Gaussian Mixture Model, EM algorithm

Kairit Sirts

28.02.2014

K-means and Gaussians

- In K-means we attach each point to its closest centroid according to formula:

$$z_i = \arg\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2$$

- What we are really computing is:

$$\|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 = \sum_{j=1}^{d} (x_{ij} - \mu_{kj})(x_{ij} - \mu_{kj}) = (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\mathbf{x}_i - \boldsymbol{\mu}_k)$$

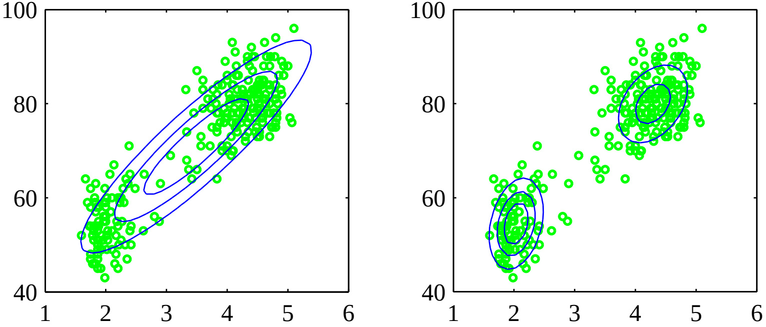
- Recall the formula for multivariate Gaussian:

$$P(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right]$$

- If we assume identity covariance $\Sigma = I$ then we are really computing Gaussian probabilities in K-means.

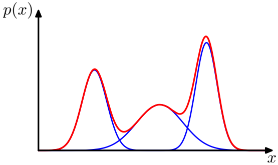
Multimodal Data

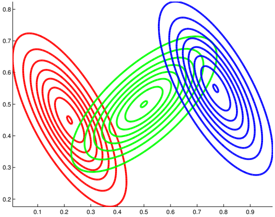
- Gaussian distribution is widely used in modeling, mainly because it has nice mathematical properties.
- In real life data is rarely Gaussian but several Gaussians might fit data quite well.

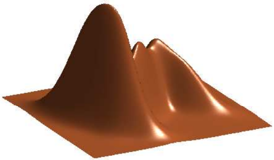


Gaussian Mixture Model

- Gaussian Mixture Model (GMM) is a linear superposition of several Gaussians.
- We introduce latent variables that indicate from wich mixture component each point comes from.
- The work with joint distribution over observed and latent variables is easier than with marginal distribution over data.







Gaussian Mixture Model

- There are K Gaussians **base** or **component distributions**:

$$p(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)$$

- and one **mixing distribution**, also called mixing coefficients:

$$\boldsymbol{\pi} : \quad \sum_{k=1}^{K} \pi_k = 1$$

- The probability of a point $\mathbf{x}_i$ is then:

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

Generative story

- **Latent variables** z_i: $z_i = k$ means component k generated point $\mathbf{x}_i$.
- Probability of being generated by a component:

$$p(z_i = k|\boldsymbol{\pi}) = \pi_k$$

- Probability of a point given we know wich component generated it:

$$p(\mathbf{x}_i|z_i = k, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)$$

- **Joint probability** of generating the component and the point from it:

$$\begin{aligned} p(\mathbf{x}_i, z_i = k|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= p(z_i = k|\boldsymbol{\pi})P(\mathbf{x}_i|z_i = k, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \pi_k\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) \end{aligned}$$

- **Marginal probability** of the point - sum out the components:

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

Inference

- We set up a **generative model** that can be used to generate data.
- But we observe only data.
- We need to learn model parameters - this is also called **inference**.
- Generation proceeds from parameters to data.
- Inference proceeds from data to parameters.

Estimating the parameters for GMM

- We need to estimate: $\boldsymbol{\pi}$, $\boldsymbol{\mu}_{\mathbf{k}}$, Σ_k, $k = 1 \dots K$
- The log-likelihood of GMM is:

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

- There are several difficulties in applying maximum likelihood framework directly to GMM:
  - Singularity: Fitting a component mean exactly on a data point leads likelihood to infinity.
  - Identifiability: K-component mixture has $K!$ equivalent solutions.
  - There is a summation inside the logarithm and thus setting derivatives of log-likelihood to zero will no longer give a closed form solution.

Iterative approach

- If we would know the component parameters and mixing proportions then we could compute the probability that the component k is responsible for the i-th point: $p(z_i = k|\mathbf{x}_i, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- If we would know the responsibilities then we could compute the estimates for mixing coefficients π_k.
- If we would know the responsibilities and mixing coefficients then we could compute the estimates for component means and variances $\boldsymbol{\mu}_k$ and Σ_k.

Expectation-Maximization

- The described iterative algorithm is often used for estimating the parameters of the models with latent variables.
- The general algorithm is called **expectation-maximization** and consists of two steps:
  - **Expectation** (E) step: compute the expected values for latent variables given some estimates for the parameters.
  - **Maximization** (M) step: maximize the parameters given the values of latent variables.
- It can be shown that EM algorithm monotonically increases the log likelihood of the observed data.

EM more formally

- Define **complete data log likelihood**:

$$\mathcal{L}_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}$$

- This cannot be computed as the latent variables z_i are unknown.
- Define **expected complete data log likelihood**:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) = E[\mathcal{L}_c(\boldsymbol{\theta}) | \mathbf{X}, \boldsymbol{\theta}^{t-1}]$$

- t is the current iteration number, Q is called **auxiliary function**.
- **E step** computes the latent values needed to compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$.
- **M step** optimizes Q with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{t} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$$

EM for GMM

- The expected complete data log likelihood is:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) &= \mathbb{E}\left[\sum_i \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta})\right] \\
&= \sum_i \mathbb{E}\left[\log\left[\prod_{k=1}^{K} (\pi_k p(\mathbf{x}_i | \boldsymbol{\theta}_k))^{\mathbb{I}(z_i = k)}\right]\right] \\
&= \sum_i \sum_k \mathbb{E}\left[\mathbb{I}(z_i = k)\right] \log\left[\pi_k p(\mathbf{x}_i | \boldsymbol{\theta}_k)\right] \\
&= \sum_i \sum_k p(z_i = k | \mathbf{x}_i, \boldsymbol{\theta}^{t-1}) \log\left[\pi_k p(\mathbf{x}_i | \boldsymbol{\theta}_k)\right] \\
&= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i | \boldsymbol{\theta}_k)
\end{aligned}
$$

- r_{ik} are the responsibilities and their values are latent.

E step for GMM

- We have to compute the values for the latent quantities in $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$
- Compute the rseponsibilities r_{ik} for each i and k:

$$r_{ik} = \frac{\pi_k p(\mathbf{x}_i | \boldsymbol{\theta}_k^{t-1})}{\sum_{k'} \pi_{k'} p(\mathbf{x}_i | \boldsymbol{\theta}_{k'}^{t-1})}$$

- Basically we compute the probability of point $\mathbf{x}_i$ being generated by a component and then normalize it with respect to all components.

M step for GMM

- Optimize Q with respect to $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$ and Σ_k.
- If $r_k = \sum_i r_{ik}$ is the weighted number of points assigned to cluster k:

$$\pi_k = \frac{r_k}{n}$$

- For $\boldsymbol{\mu}_k$ and Σ_k look only at the parts in Q that depend on them:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}_k, \Sigma_k) &= \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\theta}_k) \\ &= -\frac{1}{2} \sum_i r_{ik} \left[\log|\Sigma|_k + (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right]\end{aligned}$$

- Taking the derivatives with respect to each of them yields:

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{\sum_i r_{ik}\mathbf{x}_i}{r_k} \\ \Sigma_k &= \frac{\sum_i r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{r_k} = \frac{\sum_i r_{ik}\mathbf{x}_i\mathbf{x}_i^T}{r_k} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\end{aligned}$$

K-means and Gaussian Mixture Models

- K-means is essentially a Gaussian mixture model
- The covariances are set to the same symmetric matrix for each cluster:

$$\Sigma_1 = \cdots = \Sigma_K = \sigma^2 \mathbf{I}$$

- Mixing proportions are uniform: $\pi_k = \frac{1}{K}$
- Thus, only cluster means $\boldsymbol{\mu}_k$ must be estimated

K-means and Gaussian Mixture Models

- Consider delta-function approximation for responsibilities in E-step:

$$p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) \approx \mathbb{I}(z_i^* = k)$$
$$z_i^* = \arg\max_k p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})$$

- As the covariances are spherical and equal this reduces the E step to:

$$x_i^* = \arg\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2$$

- As the clustering is **hard** (due to delta approximation) we only have to compute regular average for means (instead of weighted average as in GMM) and the M step is:

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i:z_i=k} \mathbf{x}_i$$