# Mahalanobis Distance

Samuel Cripps V
210108060 EEE

# What is MD??

- In a given space we consider two types of distances namely Euclidean and Mahalanobis distances.
- In Mathematics Euclidean distance between two points is the length of line segment between them.
- The Euclidean distance is given by >>>

$$d(p,q) = d(q,p)$$
$$= \sqrt{(q_1 - p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- Mahalanobis distance(MD) is the distance between two points in a multivariate space.
- The MD measures distance relative to the centroid. It's expression is given below,

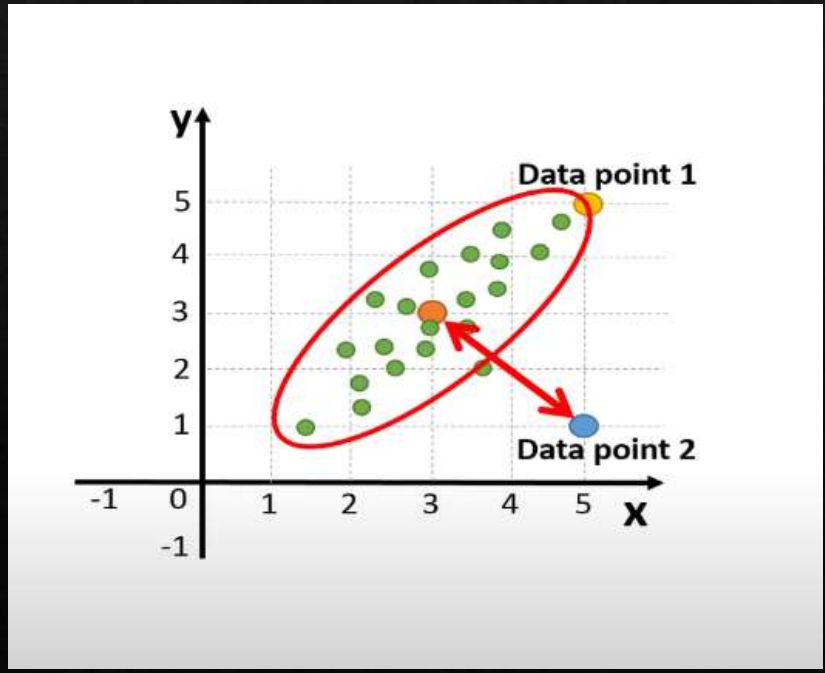
where x is the random variable vector ; u is the vector of mean values of independent variables ; S^-1 is the inverse covariance matrix of the independent variables.

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

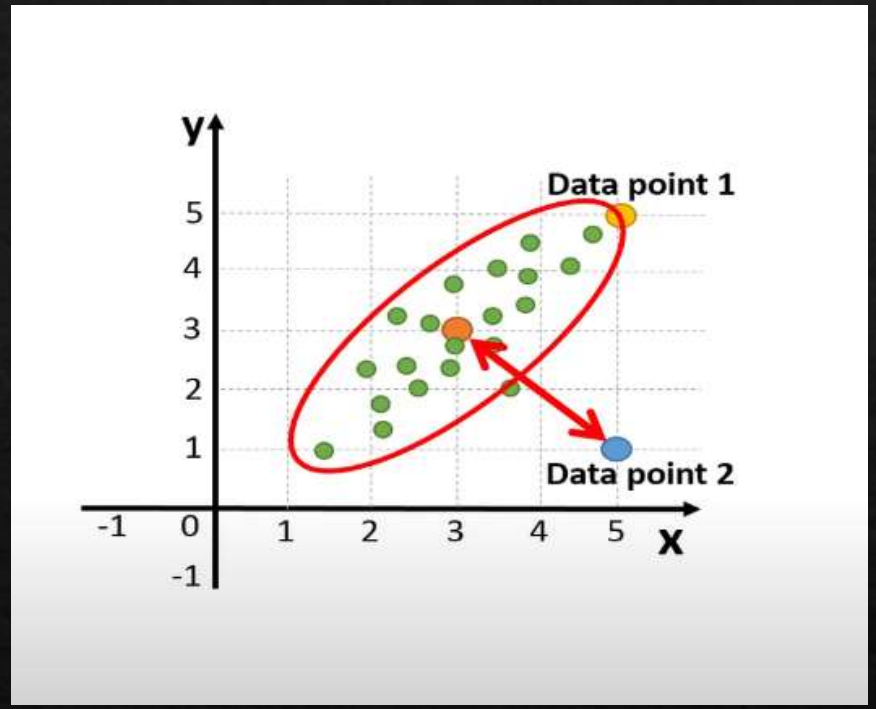
# Difference between the Euclidean's and the Mahalanobis'

- In a regular Euclidean space, the variables are represented by axes drawn at right angles and the distance between any two points can be measured by using a ruler.
- When the variables (independent) in a distribution have correlation , the axes are no longer at right angles and the measurements become impossible with a ruler. Also if the number of variables are more than three, then plotting them in a regular 3D space is impossible.
- MD solves this problem as it can measure the distance between points , even correlated points for multiple variables. (multivariate space)

- MD measures the distance from the centroid.
- Centroid is a base or central point which can be thought of as an overall mean for multivariable data. Centroid is a point in multivariate space where all the means from all the variables intersect.
- Larger the MD, farther away from the centroid the data point is.
- In the figure given, the Euclidean distances of the two data points from the centroid is same, but the Mahalanobis distances of the two data points from the centroid are very different.



- Generally MD is also defined as the distance between a point and a distribution.
- The Mahalanobis distance takes into account the correlation between the independent variables involved.
- In the given figure, clearly the Data point 1 is closer to the elliptic distribution as compared to the Data point 2. Hence the MD of point 1 is less than the MD of point 2 whereas the Euclidean distances of the Data points remain same while measured from the centroid.



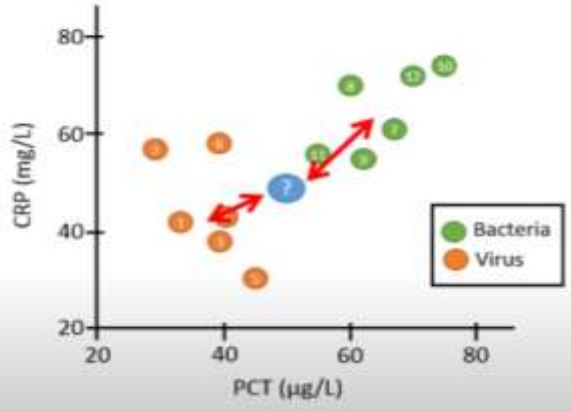
# Principle and Applications :

- The principle / idea involved in calculating the Mahalanobis distance is by measuring how many deviations away point P is from the mean of the distribution.
- The Mahalanobis distance is most commonly used in multivariate statistics. It determines whether a sample is an outlier, whether a process is in control or whether  a sample is a member of a group or not.
- It is has its application in Machine Learning where it is used to classify data and predict the outcome in Crossvalidation.

# Application In ML:

Consider the following dataset where we have to predict and classify whether the infection of a patient is Viral or Bacterial.

| Infection | CRP (mg/L) | PCT (µg/L) |
|---|---|---|
| Viral | 42 | 33 |
| Viral | 57 | 29 |
| Viral | 38 | 39 |
| Viral | 43 | 40 |
| Viral | 30 | 45 |
| Viral | 58 | 39 |
| Bacterial | 61 | 67 |
| Bacterial | 70 | 60 |
| Bacterial | 55 | 62 |
| Bacterial | 74 | 75 |
| Bacterial | 56 | 55 |
| Bacterial | 72 | 70 |

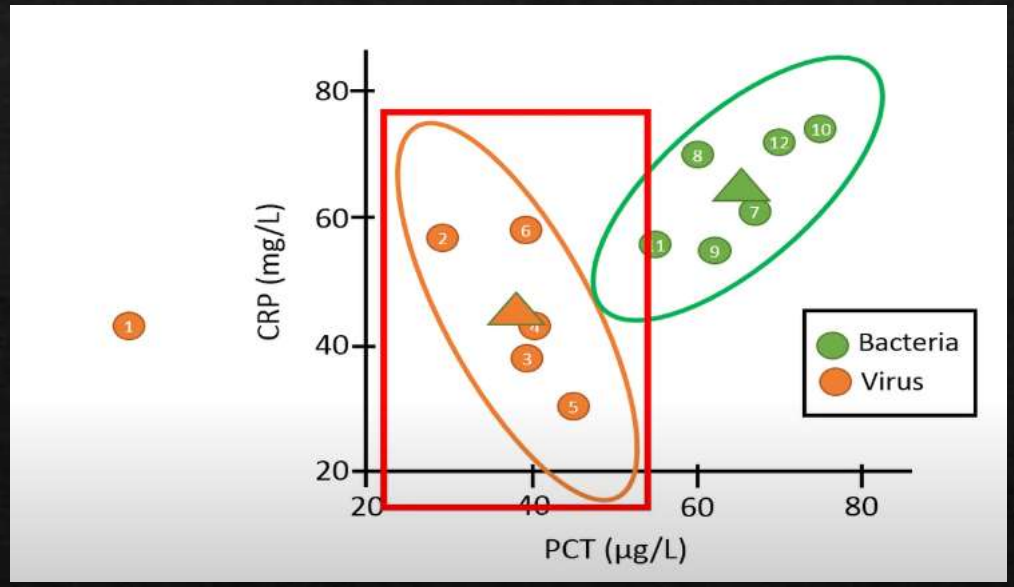



In order to classify an unknown data point into viral class or bacterial class we have to calculate the Mahalanobis Distance of that point from the centroid of the two distributions.

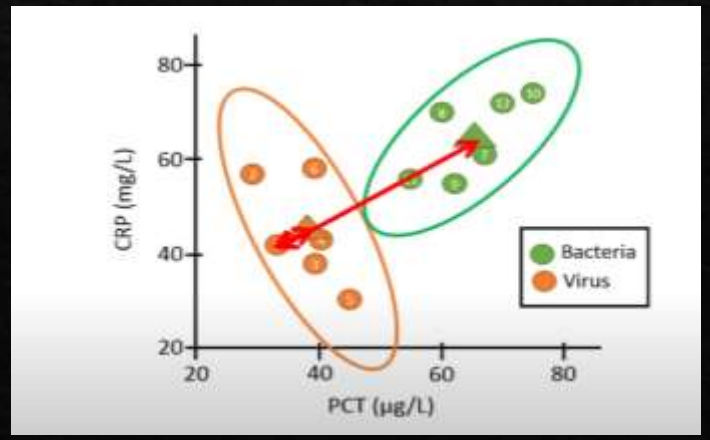
If the MD from the centroid of Viral distribution is smaller then that data point can be classified into the Viral class otherwise it goes into the Bacterial class.

# MD to predict the Classification

In order to predict a given data point from a distribution (in cross validation) we have to remove that particular data point from the distribution and then calculate the new centroid of that particular distribution. After that we have to put that point back in the distribution and measure it MD from both the centroids and decide to which class it belongs to.




| Infection | CRP (mg/L) | PCT (µg/L) | Predict |
|---|---|---|---|
| Viral | 42 | 33 | Viral |
| Viral | 57 | 29 | |
| Viral | 38 | 39 | |
| Viral | 43 | 40 | |
| Viral | 30 | 45 | |
| Viral | 58 | 39 | |
| Bacterial | 61 | 67 | |
| Bacterial | 70 | 60 | |
| Bacterial | 55 | 62 | |
| Bacterial | 74 | 75 | |
| Bacterial | 56 | 55 | |
| Bacterial | 72 | 70 | |

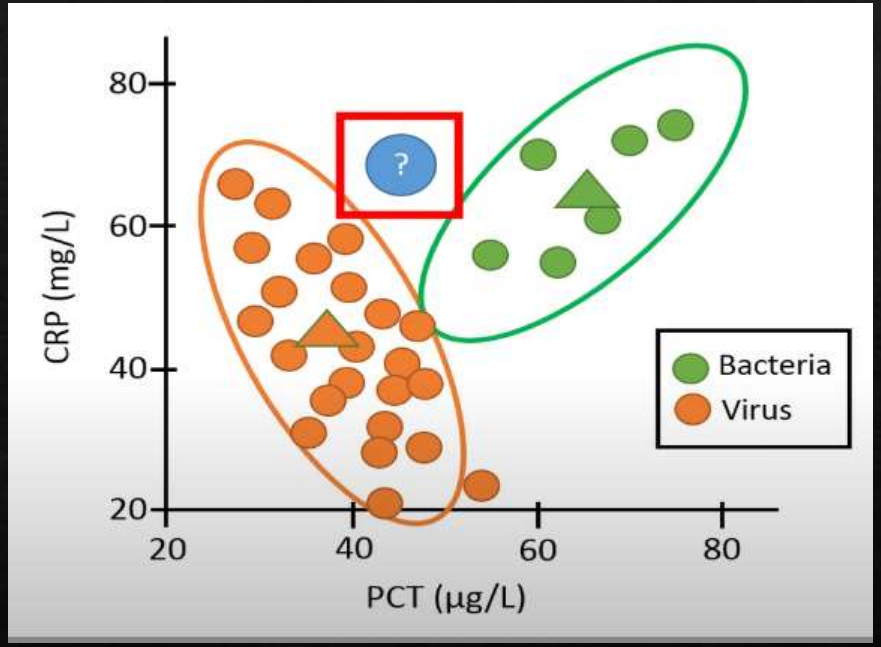



| Infection | CRP (mg/L) | PCT (µg/L) | Predict |
|---|---|---|---|
| Viral | 42 | 33 | Viral |
| Viral | 57 | 29 | Viral |
| Viral | 38 | 39 | Viral |
| Viral | 43 | 40 | Viral |
| Viral | 30 | 45 | Viral |
| Viral | 58 | 39 | Bacterial |
| Bacterial | 61 | 67 | Bacterial |
| Bacterial | 70 | 60 | Bacterial |
| Bacterial | 55 | 62 | Bacterial |
| Bacterial | 74 | 75 | Bacterial |
| Bacterial | 56 | 55 | Bacterial |
| Bacterial | 72 | 70 | Bacterial |

# Advantage :

One of the advantages of using the Mahalanobis distance in comparision is that it works fine with imbalanced data sets.

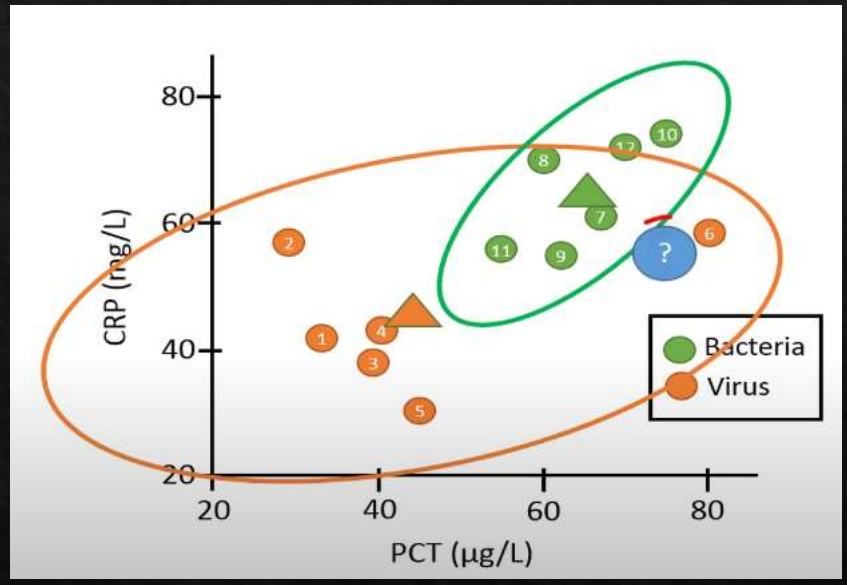
When any one of the distribution has much more data points than the other, the classification gets easier as the MD between a new observation and the two groups will not be affected by the fact that the viral group has more data because we expect that the covariance and the centroid will be roughly the same even though we add more data points.



# Disadvantages:

However one of the major disadvantage is that it is very sensitive to outliers.

Suppose a data point from the Viral set is placed close to the Bacteria set and a new example with an unknown class is located there, then it will be classified as belonging to the Viral class even though it belongs close to the distribution of Bacterial class due to the outlier (denoted by 6).



# Origin

- This distance was introduced by Indian Statistician and Scientist P.C.Mahalanobis in 1936.
- It was prompted by the problem of identifying the similarities of skulls based on measurements in 1927.
- He was also the member of first planning commission of free India.
- His statistical methods including the MD have been applied to India's social and economic problems that further India's efforts to industrialize in 1950s and 60s.

Interesting Point:

Indian Mathematical genius S.S. Ramanujan and P.C. Mahalanobis interacted with one another during their time at Cambridge.