

Unsupervised Learning (K-Means, GMM)

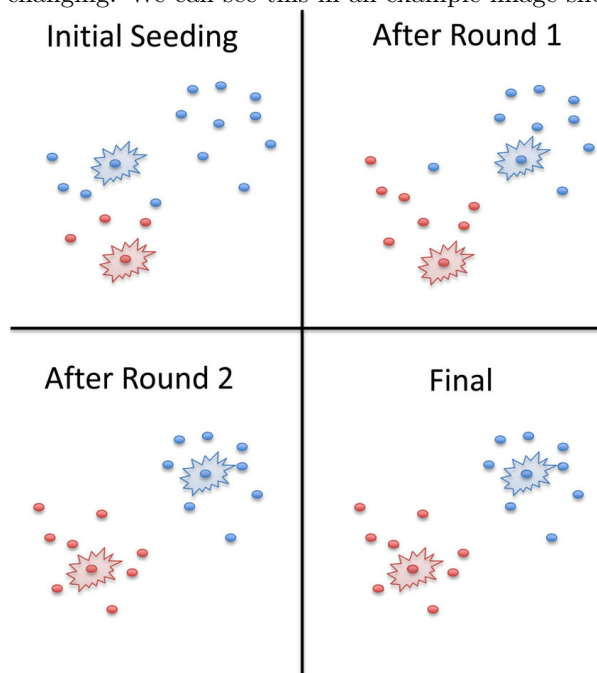
Prepared by: 2019201003, 2018201059, 2019201049

1 K-Means Summary

K-Means is an unsupervised learning algorithm that allows us to identify similar groups or clusters of data points within our data. It is an iterative algorithm that tries to partition the dataset into K pre-defined distinct clusters where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way kmeans algorithm works is as follows:

1. Specify number of clusters K .
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing. We can see this in an example image shown below



The objective function for K-Means is

$$J = \sum_{i=1}^m \sum_{k=1}^K w_{ik} \|x^i - \mu_k\|^2$$

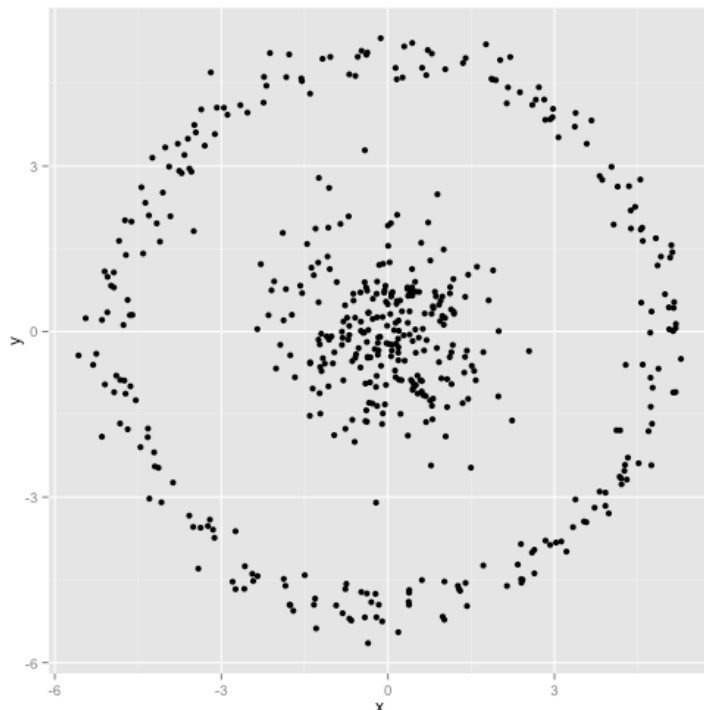
,where $w_{ik}=1$ for data point x_i if it belongs to cluster k ; otherwise, $w_{ik}=0$. Also, μ_k is the centroid of x_i 's cluster. The objective function for K-Means that we see here is non convex.

2 Limitations of K-Means Algorithm

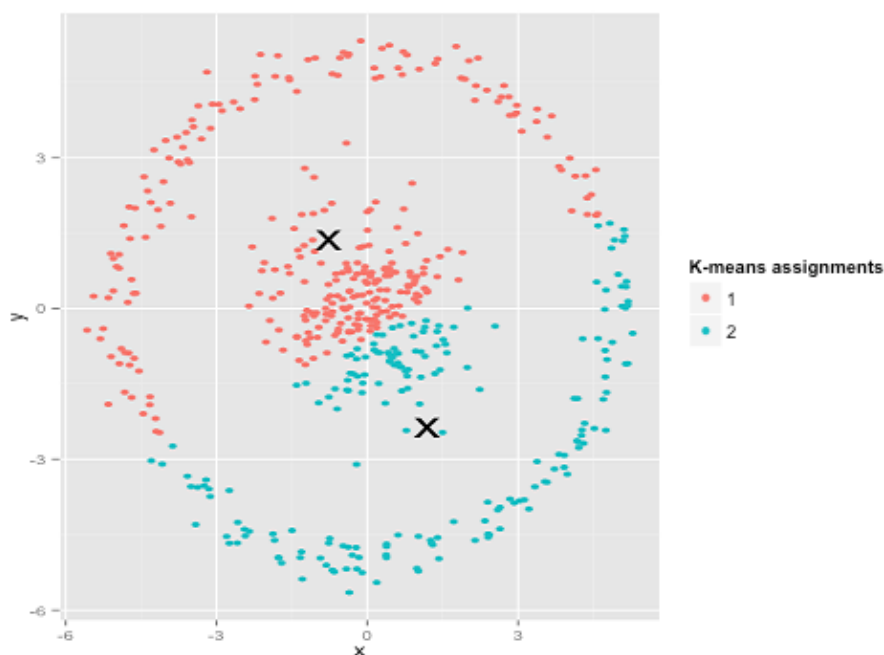
- It assumes that the clusters are spherical

This assumption essentially translates to all variables having the same variance or in other words, a diagonal covariance matrix with constant variance on the diagonal. If this is not the case, which in practice it often isn't then k-means may not be the best solution.

Let us assume we have a dataset like this

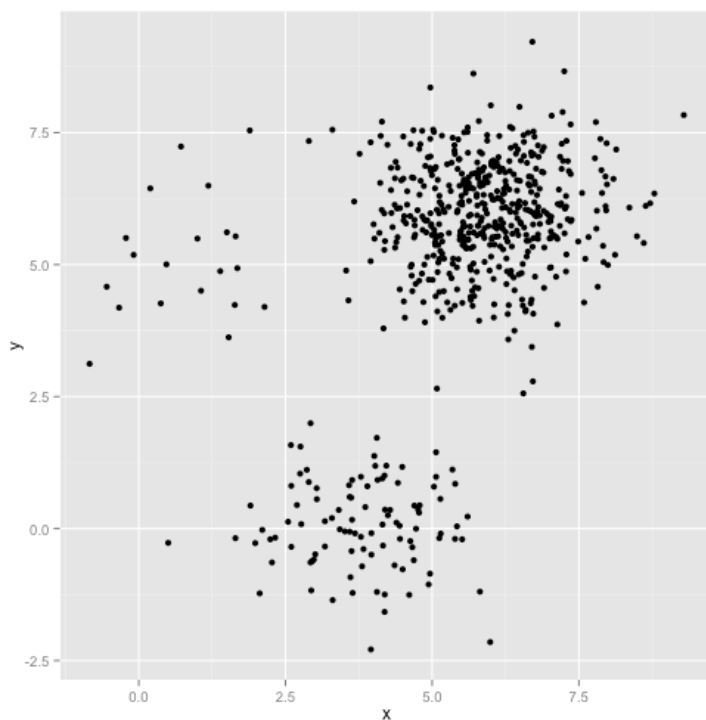


We, humans immediately recognize two natural groups of points here in the picture. But, that might not be the case with K-Means! In the picture shown below we can see how K-Means miserably fails in such a case by trying to fit a square peg in a round hole- thus trying to find nice centers with neat spheres around them.



- Unevenly Sized Clusters

Let us assume the different clusters have uneven number of points as shown below.

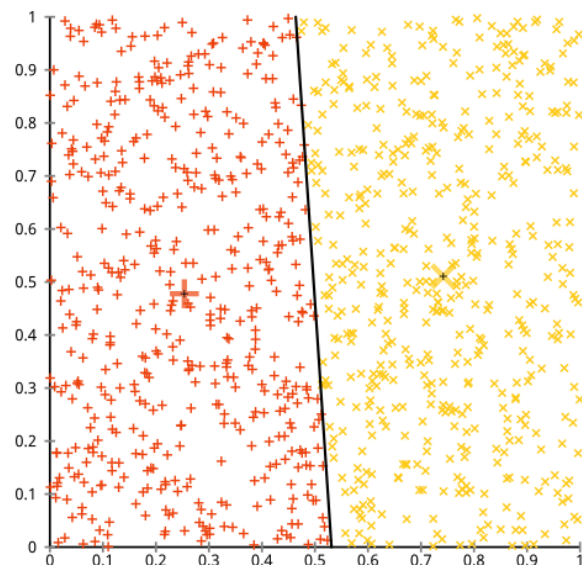


The result we get on running K-Means on this set of data is totally unexpected. In its quest to minimize the within-cluster sum of squares, the K-means algorithm gives more "weight" to larger clusters. In practice, that means it's happy to let that small cluster end up far away from any center, while it uses those centers to "split up" a much larger cluster.



- **Clustering non-clustered data**

Running k-means on uniform data also gives us clusters! So, it does not tell us when the data just does not cluster. We can see this in an example as shown below.



- **The data points are hard assigned to a cluster**

The data point is either in the cluster or it isn't. But sometimes, we are more confident about certain data points being in a cluster over others, i.e., in this method there is no uncertainty measure or probability that tells us how much a data point is associated with a specific cluster.

Now we move towards probabilistic models to help capture uncertainty in the data and to eliminate hard assignments.

3 Probability

Probability is simply *how likely something is to happen*. In probability terms, the "something" is generally referred to as an *event*.

An event is a set of outcomes of an experiment (a subset of the sample space) to which a probability is assigned.

Thus we can write

$$\text{Probability of an event happening} = \frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}}$$

Examples :

Example: the chances of rolling a "4" with a die

Number of ways it can happen: 1 (there is only 1 face with a "4" on it)

Total number of outcomes: 6 (there are 6 faces altogether)

$$\text{So the probability} = \frac{1}{6}$$

Example: there are 5 marbles in a bag: 4 are blue, and 1 is red. What is the probability that a blue marble gets picked?

Number of ways it can happen: 4 (there are 4 blues)

Total number of outcomes: 5 (there are 5 marbles in total)



$$\text{So the probability} = \frac{4}{5} = 0.8$$

Random Variable :

With probability, comes the concept of random variables. A random variable, usually written X , is a variable whose possible values are numerical outcomes of a random phenomenon.

Example: Tossing a coin: we could get Heads or Tails.

Let's give them the values **Heads=0** and **Tails=1** and we have a Random Variable "X":

<i>Random Variable</i>	<i>Possible Values</i>	<i>Random Events</i>
$X =$	{	0 ← 
	1 ← 	

In short:

$$X = \{0, 1\}$$

Note: We could choose Heads=100 and Tails=150 or other values if we want! It is our choice.

Example: How many heads when we toss 3 coins?



X = "The number of Heads" is the Random Variable.

In this case, there could be 0 Heads (if all the coins land Tails up), 1 Head, 2 Heads or 3 Heads.

So the Sample Space = {0, 1, 2, 3}

But this time the outcomes are NOT all equally likely.

The three coins can land in eight possible ways:

		$X =$ "Number of Heads"
HHH		3
HHT		2
HTH		2
HTT		1
THH		2
THT		1
TTH		1
TTT		0

Looking at the table we see just 1 case of Three Heads, but 3 cases of Two Heads, 3 cases of One Head, and 1 case of Zero Heads. So:

- $P(X = 3) = 1/8$
- $P(X = 2) = 3/8$
- $P(X = 1) = 3/8$
- $P(X = 0) = 1/8$

Types of Random Variables :

There are two types of random variables : **Discrete** and **Continuous**.

Discrete Random Variables

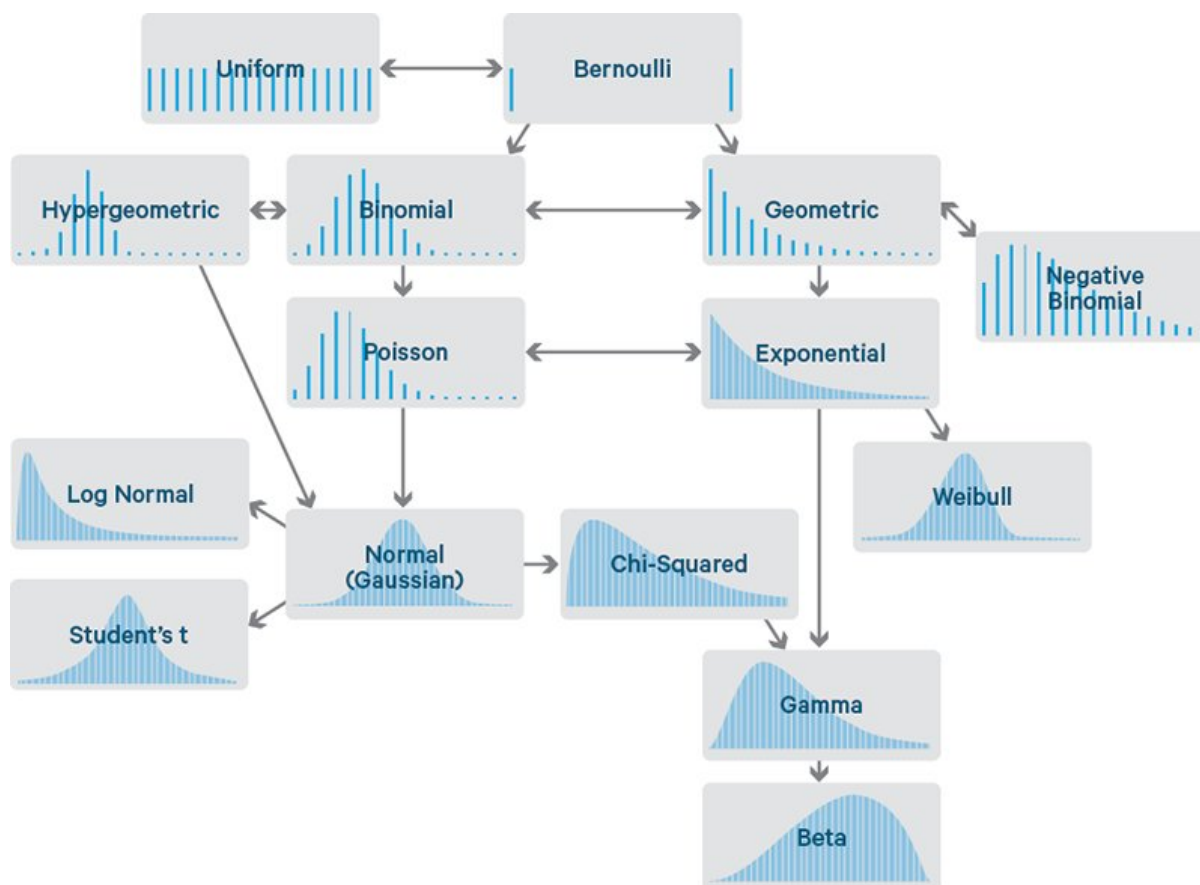
A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,... If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include : the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten, etc.

Continuous Random Variables

A continuous random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include : height, weight, the amount of sugar in an orange, the time required to run a mile, etc.

A continuous random variable is not defined at specific values. Instead, it is defined over an interval of values, and is represented by the area under a curve (in advanced mathematics, this is known as an integral). The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Some common probability distributions



Independent and Dependant events

Independent Events Two events are independent if the result of the second event is not affected by the result of the first event.

If A and B are independent events, the probability of both events occurring is the product of the probabilities of the individual events, i.e.,

$$P(A \text{ and } B) = P(A) \times P(B)$$

Example 1:

A box contains 4 red marbles, 3 green marbles and 2 blue marbles. One marble is removed from the box and then replaced. Another marble is drawn from the box. What is the probability that the first marble is blue and the second marble is green?

Because the first marble is replaced, the size of the sample space (9) does not change from the first drawing to the second so the events are independent.

$$\begin{aligned} P(\text{blue then green}) &= P(\text{blue}) \cdot P(\text{green}) \\ &= \frac{2}{9} \cdot \frac{3}{9} \\ &= \frac{6}{81} \\ &= \frac{2}{27} \end{aligned}$$

Dependent Events Two events are dependent if the result of the first event affects the outcome of the second event so that the probability is changed.

If A and B are dependent events, then the probability of both occurring is,

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

Probability of B given A

Example 2:

A box contains 4 red marbles, 3 green marbles and 2 blue marbles. One marble is removed from the box and it is not replaced. Another marble is drawn from the box. What is the probability that the first marble is blue and the second marble is green?

Because the first marble is not replaced, the size of the sample space for the first marble (9) is changed for the second marble (8) so the events are dependent.

$$\begin{aligned} P(\text{blue then green}) &= P(\text{blue}) \cdot P(\text{green}) \\ &= \frac{2}{9} \cdot \frac{3}{8} \\ &= \frac{6}{72} \\ &= \frac{1}{12} \end{aligned}$$

4 Deterministic vs Probabilistic models

Recall unsupervised learning

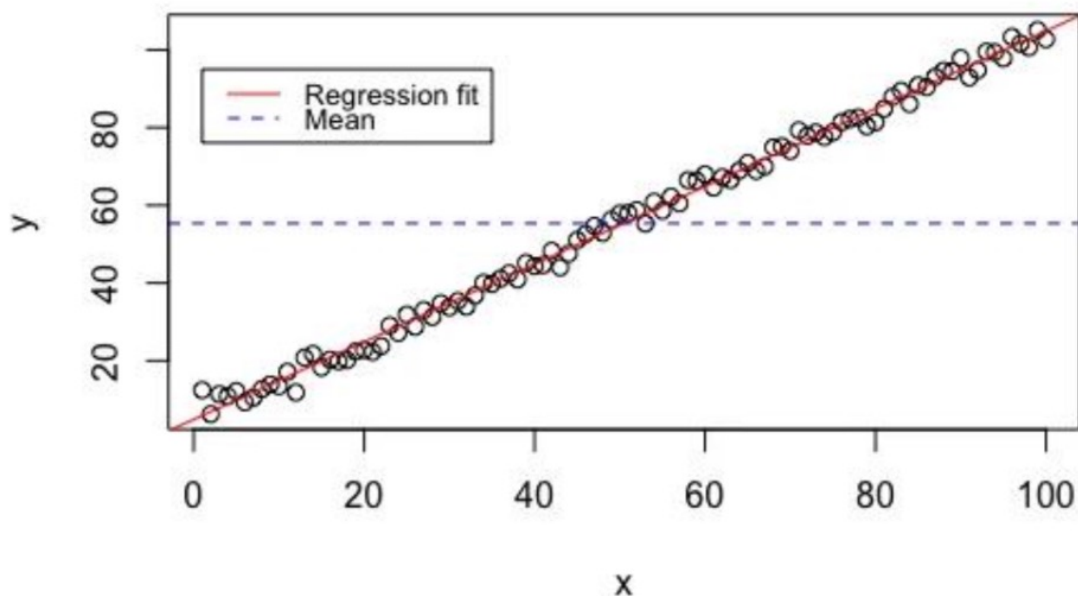
Task: Given $X \in \mathcal{X}$, learn $f(X)$.

We're essentially trying structure the given data. The function f can be either deterministic or probabilistic. Probabilistic models are better in certain cases because they help capture uncertainty.

Deterministic Models

Example :

$$f(x) = a_0 + a_1 * x$$



For each possible x , there is a single value y output by this function, hence deterministic.

Every time you run the model with the same initial conditions you will get the same results. Simple statistical statements, which do not mention or consider variation, could be viewed as deterministic models.

Probabilistic Models

Example : Sales volume (y) is 'about' 10 times advertising spending (x)

$$y = 10x + \epsilon$$

where ϵ is the predicted error to capture randomness. Such models help us capture the uncertainty.

Probabilistic Generative Models

In Probabilistic Generative models, we assume that the observed data is the realization of a distribution (which works out well in most real world scenarios).

The idea is to find that distribution function f in which the probability of occurrence of the given data is highest.

Lets try and understand the approach using an example.

Imagine we want to find a probabilistic model for this data. A coin is tossed 11 times and the output is recorded.

$$Data = HHTTHTHHTTT$$

If the probability of occurrence of a head in our model is p , the probability of occurrence of given data :

$$P(Data|p) = p * p * (1 - p) * (1 - p) * p * (1 - p) * p * p * (1 - p) * (1 - p)$$

$$L = P(Data|p) = p^5 * (1 - p)^6$$

The above term is known as likelihood. The probability of occurrence of the data, given the distribution (parameters). This the term we need to maximise to find the best distribution which can explain the given data.

what is the best value of p we can choose such that this equation gets the maximum value?

Let's see the difference between probability and likelihood, then differentiate the above equation.

What is the difference between likelihood and probability?

Probability is the percentage that a success can occur. For example, we do the binomial experiment by tossing a coin. We suppose that the event that we get the head of coin in success, so the probability of success now is 0.5 because the probability of head and tail of a coin is equal. 0.5 is the probability of a success. Likelihood is the conditional probability. With same example, we toss the coin 10 times and suppose that we get 7 successes (7 heads) and 3 failed (3 tails). The likelihood in this case is 0.1171 (assuming binomial distribution).

Meaning: 0.1171 is the probability that the above event will happen (7 successes out of 10 trials) by knowing that the probability of one success is 0.5 (single success).

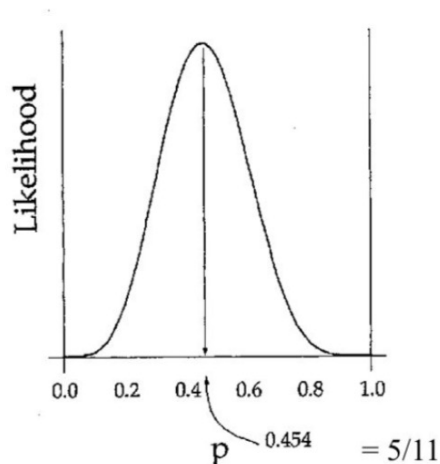
Likelihood is probability (conditional probability) of a set of successes, when the probability of a success is known. Probability is the percentage that a success can occur.

Maximum likelihood

Continuing with the equation in the previous section, differentiate L with respect to p and equate to 0.

$$\begin{aligned} dL/dp &= p^5 * 6 * (1 - p)^5 * (-1) + (1 - p)^6 * 5 * p^4 = 0 \\ 5 * p^4 * (1 - p)^6 &= 6 * p^5 * (1 - p)^5 \\ 5 * (1 - p) &= 6 * p \\ 5 &= 6 * p + 5 * p \\ p &= 5/11 \end{aligned}$$

The figure below shows how likely the given data distribution is, with respect to the value of p .



Log likelihood

The location of maxima doesn't change when we apply log to a term. Therefore we exploit this fact and make the computation easier by finding the maxima of the log of likelihood.

Depicting the application of log using the previous example:

$$\begin{aligned} \ln L &= \ln(p^5 * (1 - p)^6) \\ \ln L &= 5\ln(p) + 6\ln(1 - p) \\ d(\ln L)/dp &= 5/p + 6/1 - p = 0 \\ d(\ln L)/dp &= 5/p + 6/1 - p = 0 \\ 5 - 5p &= 6p \\ p &= 5/6 \end{aligned}$$

We arrived at the same value of maxima by differentiating the log likelihood.

Formalizing:

Likelihood can be written as

$$L(\theta) = P(\text{Data}|\theta)$$

What is the probability of occurrence of the given set of observations (data), in the distribution characterized by θ If each observation is independent, the formula can be further simplified as

$$L(\theta) = P(\text{Data}|\theta) = \prod_{i=1}^n P(X_i|\theta)$$

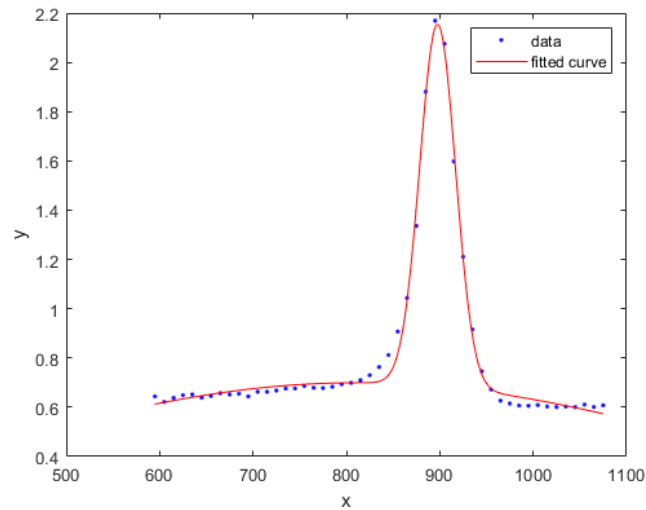
Goal is to obtain the distribution which will maximize the likelihood of the data. It can be represented mathematically as follows.

$$\theta^* = \arg \max_{\theta} P(\text{Data}|\theta)$$

5 Gaussian Distribution

1. Introduction

Assume we want to fit find the best Gaussian distribution for the data points represented below. How will we find the best mean and standard deviation?



Lets try and use maximum likelihood estimation to find the same.

2. Formula of Gaussian Distribution :

Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $-\infty < x < \infty$; $-\infty < \mu < \infty$; $\sigma > 0$

$f(x)$ → Normal Probability Distribution

x → random variable

μ → mean of distribution

σ → standard deviation of distribution

π → 3.14159

e → 2.71828

Maximum likelihood estimation of Gaussian parameters (unimodal)

The likelihood term is given as follows:

$$L = p(\mathbf{X} | \theta) = \mathcal{N}(\mathbf{X} | \theta) = \mathcal{N}(\mathbf{X} | \mu, \Sigma)$$

We need to find an optimal values for both mean and standard deviation

$$\mu_{MLE} = \underset{\mu}{\operatorname{argmax}} \mathcal{N}(\mathbf{X} | \mu, \Sigma)$$

$$\Sigma_{MLE} = \underset{\Sigma}{\operatorname{argmax}} \mathcal{N}(\mathbf{X} | \mu, \Sigma)$$

Lets maximize the log likelihood to find the best value of parameters. Remember we are besting a single best Gaussian to fit our data.

$$\begin{aligned} \sum_{n=1}^N \log(\mathcal{N}(\mathbf{x}_n | \mu, \sigma^2)) &= \sum_{n=1}^N \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp^{-\frac{1}{2} \left(\frac{(x_n - \mu)^2}{\sigma^2} \right)} \right) \\ \text{LL} &= \sum_{n=1}^N \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \left(\exp^{-\frac{1}{2} \left(\frac{(x_n - \mu)^2}{\sigma^2} \right)} \right) \right) \\ &= \sum_{n=1}^N \left(\log(1) - \log(\sqrt{2\pi\sigma^2}) + \log \left(\exp^{-\frac{1}{2} \left(\frac{(x_n - \mu)^2}{\sigma^2} \right)} \right) \right) \\ \text{LL} &= \sum_{n=1}^N \left(-\log(\sqrt{2\pi\sigma^2}) + \left(-\frac{1}{2} \left(\frac{(x_n - \mu)^2}{\sigma^2} \right) \right) \right) \\ &= \sum_{n=1}^N \left(-\frac{1}{2} \cdot \log(2\pi\sigma^2) - \frac{1}{2} \left(\frac{(x_n - \mu)^2}{\sigma^2} \right) \right) \end{aligned}$$

Simplifying the sum term

$$\begin{aligned} \text{LL} &= \sum_{n=1}^N \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2} \left(\frac{(x_n - \mu)^2}{\sigma^2} \right) \right) \\ &= -N \log(2\pi\sigma^2) * 1/2 + \sum_{n=1}^N -\frac{1}{2} \left(\frac{(x_n - \mu)^2}{\sigma^2} \right) \\ &= -N * \log(2\pi\sigma^2) * 1/2 - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 \end{aligned}$$

Paritally differentiate with respect to μ and equate to 0.

$$\begin{aligned} \frac{\partial \text{LL}}{\partial \mu} &= \frac{\partial}{\partial \mu} \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 \right) \\ &= \frac{\partial}{\partial \mu} \left(-\frac{N}{2} \log(2\pi\sigma^2) \right) + \frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 \right) \\ &= 0 + \frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 \right) \\ &= \frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 \right) \end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= \sum_{n=1}^N \frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma^2} \cdot (x_n - \mu)^2 \right) \\
&= \sum_{n=1}^N \left(\frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma^2} \right) \cdot (x_n - \mu)^2 + \left(-\frac{1}{2\sigma^2} \right) \cdot \frac{\partial}{\partial \mu} (x_n - \mu)^2 \right) \\
&= \sum_{n=1}^N \left(0 + \left(-\frac{1}{2\sigma^2} \right) \cdot \frac{\partial}{\partial \mu} (x_n - \mu)^2 \right) \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^N \frac{\partial}{\partial \mu} (x_n - \mu)^2
\end{aligned}$$

Now using chain rule

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= -\frac{1}{2\sigma^2} \sum_{n=1}^N \frac{\partial}{\partial \mu} (x_n - \mu)^2 \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^N 2(x_n - \mu) \cdot -1 \\
&= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu)
\end{aligned}$$

Equating to zero and obtain expression for μ

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu) \\
0 &= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu) \\
0 &= \sum_{n=1}^N (x_n - \mu) \\
0 &= \sum_{n=1}^N x_n - \sum_{n=1}^N \mu \\
0 &= \sum_{n=1}^N x_n - N \cdot \mu \\
N \cdot \mu &= \sum_{n=1}^N x_n \\
\mu &= \frac{1}{N} \sum_{n=1}^N x_n
\end{aligned}$$

This results tells us that the mean of the best gaussian which will fit the given data is the mean of the data itself!

Similarly when we differentiate the log likelihood WRT variance, we obtain the following result.

$$\Sigma_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\text{T}}$$

This is actually the formula of co-variance of the given data (sample co-variance).

Maximum Likelihood solution

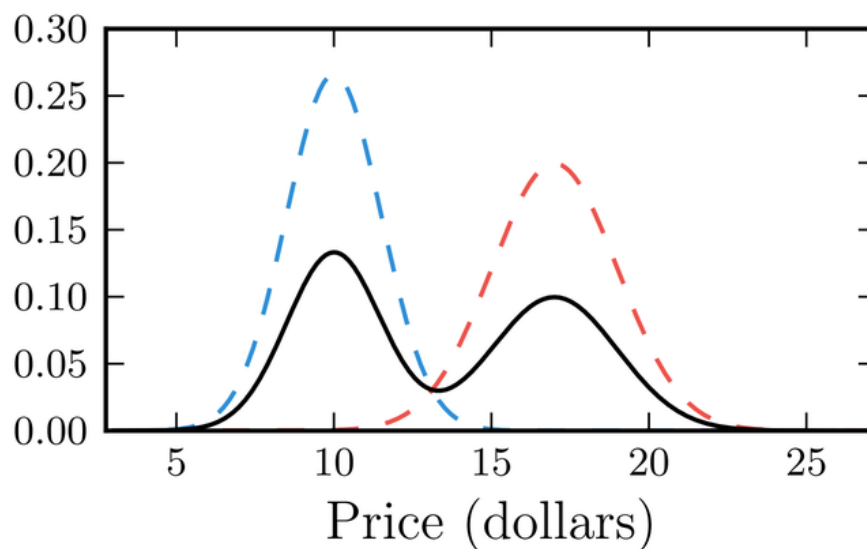
If above mentioned solution is not enough more details is given in the below link: you can refer this for better understanding

All computational details mentioned in the link given below

<http://jrmeyer.github.io/machinelearning/2017/08/18/mle.html>

Need for Gaussian mixture models

For example, suppose the price of a randomly chosen paperback book is normally distributed with mean \$ 10.00 and standard deviation \$ 1.00. Similarly, the price of a randomly chosen hardback is normally distributed with mean \$ 17 and variance 1.50. Is the price of a randomly chosen book normally distributed? The answer is no. Intuitively, we can see this by looking at the fundamental property of the normal distribution: it's highest near the center, and quickly drops off as you get farther away. But, the distribution of a randomly chosen book is bimodal: the center of the distribution is near \$13, but the probability of finding a book near that price is lower than the probability of finding a book for a few dollars more or a few dollars less. This is illustrated in Figure a.

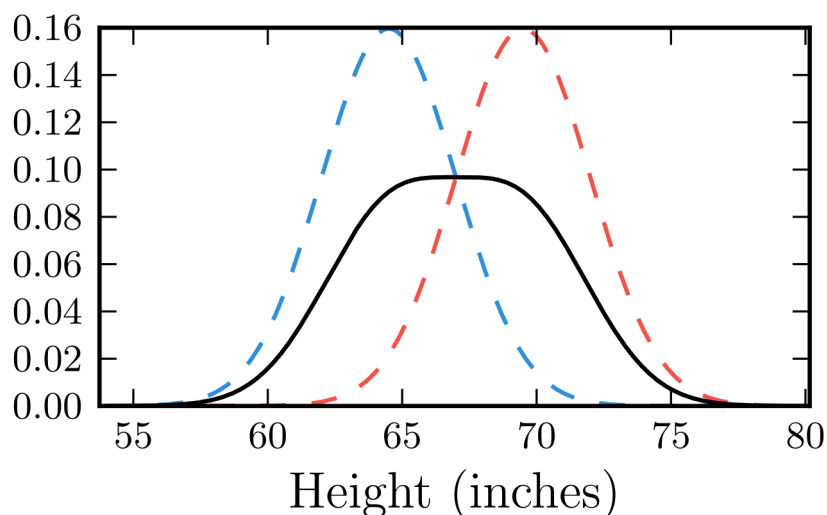


(a) Probability density for paperback books (red),hardback books (blue), and all books (black, solid)

Figure a: Two Gaussian mixture models: the component densities (which are Gaussian) are shown in dotted red and blue lines, while the overall density (which is not) is shown as a solid black line. the data within each group is normally distributed.

Another example: the height of a randomly chosen man is normally distributed with a mean around 509.5" and standard deviation around 2.5". Similarly, the height of a randomly chosen woman is normally distributed with a mean around 504.5" and standard deviation around 2.5" 1 Is

the height of a randomly chosen person normally distributed? The answer is again no. This one is a little more deceptive: because there's so much overlap between the height distributions for men and for women, the overall distribution is in fact highest at the center. But it's still not normally distributed: it's too wide and flat in the center (we'll formalize this idea in just a moment). This is illustrated in Figure 1b. These are both examples of mixtures of Gaussians: distributions where we have several groups.



(b) Probability density for heights of women (red), heights of men (blue), and all heights (black, solid)

Multivariate Gaussian Distribution

what is covariance ?

The concept of the covariance matrix is vital to understanding multivariate Gaussian distributions. Recall that for a pair of random variables X and Y , their covariance is defined as $\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$. When working with multiple variables, the covariance matrix provides a succinct way to summarize the covariances of all pairs of variables. In particular, the covariance matrix, which we usually denote as Σ , is the $n \times n$ matrix whose (i, j) th entry is $\text{Cov}[X_i, X_j]$.

Equation for Multivariate Gaussian Distribution:- The multivariate normal distribution is a multidimensional generalisation of the one-dimensional normal distribution. It represents the distribution of a multivariate random variable that is made up of multiple random variables that can be correlated with each other.

Like the normal distribution, the multivariate normal is defined by sets of parameters: the mean vector μ , which is the expected value of the distribution; and the covariance matrix Σ , which measures how dependent two random variables are and how they change together. We denote the covariance between variable X and Y as $\text{Cov}[X, Y]$. The multivariate normal with dimensionality d has a joint probability density given by:

Given $\mathbf{x} = (x_1, \dots, x_p)'$ with $x_j \in \mathbb{R} \forall j$, the **multivariate normal** pdf is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \quad (12)$$

where

- $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$ is the $p \times 1$ **mean vector**
- $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$ is the $p \times p$ **covariance matrix**

Write $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote \mathbf{x} is multivariate normal.

Gaussian Mixture Distribution Let's consider that we have K gaussian components, so a linear superposition of K gaussian component can be represented as follows

$$p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- **Log likelihood**

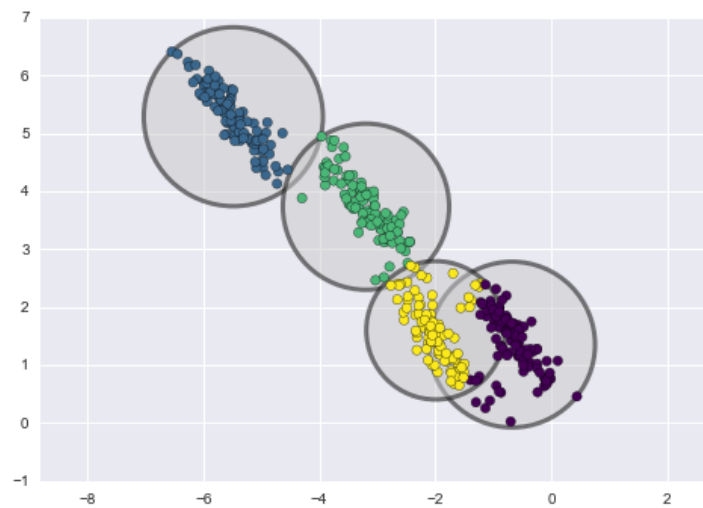
$$\ln p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

The summation term creates a problem! We cannot find the log likelihood easily as done in case of unimodal Gaussian (page 13). We will deal with this in detail in the next lecture.

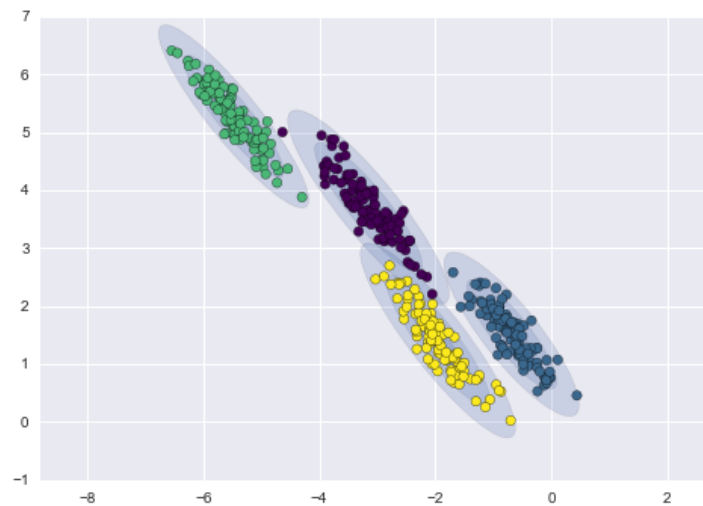
Recall the bigger picture

- What was the need for probabilistic models? Uncertainty in the data. Let's say we are clustering pictures of sky and forest, what if a picture has both? We want to be able to capture such uncertainty in the data, factor in that uncertainty in the decision making of the model.
- KNN tends to favor spherical clusters, GMM manages to eliminate that bias (examples images shown below).

Using KNN



Using GMM



References

- [1] <https://towardsdatascience.com/k-means-clustering-8e1e64c1561c>
- [2] <https://www.kaggle.com/dfoly1/k-means-clustering-from-scratch>
- [3] <https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means>

- [4] More Math in Latex, Link: https://ctan.math.illinois.edu/info/Math_into_LaTeX-4/Short_Course.pdf
- [5] <https://www.quora.com/What-is-the-difference-between-probability-and-likelihood-1>
- [6] <https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e>