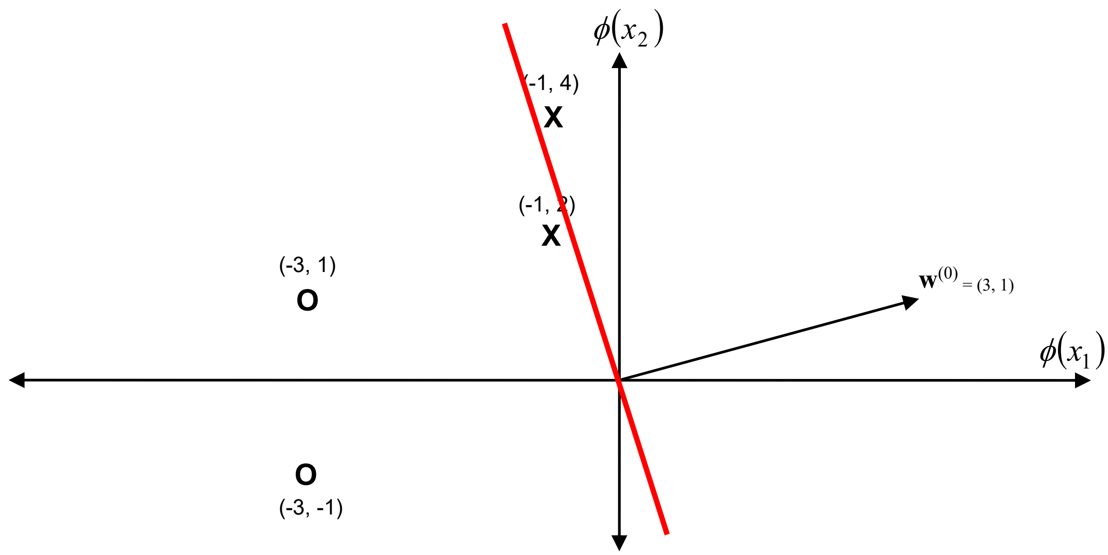


Solution:

The initial weight vector is shown below and the corresponding decision boundary $w_1x_1 + w_2x_2 = 0$ shown in red (which is perpendicular to the weight vector). For $\mathbf{w}^{(0)}$, the decision boundary is $3x_1 + x_2 = 0$. Let the target value t for points in the **X** class be 1 and for the **O** class be 0.

CSEN1083 – Data Mining

Problem Set #4



For iteration 1, based on the shown decision boundary, the y value for the points (Refer to slide 20 in Lecture 4)

(-1, 4) → y = 1/(1+exp(-w_1x_1 - w_2x_2)) = 1/(1+exp(-3*(-1) - 1*4)) = 0.73

(-1, 2) → y = 1/(1+exp(-3*(-1) - 1*2)) = 0.27

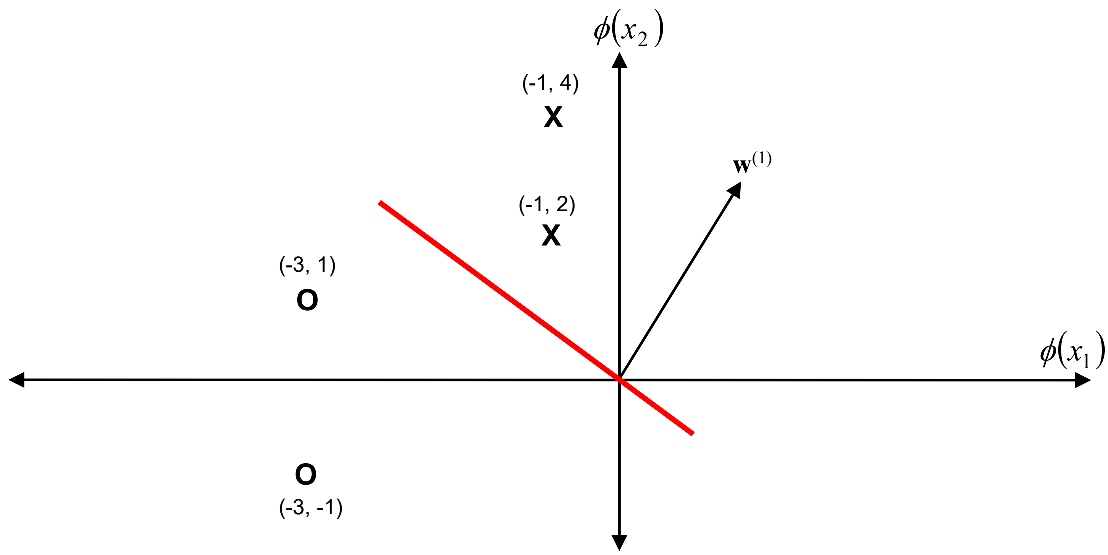
(-3, 1) → y = 1/(1+exp(-3*(-3) - 1*1)) = 0

(-3, -1) → y = 1/(1+exp(-3*(-3) - 1*-1)) = 0

The logistic regression update rule is given by $\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta\sum_{n=1}^{N}(y_n - t_n)\phi(\mathbf{x}_n)$

$$\mathbf{w}^{(1)} = \begin{bmatrix}3\\1\end{bmatrix} - 0.5\left((0.73-1)\begin{bmatrix}-1\\4\end{bmatrix} + (0.27-1)\begin{bmatrix}-1\\2\end{bmatrix} + (0-0)\begin{bmatrix}-3\\1\end{bmatrix} + (0-0)\begin{bmatrix}-3\\-1\end{bmatrix}\right) = \begin{bmatrix}2.5\\2.27\end{bmatrix}$$

Problem Set #4



Since all points are classified correctly, the algorithm stops.