COL870: Clustering Algorithms

Ragesh Jaiswal, CSE, IIT Delhi

Mixture Models

Mixture Models

- In many scenarios, it makes sense to make assumptions on how the data is generated.
- For clustering, certain probabilistic models for generating clustering instances have been studied.
- The goal is to cluster data instance that has been generated from the particular model.
- The most well studied probabilistic model is the *Gussian Mixture Model* (GMM).

- A univariate Gaussian random variable X with mean μ and variance σ^2 has the probability density function

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

- Similarly, a multivariate Gaussian r.v. $\mathbf{X} \in \mathbb{R}^n$ has the density function

$$f(\mathbf{X}) = \frac{1}{|\Sigma|^{1/2}(\sqrt{2\pi})^{n/2}} e^{\frac{-(\mathbf{X}-\mu)^T\Sigma^{-1}(\mathbf{X}-\mu)}{2}}$$

  Here $\mu \in \mathbb{R}^n$ is the mean vector and Σ is the $n \times n$ covariance matrix.

- A special case is the spherical Gaussian for which $\Sigma = \sigma^2 I_n$.
- A Gaussian distribution is denoted by $\mathcal{N}(\mu, \Sigma)$.

- A univariate Gaussian random variable X with mean μ and variance σ^2 has the probability density function

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

- Similarly, a multivariate Gaussian r.v. $\mathbf{X} \in \mathbb{R}^n$ has the density function

$$f(\mathbf{X}) = \frac{1}{|\Sigma|^{1/2}(\sqrt{2\pi})^{n/2}} e^{\frac{-(\mathbf{X}-\mu)^T\Sigma^{-1}(\mathbf{X}-\mu)}{2}}$$

  Here $\mu \in \mathbb{R}^n$ is the mean vector and Σ is the $n \times n$ covariance matrix.
- A special case is the spherical Gaussian for which $\Sigma = \sigma^2 I_n$.
- A Gaussian distribution is denoted by $\mathcal{N}(\mu, \Sigma)$.
- A Gaussian mixture model $\mathcal{M}$ is a distribution obtained from a convex combination of a bunch of such gaussians $\mathcal{N}(\mu_1, \Sigma_1), ..., \mathcal{N}(\mu_k, \Sigma_k)$. This is denoted by

$$\mathcal{M} = w_1\mathcal{N}(\mu_1, \Sigma_1) + ... + w_k\mathcal{N}(\mu_k, \Sigma_k)$$

  where $w_i \geq 0$ are called the *mixing weights* and satisfy $\sum_i w_i = 1$.

- A Gaussian mixture model $\mathcal{M}$ is a distribution obtained from a convex combination of a bunch of such gaussians $\mathcal{N}(\mu_1, \Sigma_1), ..., \mathcal{N}(\mu_k, \Sigma_k)$. This is denoted by

$$\mathcal{M} = w_1 \mathcal{N}(\mu_1, \Sigma_1) + ... + w_k \mathcal{N}(\mu_k, \Sigma_k)$$

  where $w_i \geq 0$ are called the *mixing weights* and satisfy $\sum_i w_i = 1$.
- This basically means that a a point from $\mathcal{M}$ is sampled by first choosing a component Gaussian i with probability w_i and then generating a point from $\mathcal{N}(\mu_i, \Sigma_i)$.

GMM Clustering Problem

Given a dataset with m points coming from such a GMM, recover the individual components of the mixture model.

- Partition the given points such that points drawn from the same Gaussian are in the same cluster.

GMM Clustering Problem

Given a dataset with m points coming from such a GMM, recover the individual components of the mixture model.

- Partition the given points such that points drawn from the same Gaussian are in the same cluster. Note that only probabilistic guarantees are possible where the probability is over the sample of m points.
- Another natural problem in the GMM setting is to give an estimate of the parameters of the individual component Gaussians.
- This is known as the *Parameter Estimation Problem*.
- Claim: Given a solution for one problem, the other problem may be solved and vice versa.

GMM Clustering Problem

Given a dataset with m points coming from such a GMM, recover the individual components of the mixture model.

- Suppose (for simplicity) we are given k spherical Gaussians in $\mathbb{R}^n$ with means $\mu_1, ..., \mu_k$ and variance $\Sigma = \sigma^2 I_n$.
- Under what conditions, clustering possible?

GMM Clustering Problem

Given a dataset with m points coming from such a GMM, recover the individual components of the mixture model.

- Suppose (for simplicity) we are given k spherical Gaussians in $\mathbb{R}^n$ with means $\mu_1, ..., \mu_k$ and variance $\Sigma = \sigma^2 I_n$.
- Under what conditions, clustering possible?
- Some *centre separation* is required for clustering.
- Centre Separation: A mixture of k identical spherical Gaussians satisfies c-separation if $\forall i \neq j$, $\Delta_{i,j} = ||\mu_i - \mu_j|| > c\sigma\sqrt{n}$.
- For what values of c can we come up with good clustering algorithms?
- Let x_1 and x_2 be two points sampled from the same Gaussian $\mathcal{N}(\mu_i, \sigma^2 I_n)$. What is $\mathbf{E}[||x_1 - x_2||^2]$?
- Let x and y be two points sampled from the Gaussians $\mathcal{N}(\mu_i, \sigma^2 I_n)$ and $\mathcal{N}(\mu_j, \sigma^2 I_n)$ respectively. What is $\mathbf{E}[||x - y||^2]$?

GMM Clustering Problem

Given a dataset with m points coming from such a GMM, recover the individual components of the mixture model.

- Suppose (for simplicity) we are given k spherical Gaussians in $\mathbb{R}^n$ with means $\mu_1, ..., \mu_k$ and variance $\Sigma = \sigma^2 I_n$.
- Centre Separation: A mixture of k identical spherical Gaussians satisfies c-separation if $\forall i \neq j$, $\Delta_{i,j} = ||\mu_i - \mu_j|| > c\sigma\sqrt{n}$.
- For what values of c can we come up with good clustering algorithms?
- Let x_1 and x_2 be two points sampled from the same Gaussian $\mathcal{N}(\mu_i, \sigma^2 I_n)$. What is $\mathbf{E}[||x_1 - x_2||^2]$?
- Let x and y be two points sampled from the Gaussians $\mathcal{N}(\mu_i, \sigma^2 I_n)$ and $\mathcal{N}(\mu_j, \sigma^2 I_n)$ respectively. What is $\mathbf{E}[||x - y||^2]$?
- Claim (informal): For $c = 2$, the Gaussians are well separated and can be clustered using some distance based algorithms.

GMM Clustering Problem

Given a dataset with m points coming from such a GMM, recover the individual components of the mixture model.

- Suppose (for simplicity) we are given k spherical Gaussians in $\mathbb{R}^n$ with means $\mu_1, ..., \mu_k$ and variance $\Sigma = \sigma^2 I_n$.
- Centre Separation: A mixture of k identical spherical Gaussians satisfies c-separation if $\forall i \neq j$, $\Delta_{i,j} = ||\mu_i - \mu_j|| > c\sigma\sqrt{n}$.
- For what values of c can we come up with good clustering algorithms?
- Claim (informal): For $c = 2$, the Gaussians are well separated and can be clustered using some distance based algorithms.
- $c = 1$ denotes the case when the Gaussians have significant overlap.
    - If we have points in the sample that close to the centers, then we might be able to cluster.
    - The issue is that it is highly unlikely that there will be a point closer than $1/2\sigma\sqrt{n}$ from the centre unless the sample size is $2^{O(n)}$.

GMM Clustering Problem

Given a dataset with m points coming from such a GMM, recover the individual components of the mixture model.

- Suppose (for simplicity) we are given k spherical Gaussians in $\mathbb{R}^n$ with means $\mu_1, ..., \mu_k$ and variance $\Sigma = \sigma^2 I_n$.
- Centre Separation: A mixture of k identical spherical Gaussians satisfies c-separation if $\forall i \neq j$, $\Delta_{i,j} = ||\mu_i - \mu_j|| > c\sigma\sqrt{n}$.
- For what values of c can we come up with good clustering algorithms?
- Claim (informal): For $c = 2$, the Gaussians are well separated and can be clustered using some distance based algorithms.
- $c = 1$ denotes the case when the Gaussians have significant overlap.
  - If we have points in the sample that close to the centers, then we might be able to cluster.
  - The issue is that it is highly unlikely that there will be a point closer than $1/2\sigma\sqrt{n}$ from the centre unless the sample size is $2^{O(n)}$.
  - One solution is to project the points to a random subspace of dimension $d << n$. Now we might need the sample to be of size only $2^{O(d)}$.

End