Roll No.:___________________

CS 557: STATISTICAL PATTERN RECOG & LEARNING

Solutions Midterm exam 1, Fall 2016

QUESTION 1 (Marks: 5+5+2+3)

We have two classes, where the likelihood of data is modelled by Gaussian distribution. Suppose you are given the following statistics with two attributes x_1 and x_2:

For class 1: $\mu_1 = (0,0)$, $\Sigma_1 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$

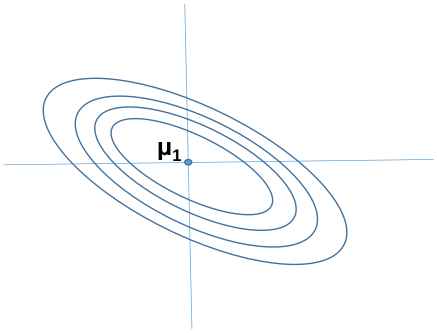
For class 2: $\mu_2 = (2,1)$, $\Sigma_2 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$

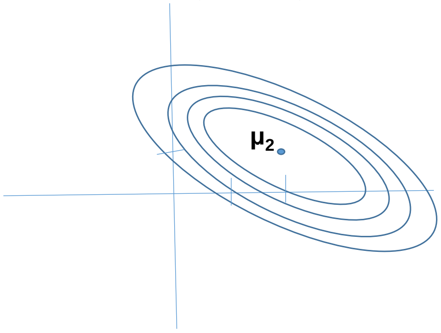
SOLUTION Part a

What do the contours of the two distributions look like. Draw them separately on two graphs.

SOLUTION

(note the covariance component is negative, that is why the ellipses are rotated)





Part b

Find the decision boundary between the two classes assuming that we use Maximum aposteriori classifier and both classes are equally likely. Show all working.

SOLUTION

We can see that as the two classes are equally likely so MAP classification is the same as ML classification. Also, plugging in the values of the two distributions in the following expression

P(C$_i$|**x**) = 1/(sqrt(2*pi) * 1/| Σ_i| *exp(-1/2 *(**x**- μ_i)$^T \Sigma_i^{(-1)}$ (**x**- μ_i))+ P(C$_i$)

(where i=1,2 for the above two classes)

As | Σ_1|=| Σ_2|

so when classifying examples, the only deciding factor that remains is the Mahalonobis distance of an example point from the corresponding mean of a class. The decision boundary can be found by solving for **x** in the following equation:

$$(\mathbf{x}-\mu_1)^T \Sigma_1^{(-1)} (\mathbf{x}-\mu_1) = (\mathbf{x}-\mu_2)^T \Sigma_2^{(-1)} (\mathbf{x}-\mu_2)$$

Plugging in all values from the above (do the working yourself), we get the equation of the line for the decision boundary:

$10x_1+6x_2-13=0$

Part c

On a separate graph plot both means and the decision boundary. Clearly write the coordinates of the points that meet the graph at any of the coordinate axis.

SOLUTION

(draw yourself). The points (0,13/6) and (13/10,0) lie on the decision boundary. Also, the boundary passes through the mid point of the line joining the two means, i.e., the point (1,1/2)

Part d

What is the classification of the points (1,1), (0.5,0.5) and (-1,-1). (No marks without proper working.)

SOLUTION

check the Mahalonobis distance of each of the above from the means of the corresponding classes(do the working yourself)

(1,1) belongs to Class 2

(0.5,0.5) belongs to Class 1

(-1,-1) belongs to Class 1

QUESTION 2 (Marks: 5+5+5)

We have three types of documents fiction, politics and sports. We check whether the three words, i.e., *player*, *game*, *winner* occur in a document or not. Here are some statistics that we gather:

- Word *player* occurs in 40% of fiction, 60% of politics and 80% of sports documents.
- Word *game* occurs in 20% of fiction, 70% of politics and 90% of sports documents
- Word *winner* occurs in 50% of fiction, 30% of politics and 70% of sports documents
- There are 50% fiction documents, 20% politics documents and 30% sports documents

PART a

What is the probability that we observe no occurrence of *player* but occurrences of *game* and *winner* in any document if we use **naive Bayes' assumption**. Write the formula you are using as well.

SOLUTION

We have to find:

P(~player,game,winner)

Use the sum rule of probability

P(~player,game,winner) = P(~player,game,winner|fiction)P(fiction)
+
P(~player,game,winner|politics)P(politics)
+
P(~player,game,winner|sports)P(sports)

(use naive Bayes' assumption)

P(~player,game,winner) = P(~player|fiction)P(game|fiction)P(winner|fiction)P(fiction)
+
P(~player|politics)P(game|politics)P(winner|politics)P(politics)
+
P(~player|sports)P(game|sports)P(winner|sports)P(sports)

Plug in all the values from above

P(~player,game,winner) = .6*.2*.5*.5+.4*.7*.3*.2+.2*.9*.7*.3
= 0.0846

PART b

If we observe no occurrences of *player* and *game* but an occurrence of *winner* in a document then which category of documents does it belong to**? You have to use Naive Bayes' assumption and MAP classification**. Show all working.

SOLUTION

Using MAP we choose the class for which $P(C_i|\mathbf{x})$ is maximum

(do the working yourself and write the mathematical expressions also

P(fiction|player,game,~winner) = .6*.8*.5*.5 / P(player,game,~winner)

P(politics|player,game,~winner) = .4*.7*.3*.2 / P(player,game,~winner)

P(sports|player,game,~winner) = .2*.9*.7*.3 / P(player,game,~winner)

You don’t need to compute the above. We can see that P(fiction|player,game,~winner) is the highest, so the MAP class is fiction.

PART c

Suppose we are allowed to modify the statistics related to the word winner found in documents related to politics. What should be the minimum probability P(word=*winner* | document=politics) for a document to be classified as politics when we observe all three words *winner*,*player* and *game* in it and the classification is done via **naive Bayes’ and maximum likelihood classification**. Show all working.

SOLUTION

Let P(winner|politics) = w
Then for **ML classification** we would like
P(winner,player,game|politics) to be higher than both P(winner,player,game|fiction) and P(winner,player,game|sports)

P(winner,player,game|politics) = w*.6*.7 = 0.42w
P(winner,player,game|fiction) = .4*.2*.5
P(winner,player,game|sports) = .8*.9*.7

Sorting the above, we want:

P(winner,player,game|politics) > P(winner,player,game|sports) > P(winner,player,game|fiction)

0.42w > 0.504

For the above to be true w has to be greater than 1, which is not possible as w represents a probability. Hence, no matter what the value of P(winner|politics), we choose, a document with all three words winner,player and game cannot be classified as belonging to politics.