5. Models based on distances

Distance is the amount of space between two samples. Formally, a distance is a function with the following characteristics:

- It is not negative. $D(x, y) \geq 0,\ \forall x, y$
- It is symmetric. $D(x, y) = D(y, x),\ \forall x, y$
- It satisfies the triangle inequality. $D(x, y) \leq D(x, z) + D(z, y),\ \forall x, y, z$
- The distance between a sample and itself is 0. $D(x, x) = 0,\ \forall x$

Some distances are:

Vectors	
Euclidean distance	$\|\|A-B\|\|_2 = \sqrt{\sum_i (A_i - B_i)^2}$
Manhattan distance	$\|\|A-B\|\|_1 = \sum_i \|A_i - B_i\|$
Maximum distance	$\|\|A-B\|\|_\infty = \max_i \|A_i - B_i\|$
Mahalanobis distance	$D_{Mahalanobis}(A,B) = \sqrt{(A-B)^T \Sigma^{-1} (\mathrm{A} - \mathrm{B})}$ Σ is the covariance matrix
Cosine similarity	$\mathrm{Cos\_sim}\ (\mathrm{A,B}) = \frac{A.B}{\|\|A\|\|\|\|B\|\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i}\sqrt{\sum_{i=1}^{n} B_i}}$
Words	
Hamming distance	Levenshtein distance
Probability distributions	
Kullback – Leibler divergence	$D_{KL}(P, Q) = \sum_i P(i) \ln\left(\frac{P(i)}{Q(i)}\right)$

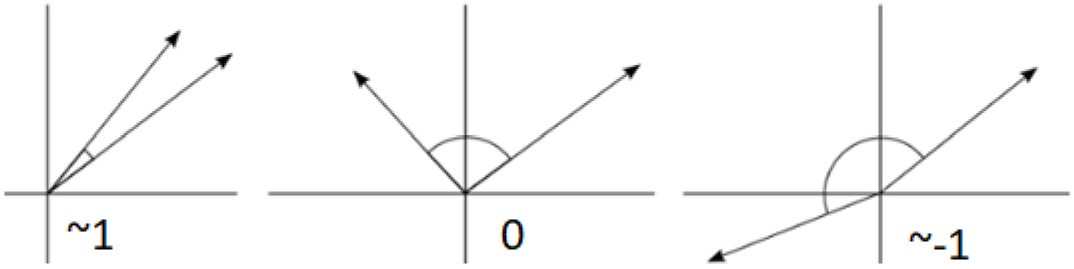
Cosine similarity

It measures the cosine of the angle between two vectors A and B. In other words, it measures the similarity between vector directions.

$$\cos(\theta) = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i}\sqrt{\sum_{i=1}^{n} B_i}}$$

The value can be between -1 and 1:

- -1, it means the vectors are opposite
- 0, it means the vectors are orthogonal
- 1, it means the vectors have the same direction



Mahalanobis distance

Explanation with an example:

Imagine that a fisher wants to measure the similarity among salmons because he wants to classify them into two groups for selling the bigger ones at a higher price. For each salmon, he measures the width and the length. Each salmon can be represented as a vector whose entries are these measures $\vec{x_i} = [x_{1i}, x_{2i}]^T$.

The length is a random variable with values between 50 and 100cm, whereas the width values are between 10 and 20cm. If the fisher uses a Euclidean distance, the length will have more importance than the width. For that reason, he decides to use the following equation:

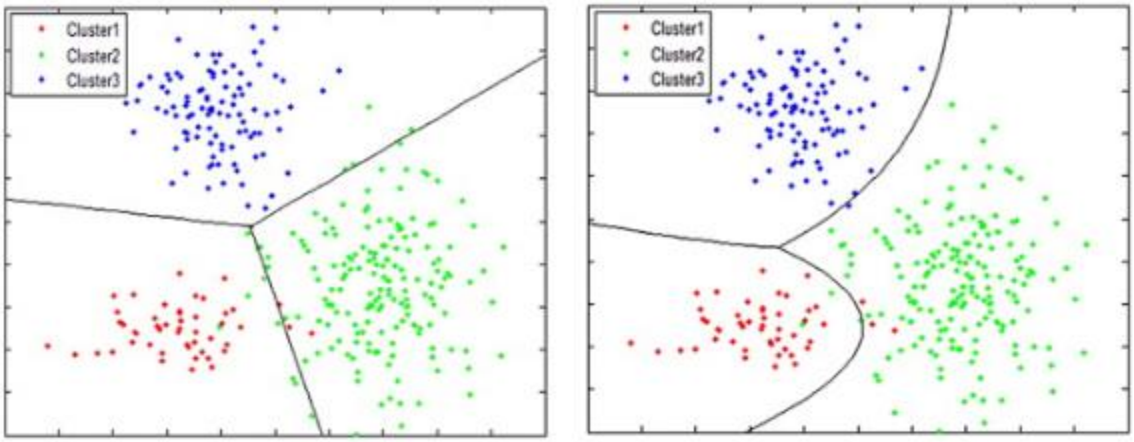
$$Mahalanobis\_distance(x_1, x_2) = \sqrt{\left(\frac{x_{11} - x_{12}}{\sigma_1}\right)^2 + \left(\frac{x_{21} - x_{22}}{\sigma_2}\right)^2} = \sqrt{(\vec{x_i} - \vec{x_2})^T S^{-1} (\vec{x_i} - \vec{x_2})}$$

where $S = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$

In general, the equation of Mahalanobis distance is:

$$Mahalanobis\_distance(x_1, x_2) = \sqrt{(\vec{x_i} - \vec{x_2})^T \Sigma^{-1} (\vec{x_i} - \vec{x_2})}$$

where Σ is the covariance matrix



(a) Clusters boundaries by Euclidean distance

(b) Clusters' boundaries by Mahalanobis distance

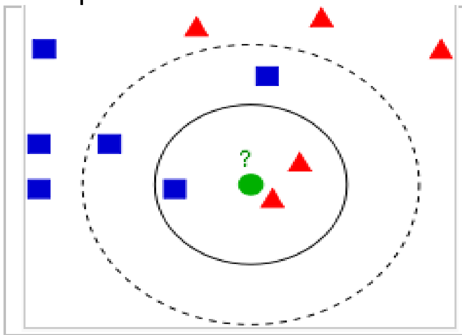
5.1 k - Nearest Neighbors (KNN)

Supervised learning: Classification
Variable type: all

It is a simple classifier that assigns the label that corresponds to the mode of the k nearest neighbors. It is sensible to the value of k.
Disadvantage: It has high complexity. To predict a label, it calculates the distance against the sample and all the training samples to find the k nearest neighbors.

```
Input: samples, k (number of neighbors) and the point to be predicted
Begin
   Get the k nearest neighbors to the point to be predicted
   Return the label that corresponds to the mode of the k nearest neighbors' labels
End
```

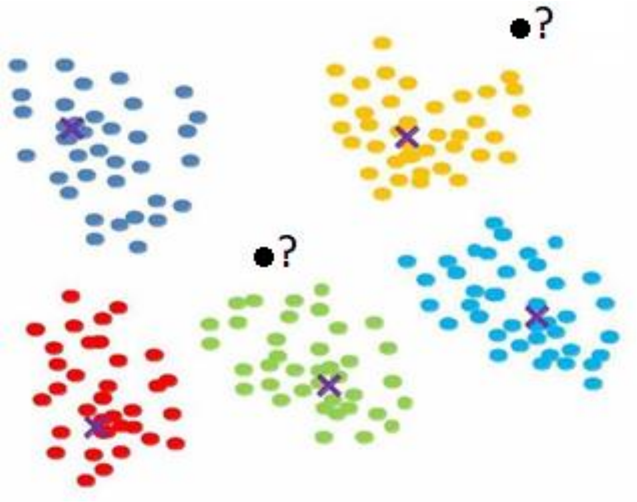
Example:

	3 – nearest neighbors The circle gets the label of red triangle 5 – nearest neighbors The circle gets the label of blue square 11 – nearest neighbors The circle gets the label of red triangle

5.2 Nearest Centroid

Unsupervised learning: Classification
Variable type: continuous

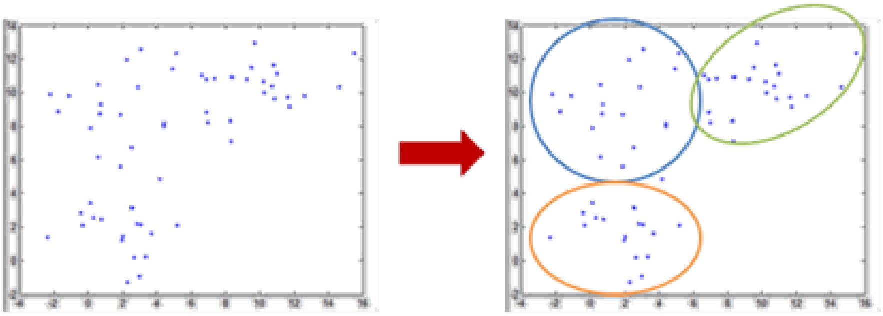
It is a simple classifier that represents each class by the centroid of its samples. It assigns the label corresponding to the class whose centroid is the nearest to the sample to be predicted.
Disadvantage: The problem is that it assumes unimodal distributions on the classes.



5.3 Hierarchical clustering

Unsupervised learning: Clustering
Variable type: all

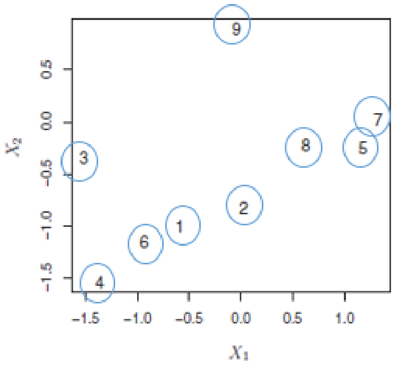
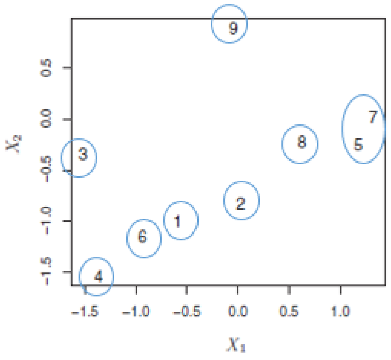
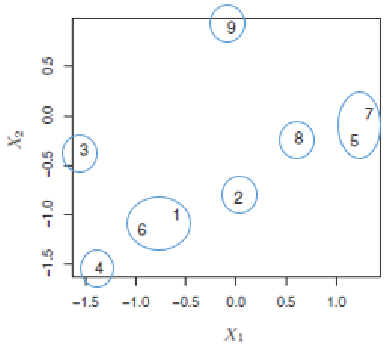
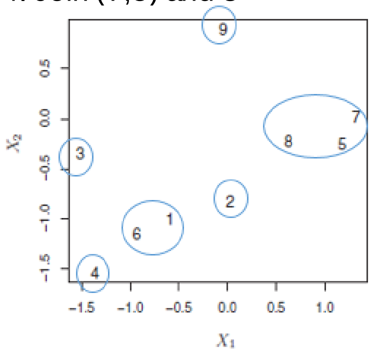
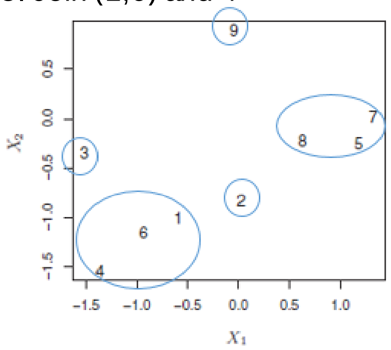
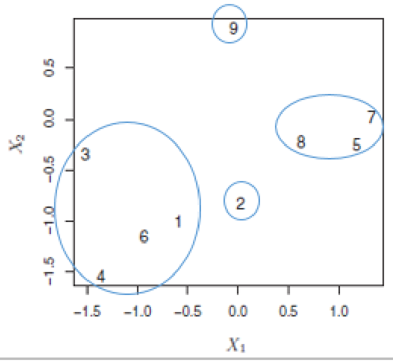
Remembering, clustering consists of group samples based on the features.

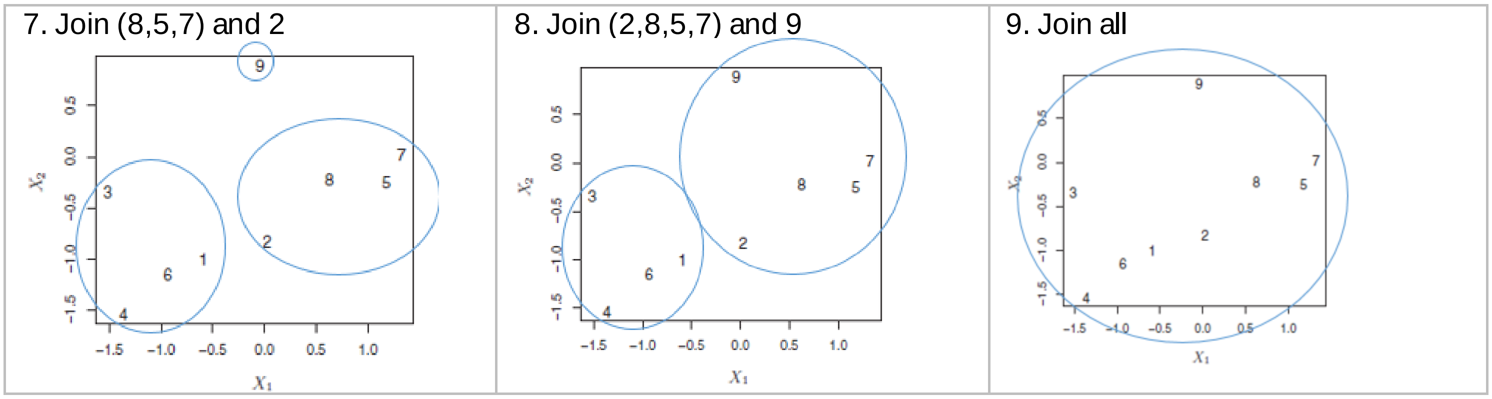


Hierarchical clustering works iteratively. The algorithm is:

```
Input: samples
Begin
    Each sample is a cluster
    Repeat until there is only one cluster
         Join the nearest two clusters
End
```

Example:

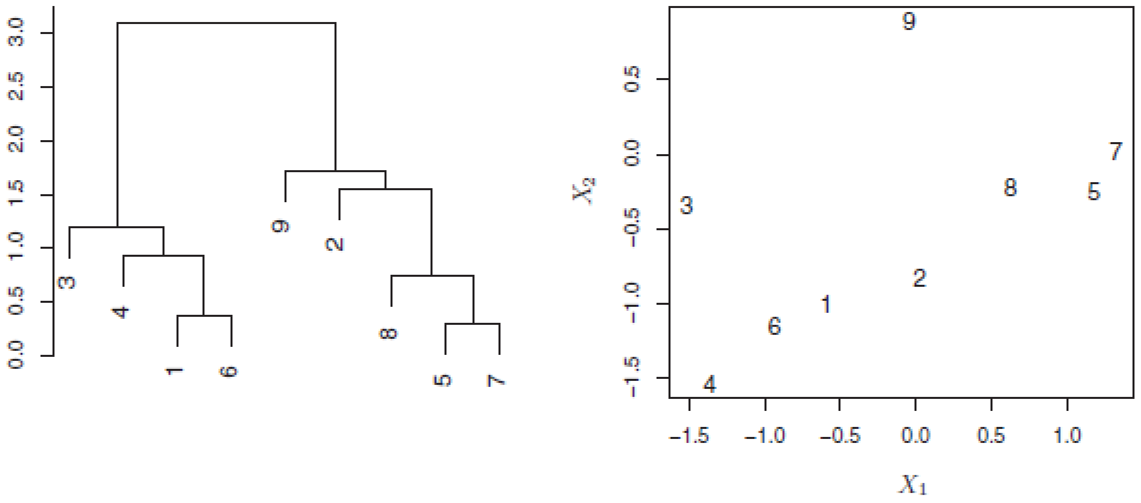
1. All the samples are different clusters 	2. Join 5 and 7 	3. Join 1 and 6 
4. Join (7,5) and 8 	5. Join (1,6) and 4 	6. Join (1,6,4) and 3 


7. Join (8,5,7) and 2

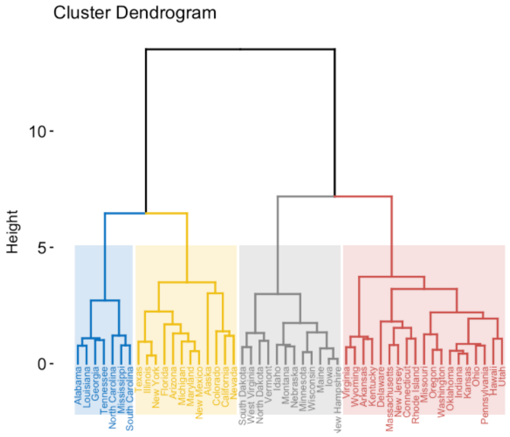
8. Join (2,8,5,7) and 9

9. Join all

We can represent the clustering process with a dendrogram that is a binary tree where the length of the branch represents the distance where the samples were joined.



The dendrogram can be used to analyze the number of clusters. In addition, by pruning the tree, the clusters can be found.

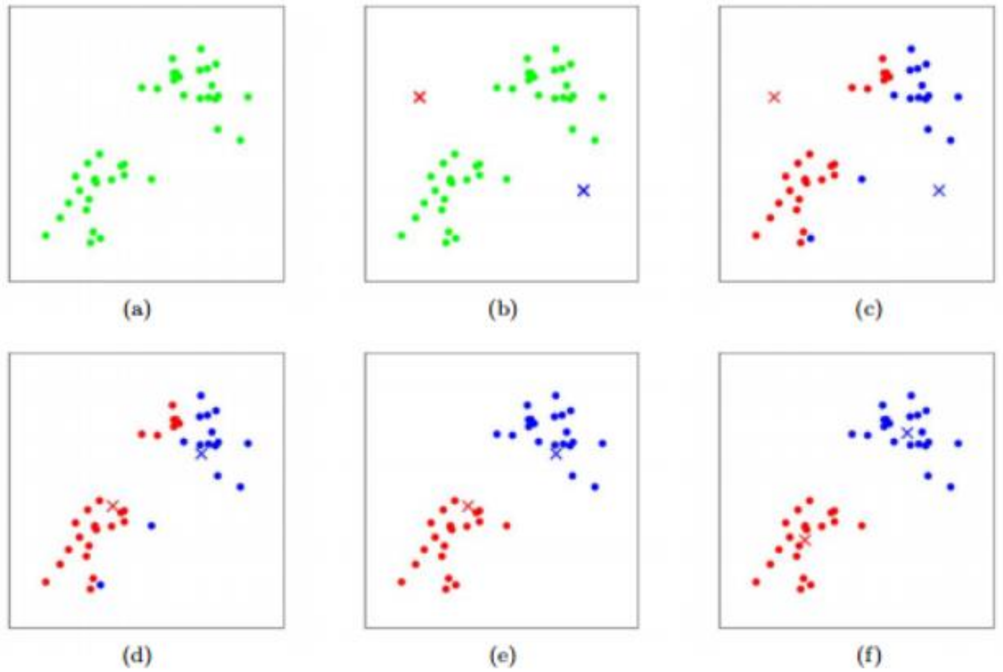


5.4 k - Means

Unsupervised learning: Clustering
Variable type: continuous

K-means finds k groups in the unlabeled data. An important parameter is the number of clusters we expect to find, called k. The algorithm is:

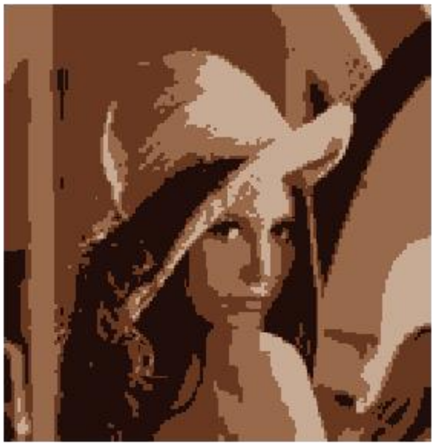
```
Input: samples and k (the number of clusters)
Begin
    Randomly select k prototypes
    Repeat until the prototypes do not move
        Assign the samples to the nearest prototype
        Update the prototypes as the centroid of the samples
End
```



(a) (b) (c)

(d) (e) (f)

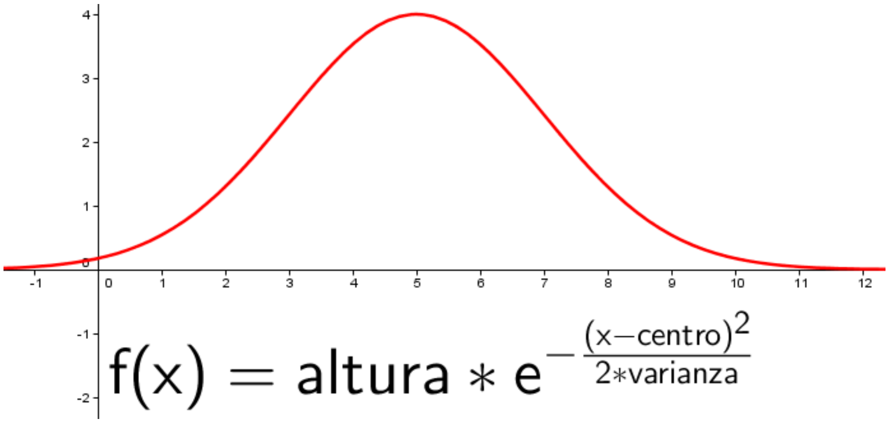
For example, we can use k-means for grouping the image's colors



5.5 Gaussian Mixture Models (GMM)

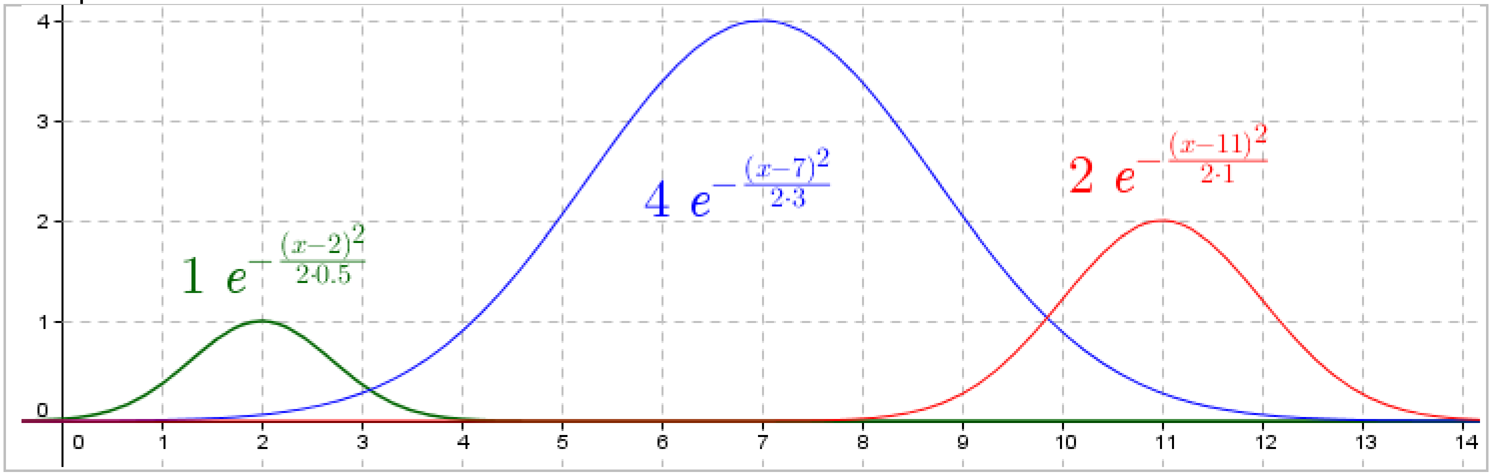
Unsupervised learning: Clustering
Variable type: continuous

A Gaussian is a function:



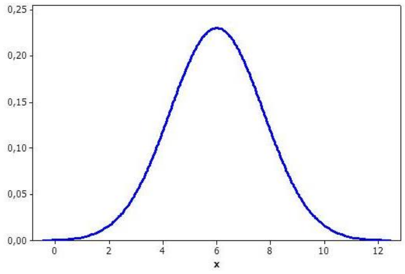
$$f(x) = \text{altura} * e^{-\frac{(x-\text{centro})^2}{2*\text{varianza}}}$$

Examples:

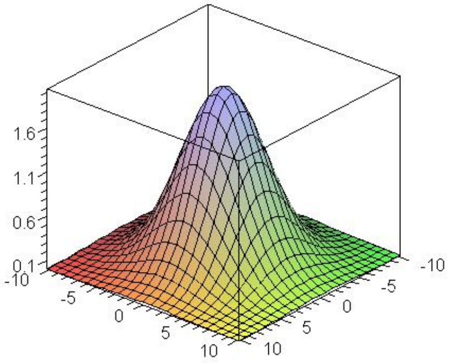


It can be seen in several dimensionality spaces:

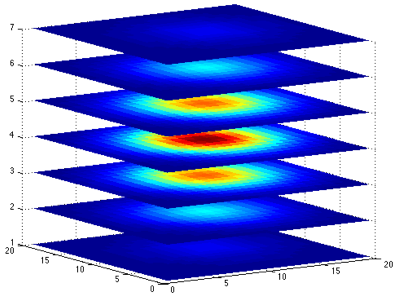
1-D



2-D

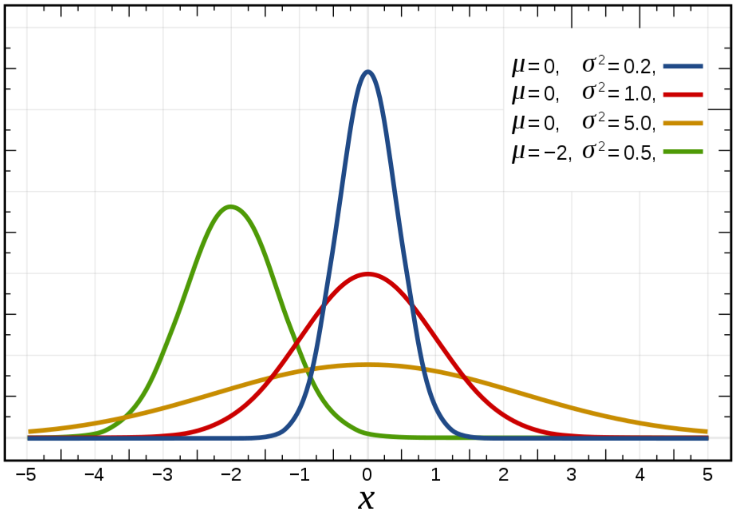


3-D

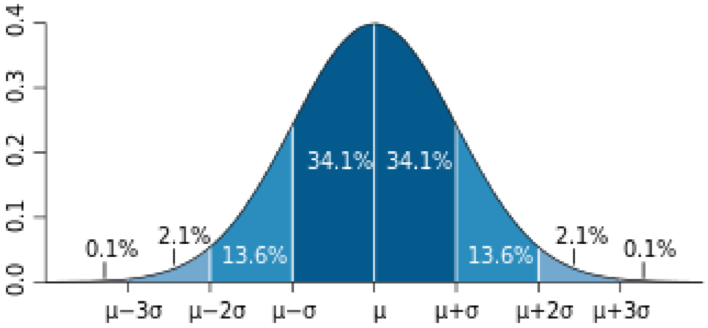


The Normal distribution is represented with a Gaussian:

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

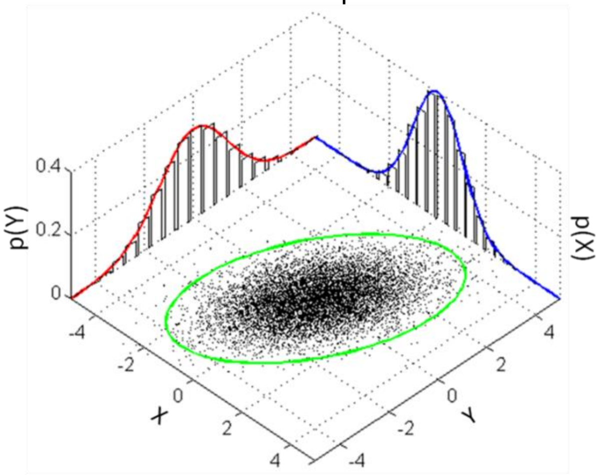


$$\mu = mean$$
$$\sigma^2 = variance$$



A Multivariate normal distribution is represented with a Gaussian in a d-dimensional space.

$$P(X|\mu,\sigma^2) = \frac{e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)}}{\sqrt{(2\pi)^d|\Sigma|}}$$

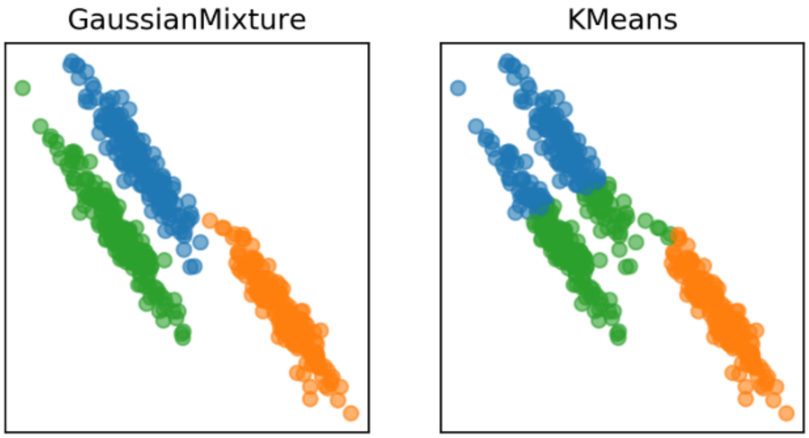


$$d = dimensionality\ of\ X$$
$$\mu = centroid$$
$$\sum = covariance\ matrix$$
$$|\Sigma| = \det \Sigma$$

The clustering algorithm to find the k groups in the data is similar to k – means algorithm, it is called Expectation Maximization (EM).

Expectation Maximization algorithm:

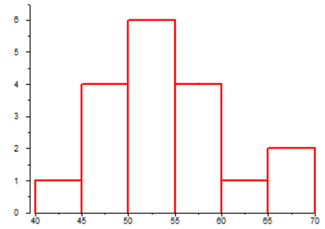
```
Input: X (training matrix nSamples x nFeatures) and k (number of Gaussians)

Randomly calculate K prototypes (centroids and covariance matrices)
Repeat until convergence
    Assign the samples to the Gaussian with more likelihood
    Calculate the parameters of the K Gaussians based on the samples assigned to them
```



5.6 Probability distributions

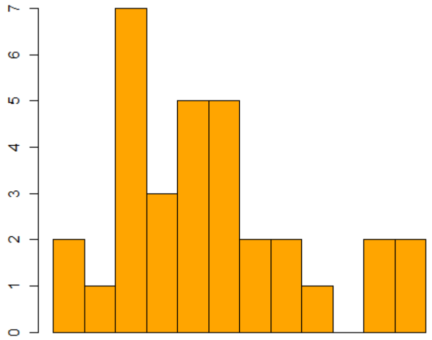
Histograms

A histogram is a representation of a variable, commonly it is plotted similar to a bar plot. The bar length represents the frequency of a value. The variables need to be discrete to calculate the frequencies, or we must discretize them.

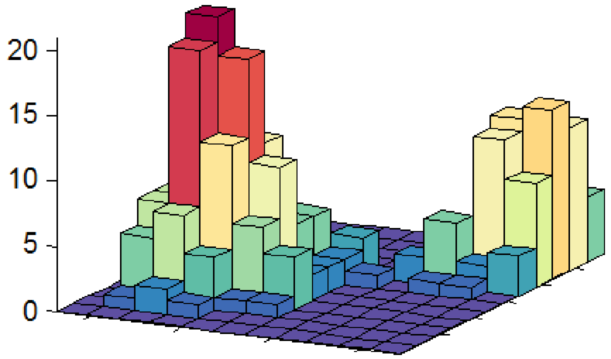


The histograms can be represented in several dimensions:

1D



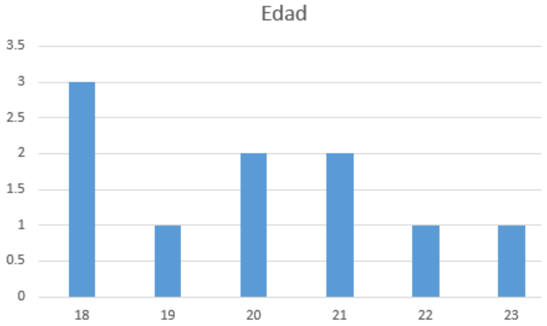
2D



Example in 1D

Ages: 20, 18, 19, 20, 21, 18, 23, 18, 21, 22

Age	Frecuency
18	3
19	1
20	2
21	2
22	1
23	1



Example in 2D:

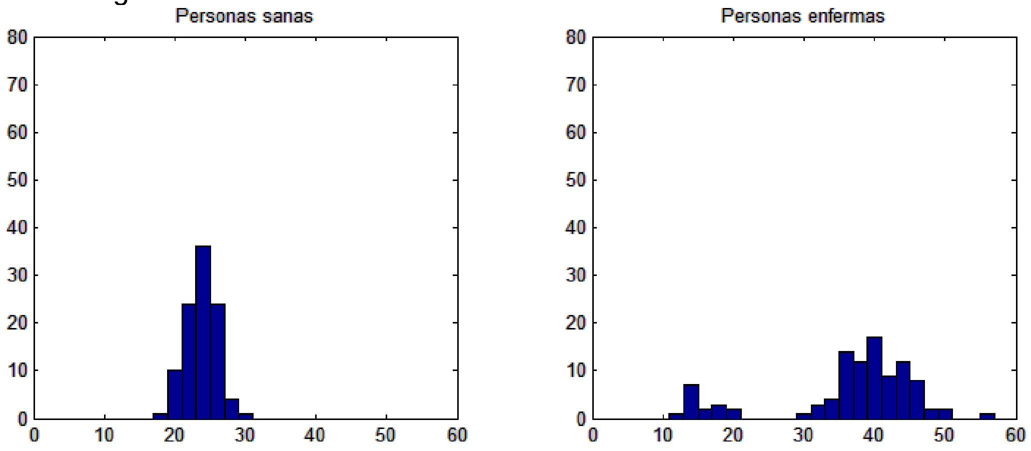
	Mathematics	Computation	Mechatronics	Pedagogy
Women	10	6	4	50
Men	12	28	30	2

Classification

1. Create a histogram for each class using the training dataset
2. Classify new data using the likelihood with each histogram

Example: Classification of healthy and unhealthy people based on Body Mass Index

Step 1. Create the histograms



Step 2. Classify new data

Person A: BMI = 35

Likelihood with healthy people: 0 Likelihood with unhealthy people: 12

Labeled as: unhealthy person

Person B: IMC = 20

Likelihood with healthy people: 10 Likelihood with unhealthy people: 2

Labeled as: healthy person

Distance between distributions

The distance between two histograms can be performed using Kullback – Leibler divergence

$$D_{KL}(P,Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

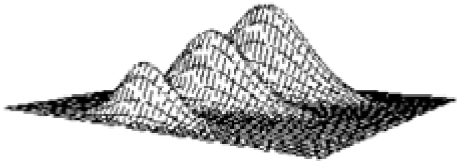
		P(X)		Q(X)
A		0.9		0.1
B		0.5		0.8
C		0.5		0.1

$$D_{KL}(P,Q) = 0.9 * \log\left(\frac{0.9}{0.1}\right) + 0.05 * \log\left(\frac{0.05}{0.8}\right) + 0.05 * \log\left(\frac{0.05}{0.1}\right)$$

Gaussian Mixture Models

A Gaussian Mixture Model can model complex distribution using a Gaussian linear combination. Examples:

GMM in 2D



GMM in 3D

The probability density function of a GMM is defined as:

$$\boldsymbol{P(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x|u_k, \Sigma_k)}$$

where:

K is the Gaussian number
$\mathcal{N}(x|u_k,\Sigma_k)$ is the probability density function of the Gaussian k
π_k is the weight of the Gaussian k
u_k is the mean vector of the Gaussian k
Σ_k is the covariance matrix of the Gaussian k

Classification

1. Create a GMM for each class using the training dataset
2. Classify new data using the likelihood with each GMM

Distance between distributions

The distance between two Multivariate Normal distributions $\mathcal{N}(\mu,\Sigma)$ can be performed using Bhattacharyya distance:

$$D_B = \frac{1}{8}(\mu_i - \mu_j)^T E^{-1}(\mu_i - \mu_j) + \frac{1}{2}\ln\left(\frac{\det E}{\sqrt{\det E_i \det E_j}}\right)$$

where:

μ_i and E_i are the mean and the covarance matrix of the Gaussian
$E = \frac{E_i + E_j}{2}$

The distance between two GMM distributions can be performed by a weighted sum of the distances between all the pair of Gaussians.