CSE/ECE-471: Statistical Methods in AI **Spring 2020**

Limitations of K-Means, Likelihood, GMM

Prepared by: Mudit Malpani(2019201063), Sanjana Sunil (20171027), Surbhi (2019202002)

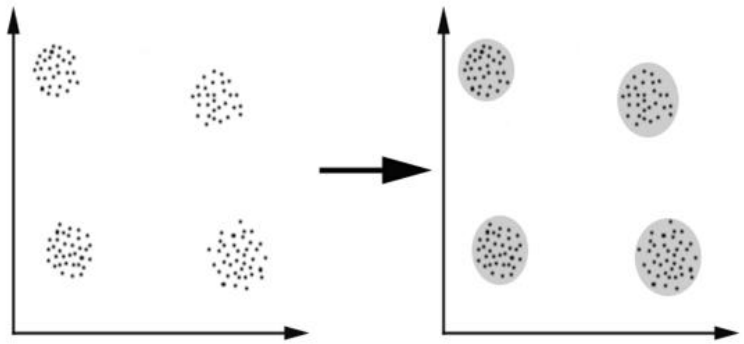
In this note, we discuss limitations of k-means algorithm, probabilistic models, likelihood and GMM.

Contents

1 K Means Algorithm

1.1 Introduction to Clustering

Clustering is an instance of unsupervised learning used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different.



Clustering as a summary of input data

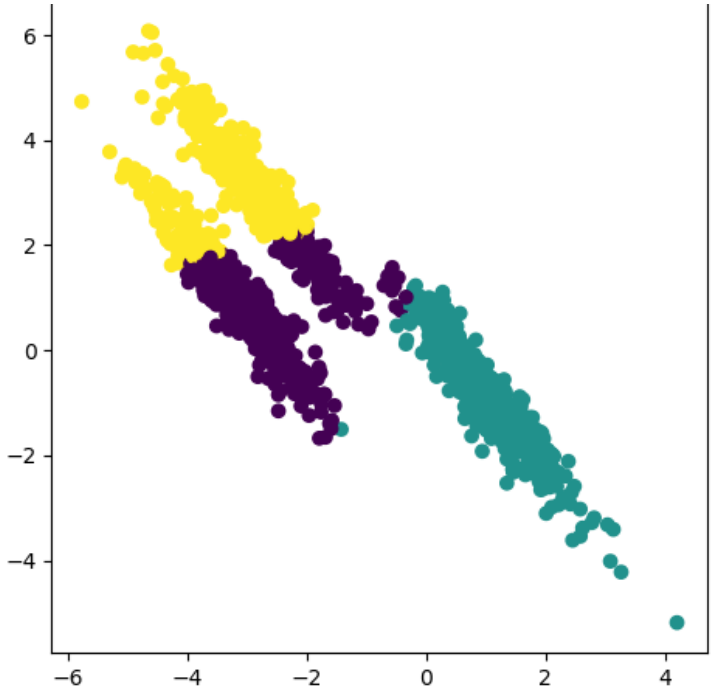
One of the main reasons to perform clustering is to provide a summary of the data.

1.2 Algorithm

K Means algorithm tries to partition the data set into K predefined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. 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.

1.3 Limitations of K-means

- The value of K needs to be chosen.
- In K-Means, we take Euclidean distance, so it prefers spherical cluster boundaries even though that might not provide the best clustering.

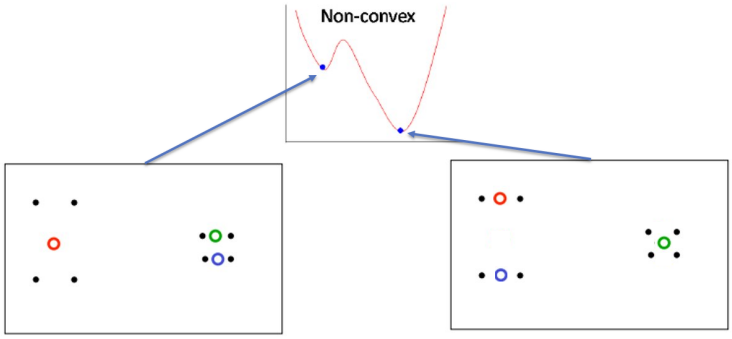


Euclidean distance- spherical cluster boundaries

- If x_i, $i = 1, 2, ..., n$ are the data points and μ_j, $j = 1, 2, ..., k$ are the mean values or the centers, then the objective function of K-means algorithm is to minimize the sum of distances between the data point and the centers, that is minimize the following function:

$$\sum_{i=1}^{n} \min_{j=1,2,...,k} \|x_i - \mu_j\|^2$$

  However, this is a non-convex function which means it can have multiple local minima which K-means algorithm might get stuck at.



Two different minimas of the non-convex function and the different end results

  This is dependent on the initialization of centers. We can use K-means++ algorithm to improve this initialization.

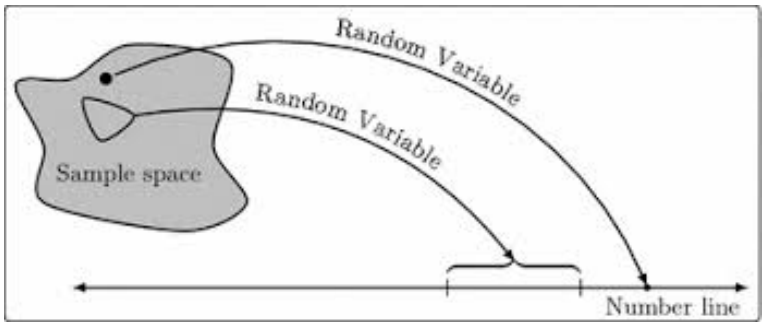
- K-means algorithm does hard clusterings, that is it either falls into a particular cluster or it does not. There is no in between. This might not always be ideal. For example, consider the clustering of movies into different genres. One movie could fall into different genres.

- Algorithm fails for non-linear data.

2 Basics of Probability

2.1 Random variable

Random variable is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.



Random variables

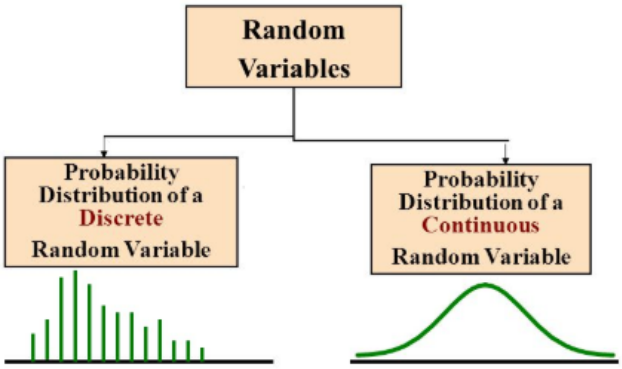
2.1.1 Discrete random variable

A discrete random variable is one which may take on only a countable number of distinct values. 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 probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function *e.g.* Discrete Uniform Distribution, Binomial Distribution etc.

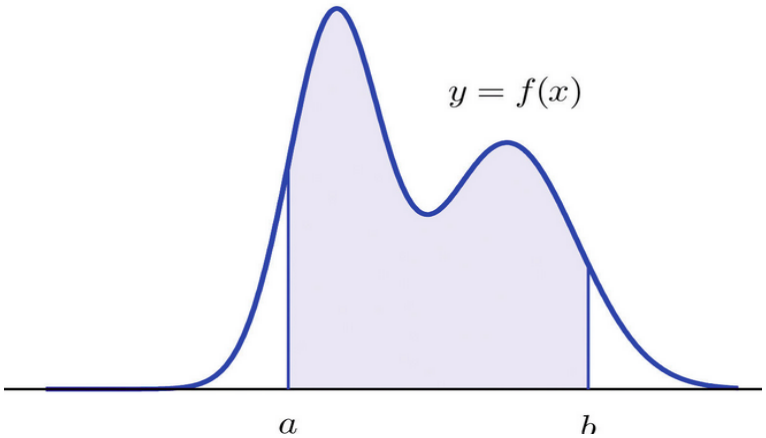
2.1.2 Continuous random variable

A continuous random variable is one which takes an infinite number of possible values. A continuous random variable is not defined at specific values. Instead, it is defined over an interval of values and its probability distribution represented by the area under a curve *e.g.* Normal Distribution.



In a discrete probability distribution, we can calculate the value of probability at a particular point, for example probability that a random variable X equals 2 or $Pr(X = 2)$. However, we cannot do this

for a continuous distribution since this would tend to zero. So we take the probability over an area for example for example, $Pr(2 - \delta <= X < -2 + \delta)$.



The area of the shaded region represents probability $Pr(a < X < b)$. $y = f(x)$ is called probability density function or pdf.

2.2 Random Vector

Random vector is a n-dimensional vector and it consists of n random variables.

2.3 Probability Events

Independent events are events that do not affect the outcome of one other. If events A and B are independent, then the probability of both occurring is $P(A \cap B) = P(A)P(B)$. For example, tossing a coin two times in a row. Each event is independent.

Dependent events are those in which the outcome of one event affects the outcome of the other. If events A and B are dependent, the probability of both occurring is $P(A \cap B) = P(A)P(B|A)$. For example, probability of drawing 4 kings in a row from a deck of cards without replacing them. Each event is dependent on what was previously drawn.

3 Probabilistic Models

In unsupervised learning, given $X \in \mathcal{X}$, we need to learn $f(X)$. This $f(X)$ can be either deterministic or probabilistic.

3.1 Deterministic vs. Probabilistic Model

In deterministic models, there are no uncertainties and a given input will always produce the same output. It hypothesizes an exact relationship between variables. For example, the conversion between Celsius and Fahrenheit follow an exact formula and is a deterministic model.

Probabilistic models however model uncertainty and gives a distribution over all possible outcomes. The model can give different results even with the same initial conditions due to some randomness. For example, the sales of a company increases when they show more advertisements. This could be an almost linear relationship, but not exactly, due to the presence of various other random factors.

3.2 Parameters of a model

Parameters, represented by θ, defines what a model looks like. For example, if we had a linear model $y = mx + c$, if x is the number of advertisements and y is the sales by the company, then m and c are the parameters of the linear model.

Consider another example where a coin is tossed and either heads or tails can come up. The parameters for this model would be p which could represent the probability of getting heads, and $(1 - p)$ probability of getting tails.

Gaussian distribution has two parameters - the mean μ and the standard deviation σ.

3.3 Likelihood

We assume that the observed data is the realization of some probabilistic model. Likelihood, denoted by $L(\theta)$, measures how much some particular values of parameters θ of the model can support the observed data. It is the values of the parameters that maximise the probability of some data occurring.

$$L(\theta) = Pr(Data|\theta)$$

If each observation X_i of the data is independent, and n is the total number of observations, the likelihood would become:

$$\begin{aligned} L(\theta) &= Pr(Data|\theta) \\ &= \prod_{i=1}^{n} Pr(X_i|\theta) \end{aligned}$$

For example, say we are tossing a coin and we don't know if it's fair or not. Let the probability of heads occurring be p. This is the parameter of the model. Now, say we tossed the coin 7 times and obtained the following data $HHHTTHH$. Since each of these events are equally likely, we get the following likelihood.

$$L(p) = p^5 * (1 - p)^2$$

3.3.1 Maximum Likelihood Estimation

Maximum likelihood estimation is the process of finding the parameters of the model, such that it fits the data the best. This is equivalent to finding the maximum likelihood. Let θ^* be the values of the parameters that maximise the likelihood.

$$\theta^* = \underset{\theta}{\text{argmax}} Pr(Data|\theta)$$

To find the maxima of the likelihood function, we can use differentiation. If there were multiple parameters, we would partially differentiate with respect to each parameter.

$$L'(\theta) = 0$$

Continuing on with the previous coin tossing example, we can find the maxima of $L(p) = p^5 * (1-p)^2$ by taking it's derivative with respect to parameter p and we get the value of p^* to be $\frac{5}{7}$.

3.3.2 Log likelihood

Until now, we took the derivative of product of terms, which can be computationally expensive. To avoid this, we instead take the logarithm on both sides of the equation before taking the derivative, so that the product of terms get transformed into a sum of terms.

Note that applying logarithm will still not change the maxima of the likelihood function because logarithm is a monotonously increasing function. To see this more clearly, let θ^* be the values of the parameters that maximise the likelihood, that is $L(\theta^*) > L(\theta)$ for all other values of θ. A monotonous function has the property that if $x > y$, then $f(x) > f(y)$. This implies that $f(L(\theta^*)) > f(L(\theta))$. Therefore, maximising the likelihood is equivalent to maximising the log likelihood.

In the coin tossing example, we can take the logarithm on both sides of the equation and then take the derivative to get the maxima.

$$\begin{aligned} \ln L(p) &= \ln(p^5 * (1-p)^2) \\ &= 5\ln p + 2\ln(1-p) \\ \frac{\mathrm{d}\ln L(p)}{\mathrm{d}p} &= \frac{5}{p} - \frac{2}{(1-p)} = 0 \end{aligned}$$

We get the same value of p^* to be $\frac{5}{7}$.

3.3.3 Maximum Likelihood Estimation of Gaussian Parameters

Recall that the Gaussian distribution is as follows:

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

If the observations are independent, we can calculate the likelihood as follows.

$$\begin{aligned} L(\mu,\sigma) &= \prod_{i=1}^{n} Pr(x_i|\mu,\sigma) \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{i=1}^{n} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \end{aligned}$$

Taking the log likelihood, we get:

$$\ln L(\mu, \sigma) = \ln \frac{1}{(\sigma\sqrt{2\pi})^n} + \sum_{i=1}^{n} \frac{-(x_i - \mu)^2}{2\sigma^2}$$

To get the maximum likelihood, we need to partially differentiate with respect to each of μ and σ.

- Differentiating with respect to μ

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = 0$$
$$\implies 0 + \sum_{i=1}^{n} \frac{-1}{2\sigma^2} \frac{\partial (x_i - \mu)^2}{\partial \mu} = 0$$

  Solving, we get the value of parameter μ that gives the maximum likelihood.

$$\mu^* = \frac{\sum_{i=1}^{n} x_i}{N}$$

  μ^* is called the sample mean.

- Differentiating with respect to σ

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} = 0$$
$$\implies \frac{-N}{\sigma} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^3} = 0$$

  Solving, we get the value of parameter σ that gives the maximum likelihood.

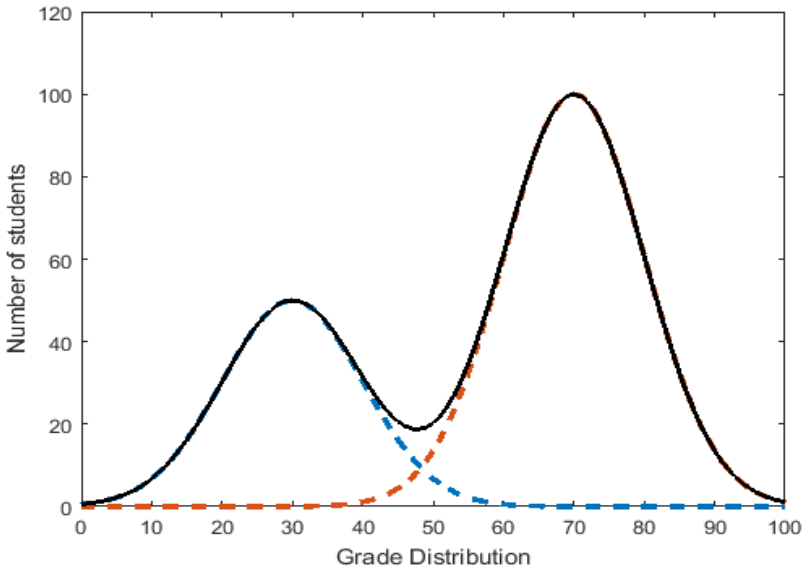
$$(\sigma^2)^* = \frac{\sum_{i=1}^{n} (x_i - \mu^*)^2}{N}$$

  $(\sigma^2)^*$ is the sample variance.

4 Gaussian Mixture Model

4.1 Multimodal Distribution

Data having continuous probability distribution with multiple modes/peaks is said to have Multimodal Distribution. The multimodality may indicate that the sample is not homogeneous and the observations come in fact from two or more "overlapping" distributions. For example consider this graph of Grade Distributions vs Number of students. We observe 2 peaks. One at 30 and other at 70. This suggests 2 distributions added together with mean as 30 and 70.

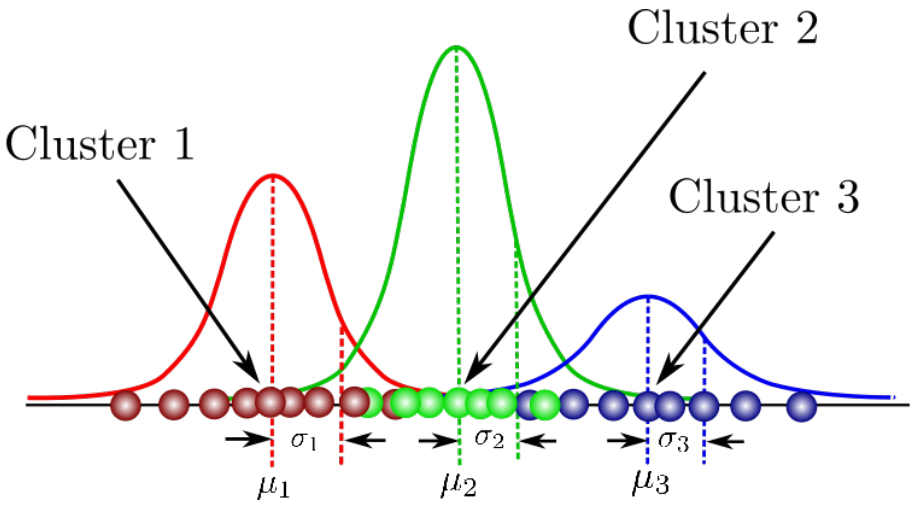


Mixture of two distributions

We will study one such distribution called Gaussian Mixture Model, in detail.

4.2 Gaussian Mixture Model

One of the most common Multimodal distributions is Gaussian Mixture Model. Its a group of normal distributions added together.



Mixture of three normal distributions

K is the total number of clusters. Each k^{th} cluster is comprised of the following:

- A mean μ that defines its centre.
- A covariance σ that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario.
- A mixing probability π that defines how big or small the Gaussian function will be. In other words its proportional to the number of points in the given cluster.

Single Gaussian Distribution:

$$\mathcal{N}(x|\mu_1, \sigma_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}}$$

The convex combinations for K Gaussian Distributions is the sum of the individual Gaussian distributions multiplied by their mixing probability or importance:

$$Pr(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k)$$

Normalization Condition:

$$\sum_{k=1}^{K} \pi_k = 1$$

4.3 Maximum Likelihood Estimation

4.3.1 Likelihood of Gaussian Mixture Model

As seen before, the likelihood for independent events will be the product of the probabilities for each data point. If there are n data points, then the likelihood of a Gaussian Mixture Model is:

$$L(\pi, \mu, \sigma) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \sigma_k)$$

Subject to:

$$\sum_{k=1}^{K} \pi_k = 1$$

4.3.2 Log Likelihood

Just like above, we do a log of this likelihood to make our computation relatively easier.

$$\ln L(\pi, \mu, \sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \sigma_k) \right)$$

Since it contains logarithm of sum of terms, its difficult to solve it in this form. We somehow need product of terms.

This is discussed in detail in the next class but we shall present a brief idea here. We get a product of terms this by introducing another latent variable z_{nk}. This variable should be 1 when data point x_n comes from cluster k, otherwise 0. If this information is known, the probability can be written as:

$$Pr(X, Z|\pi, \mu, \sigma) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \sigma_k)^{z_{nk}}$$

Taking log again:

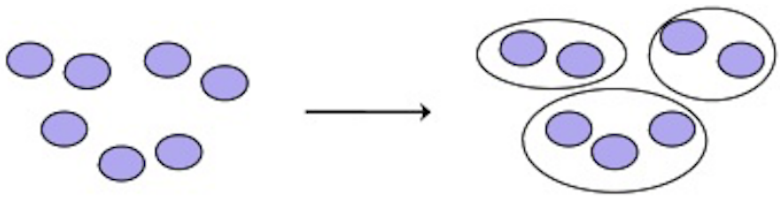
$$\ln Pr(X, Z|\pi, \mu, \sigma) = \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \left(\pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \sigma_k)^{z_{nk}}\right)$$

This becomes much easier to solve for maximum value after further simplification.

4.3.3 Brief Introduction to EM Algorithm

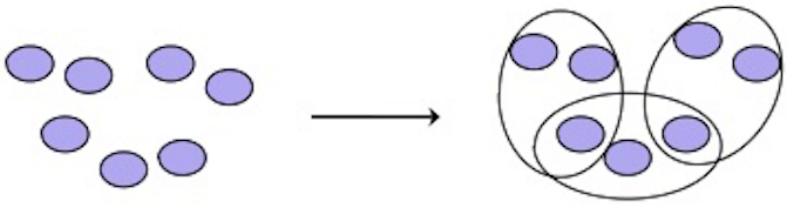
EM algorithm is done because we do not yet know the latent variables and need to find it.

With z_{nk} being either 0 or 1, its similar to how we did in case of K-means deterministic clustering. This is called *hard clustering*.



Hard Clustering

Here we pick up a slightly different approach. First we take random guesses for mean, variance and mixing weights of the clusters. Now lets introduce another variable $y_k(x_n)$ as the *expected value* of the latent variables which is a continuous value between 0 and 1. This is called *soft clustering*.



Soft Clustering

Then we update our mean, variance and mixing weights based on the expectation and get the new estimate of clusters, and then repeat the procedure just like with K-means. Hence EM algorithm is a multi-step process and improves/builds up on the last iteration. Finally we end up the probabilities of each data point belonging to each cluster.

This is a very short summary of how the EM or Expectation Maximization algorithm essentially works.

References

[1] An Introduction to Probabilistic modeling, Link: https://ethz.ch/content/dam/ethz/special-interest/bsse/borgwardt-lab/documents/slides/CA10_probabilitytheory.pdf

[2] McClave, James T., et al. Statistics for business and economics. Boston: Pearson, 2014.

[3] https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

[4] http://www.stat.yale.edu/Courses/1997-98/101/ranvar.htm

[5] https://scikit-learn.org/stable/modules/mixture.html

[6] https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95