Gaussian Mixture Models

30th March 2023

Introduction

- Gaussian Mixture Models is a "soft" clustering algorithm, where each point probabilistically "belongs" to all clusters.
- This is different than k-means where each point belongs to one cluster ("hard" cluster assignments).
- When we have log of a sum, there is no way to reduce it. This problem occurs within the log likelihood for GMM, so it is difficult to maximize the likelihood.
- The Expectation-Maximization (EM) procedure is a way to handle $\log \sum$.
- We can maximize the auxiliary function, which leads to an increase in the likelihood. We repeat this process at each iteration (constructing the auxiliary function and maximizing it), leading to a local maximum of the log likelihood for GMM.

Gaussian Mixture Model (GMM)

$P(z_i = k) = w_k$

$x_i \mid z_i = k \sim N(u_k, \Sigma_k)$

$\sum_k w_k = 1$

- Here is GMM's generative model:
  - First, generate which cluster i is going to be generated from:

$$z_i \mid \mathbf{w} \sim \text{ Categorical } (\mathbf{w})$$

which means that w_k is the probability that i 's cluster is k. That is,

$$P(z_i = k \mid \mathbf{w}) = w_k$$

Here, w_k are called the mixture weights, and they are a discrete probability distribution: $\sum_k w_k = 1, 0 \leq w_k \leq 1$.

  - Then, generate $\mathbf{x}_i$ from the cluster's distribution:

$$\mathbf{x}_i \mid z_i = k \quad \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Gaussian Mixture Model (GMM)

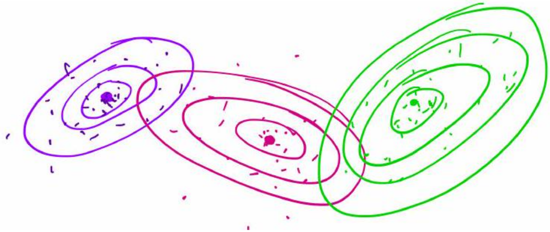
- Recap the notation:

$$\begin{aligned}
\mathbf{x}_i &\rightarrow \text{data} \\
z_i &\rightarrow \text{cluster assignment for } i \\
\boldsymbol{\mu} &\rightarrow \text{center of cluster } k \\
\boldsymbol{\Sigma}_k &\rightarrow \text{spread of cluster } k \\
w_k &\rightarrow \text{proportion of data in cluster } k \text{ (mixture weights)}
\end{aligned}$$

- The formula for the normal distribution:

$$p(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})\right)$$

- Here is a picture of the generative process,
- First generated the cluster centers and covariances, and then generated points for each cluster, where the number of points is proportional to the mixture weights.



Likelihood

- $$\text{likelihood} = P\left(\{\mathbf{X}_1, \ldots, \mathbf{X}_n\} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \mid \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

  where

  $\mathbf{w} = [w_1, \ldots, w_k], \boldsymbol{\mu} = [\mu_1, .., \mu_k], \boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k]$. .

- Denote the collection of these variables as θ.

$$\begin{aligned}
\text{likelihood}(\theta) &= \prod_i P\left(\mathbf{X}_i = \mathbf{x}_i \mid \theta\right) \\
&= \prod_i \sum_{k=1}^{K} P\left(\mathbf{X}_i = \mathbf{x}_i \mid z_i = k, \theta\right) P\left(z_i = k \mid \theta\right) \qquad \text{(law} \\
&= \prod_i \sum_{k=1}^{K} N\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) w_k
\end{aligned}$$

Likelihood

- Taking the $\log$,

$$\begin{aligned}
&\log \text{ likelihood } (\theta) \\
&= \log \prod_i \sum_k P\left(\mathbf{X}_i = \mathbf{x}_i \mid z_i = k, \theta\right) P\left(z_i = k \mid \theta\right) \\
&= \sum_i \log \sum_k P\left(\mathbf{X}_i = \mathbf{x}_i \mid z_i = k, \theta\right) P\left(z_i = k \mid \theta\right).
\end{aligned}$$

- We might think this problem is specific just to the one we're working on (Gaussian mixture models) but the problem is much more general!
- Every time we have a latent variable like $\mathbf{z}$, the same problem happens.
- This problem is rather difficult to be minimized directly!

Expectation Maximization

- EM creates an iterative procedure where we update the z_i' 's and then update $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\mathbf{w}$. It is an alternating minimization scheme similar to k-means.
  - E-step: compute cluster assignments (which are probabilistic)
  - M-step: update θ (which are the clusters' properties)
- Incidentally, if we looked instead at the "complete" $\log$ likelihood $p(\mathbf{x}, | \mathbf{z}, \theta)$ (meaning that you know the z_i' 's), there is no sum and the issue with the sum and the log goes away! This is because we no longer need to sum over k, we already know which cluster k unit i is in.

Expectation Maximization

- Let's start over from scratch. We are now in a very general setting. The data are still drawn independently, and each data has a hidden variable associated with it. Notation for data and hidden variables is:

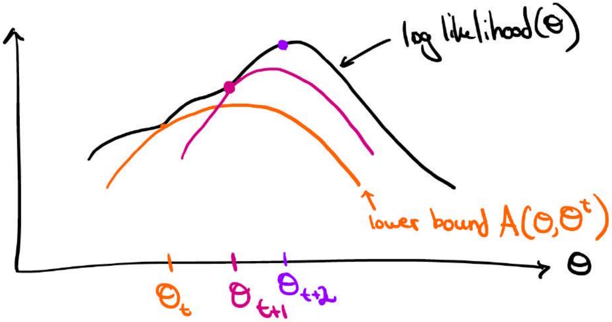
$$
\begin{aligned}
x_1, \ldots, x_n & \text{ data} \\
z_1, \ldots, z_n & \text{ hidden variables, taking values } k = 1 \ldots K \\
\theta & \text{ parameters}
\end{aligned}
$$

- Then,

$$
\begin{aligned}
\log \text{ likelihood } (\theta) &= \log P\left(X_1, \ldots, X_n = x_1, \ldots, x_n \mid \theta\right) \\
&= \sum_i \log P\left(X_i = x_i \mid \theta\right) \qquad \text{(by independence)} \\
&= \sum_i \log \sum_k P\left(X_i = x_i, Z_i = k \mid \theta\right) \qquad \text{(hidden variables)} \\
&= \sum_i \log \sum_k P\left(Z_i = k \mid \theta\right) P\left(X_i = x_i \mid Z_i = k, \theta\right)
\end{aligned}
$$

Expectation Maximization

- The idea of Expectation Maximization (EM) is to find a lower bound on likelihood (θ) that involves $P(\mathbf{x}, \mathbf{z} \mid \theta)$. Maximizing the lower bound always leads to higher values of $\text{likelihood}(\theta)$.
- The figure below illustrates a few iterations of EM.



Expectation Maximization

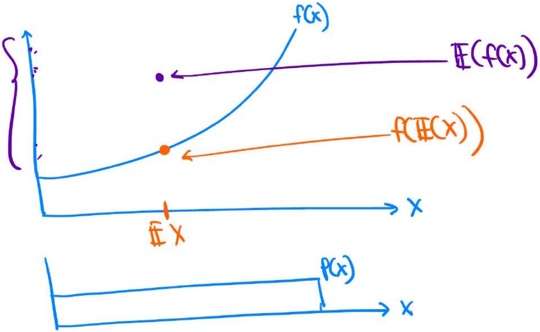
- Let us write out the procedure for constructing A, starting with the log likelihood.

$$\begin{aligned}\log \text{ likelihood } (\theta) &= \sum_i \log \sum_k P\left(X_i = x_i, Z_i = k \mid \theta\right) \qquad \text{(from above)} \\ &= \sum_i \log \sum_k P\left(Z_i = k \mid x_i, \theta_t\right) \frac{P\left(X_i = x_i, Z_i = k \mid \theta\right)}{P\left(Z_i = k \mid x_i, \theta_t\right)}\end{aligned}$$

- The weighted average $\sum_k P\left(Z_i = k \mid x_i, \theta_t\right) \langle \text{ stuff } \rangle$ can be viewed as an expectation because it's a sum of elements weighted by probabilities that add up to 1.
- We will call it $\mathbb{E}_z$.

$$\log \text{ likelihood } (\theta) = \sum_i \log \mathbb{E}_z \frac{P\left(X_i = x_i, Z_i = k \mid \theta\right)}{P\left(Z_i = k \mid x_i, \theta_t\right)}$$

- We will now use Jensen's inequality for convex functions, which allows us to switch a log and an expectation.



- Lemma (Jensen's Inequality). If f is convex, then $f(\mathbb{E}X) \leq \mathbb{E}(f(X))$.
- If f is convex, $-f$ is concave, thus $-f(\mathbb{E}X) \geq -\mathbb{E}(f(X)) = \mathbb{E}(-f(X))$. Here, $-f(x) = \log(x)$ which is concave, thus, $\log(\mathbb{E}X) \geq \mathbb{E}\log X$.

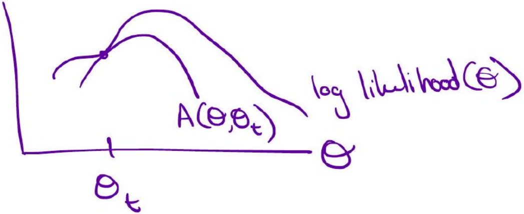
- Back to where we were:

$$\begin{aligned}
\log \ \text{likelihood} \ (\theta) &= \sum_i \log \mathbb{E}_z \frac{P\left(X_i = x_i, Z_i = k \mid \theta\right)}{P\left(Z_i = k \mid x_i, \theta_t\right)} \\
&\geq \sum_i \mathbb{E}_z \log \frac{P\left(X_i = x_i, Z_i = k \mid \theta\right)}{P\left(Z_i = k \mid x_i, \theta_t\right)} \qquad \text{(Jensen's inequality)} \\
&= \sum_i \sum_k P\left(Z_i = k \mid x_i, \theta_t\right) \log \frac{P\left(X_i = x_i, Z_i = k \mid \theta\right)}{P\left(Z_i = k \mid x_i, \theta_t\right)} =: A\left(\theta, \theta_t\right).
\end{aligned}$$

- $A\left(\cdot, \theta_t\right)$ is called the auxiliary function.

Sanity check

- Let's make sure that $A\left(\theta_t, \theta_t\right)$ is log likelihood $\left(\theta_t\right)$.



$$A(\theta_t, \theta_t) = \sum_i \sum_k P(Z_i = k | x_i, \theta_t) \log \frac{P(X_i = x_i, Z_i = k | \theta_t)}{P(Z_i = k | x_i, \theta_t)}$$

- From the definition of conditional probability, $P\left(X_i = x_i, Z_i = k \mid \theta_t\right) =$ $P\left(Z_i = k \mid x_i, \theta_t\right) P\left(X_i = x_i \mid \theta_t\right)$.

Sanity check

- Plugging this in,

$$A\left(\theta_t,\theta_t\right)=\sum_i\sum_k P\left(Z_i=k \mid x_i,\theta_t\right)\log P\left(X_i=x_i \mid \theta_t\right)$$

- Note that $\sum_k P\left(Z_i=k \mid x_i,\theta_t\right)=1$ because this is a sum over a whole probability distribution, and the other term doesn't depend on k. So,

$$\begin{aligned}A\left(\theta_t,\theta_t\right)&=\sum_i \log P\left(X_i=x_i \mid \theta_t\right)=\log\prod_i P\left(X_i=x_i \mid \theta_t\right)\\&=\log \mathrm{likelihood}\left(\theta_t\right).\end{aligned}$$

Back to EM

- Recall our auxiliary function, which is a function of θ.

$$A(\theta, \theta_t) := \sum_i \sum_k P(Z_i = k \mid x_i, \theta_t) \log \frac{P(X_i = x_i, Z_i = k \mid \theta)}{P(Z_i = k \mid x_i, \theta_t)}.$$

  - E-step: compute $P(Z_i = k \mid x_i, \theta_t) =: \gamma_{ik}$ for each i, k.
  - M-step:

$$\max_\theta A(\theta, \theta_t) = \sum_i \sum_j \gamma_{ik} \log \frac{P(X_i = x_i, Z_i = k \mid \theta)}{\gamma_{ik}}$$

- The term in the denominator doesn't depend on θ so it is not involved in the maximization. Thus it becomes:

$$\max_\theta \sum_i \sum_j \gamma_{ik} \log P(X_i = x_i, Z_i = k \mid \theta)$$

- Take the derivative and set it to 0.

Back to GMM

- Let us now apply EM to GMM. Here is a reminder of the notation:

$$\begin{aligned} w_{kt} &= \text{ probability to belong to cluster } k \text{ at iteration } t \\ \boldsymbol{\mu}_{kt} &= \text{ mean of cluster } k \text{ at iteration } t \\ \boldsymbol{\Sigma}_{kt} &= \text{ covariance of } k \text{ at iteration } t \end{aligned}$$

  and θ_t is the collection of $(w_{kt}, \boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt})$'s at iteration t.
- **E-step**: Using Bayes Rule

$$P\left(Z_i = k \mid \mathbf{x}_i, \theta_t\right) = \frac{P\left(\mathbf{X}_i = \mathbf{x}_i \mid z_i = k, \theta_t\right) P\left(Z_i = k \mid \theta_t\right)}{P\left(\mathbf{X}_i = \mathbf{x}_i \mid \theta_t\right)}.$$

- The denominator equals a sum over k of terms like those in the numerator, by the law of total probability.

$$P\left(Z_i = k \mid \mathbf{x}_i, \theta_t\right) = \frac{N\left(\mathbf{x}_i; \boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt}\right) w_{kt}}{\sum_{k'} N\left(\mathbf{x}_i; \boldsymbol{\mu}_{k't}, \boldsymbol{\Sigma}_{k't}\right) w_{k't}} =: \gamma_{ik}$$

Back to GMM

Advantages over Kmeans

1. Cluster assignments are probabilistic
2. Consider different covariance structures in different clusters
3. Consider the prior information $P(Z_i=k|\theta)$

- $$P\left(Z_i = k \mid \mathbf{x}_i, \theta_t\right) = \frac{N\left(\mathbf{x}_i; \boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt}\right) w_{kt}}{\sum_{k'} N\left(\mathbf{x}_i; \boldsymbol{\mu}_{k't}, \boldsymbol{\Sigma}_{k't}\right) w_{k't}} =: \gamma_{ik}$$ ✗
- This is similar to k-means where we assign each point to a cluster at iteration t.
- Here, though the cluster assignments are probabilistic. (We could have indexed γ_{ik} also by t since it changes at each t, but instead we will just replace its value at each iteration for notation convenience.)

Back to GMM

- **M-step**: Here is the auxiliary function we will maximize:

$$\max_{\theta} A(\theta, \theta_t) = \sum_i \sum_j \gamma_{ik} \log P(X_i = x_i, Z_i = k \mid \theta)$$

- Update θ, which is the collection $\mathrm{w}, \boldsymbol{\mu}, \Sigma$, by setting derivatives of A to 0 , with one constraint: $\sum_k w_k = 1$.
- After a small amount of calculation (skipping steps here, setting the derivatives to zero and solving), the result for the cluster means is:

$$\boldsymbol{\mu}_{k,t+1} = \frac{\sum_i \mathbf{x}_i \gamma_{ik}}{\sum_i \gamma_{ik}} \quad *$$

- which is the mean of the $\mathbf{x}_i$'s, weighted by the probability of being in cluster k.

Back to GMM

- Setting the derivatives of the auxiliary function to 0 to get $\Sigma_{k,t+1}$:

$$\Sigma_{k,t+1} = \frac{\sum_i \gamma_{ik} \left(\mathbf{x}_i - \boldsymbol{\mu}_{k,t+1}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_{k,t+1}\right)^T}{\sum_i \gamma_{ik}}.$$

- The update for w is tricker because of the constraint. We need to do constrained optimization. The Lagrangian is:

$$L\left(\theta, \theta_t\right) = A\left(\theta, \theta_t\right) + \lambda \left(1 - \sum_k w_k\right)$$

where λ is the Lagrange multiplier.

Back to GMM

- Remember that w_k is part of θ. Taking the derivative, and using index k' so as not to be confused with the sum over k :

$$\begin{aligned}\frac{\partial L(\theta, \theta_t)}{\partial w_{k'}} &= \frac{\partial A(\theta, \theta_t)}{\partial w_{k'}} - \lambda \\ &= \frac{\partial}{\partial w_{k'}} \left(\sum_i \sum_k \gamma_{ik} \log P(\mathbf{X}_i = \mathbf{x}_i, Z_i = k \mid \theta) \right) - \lambda.\end{aligned}$$

-

$$P(\mathbf{X}_i = \mathbf{x}_i, Z_i = k \mid \theta) = P(Z_i = k \mid \mathbf{w}) \cdot P \bigl(\mathbf{X}_i = \mathbf{x} \mid Z_i = k, \boldsymbol{\mu}_{k,t+1}, \boldsymbol{\Sigma}_k$$

$$= w_k \cdot N \left(\mathbf{x}; \boldsymbol{\mu}_{k,t+1}, \boldsymbol{\Sigma}_{k,t+1} \right).$$

Back to GMM

- Plugging it back

$$\frac{\partial L(\theta,\theta_t)}{\partial w_{k'}} = \sum_i \frac{\partial}{\partial w_{k'}} \left[\gamma_{ik'} \log \left[w_{k'} N\left(\mathbf{x}; \boldsymbol{\mu}_{k',t+1}, \boldsymbol{\Sigma}_{k',t+1}\right)\right]\right] - \lambda$$
$$= \sum_i \frac{\partial}{\partial w_{k'}} \left[\gamma_{ik'} \log\left(w_{k',t+1}\right)\right] + \frac{\partial}{\partial w_{k'}} \left[N\left(\mathbf{x}; \boldsymbol{\mu}_{k',t+1}, \boldsymbol{\Sigma}_{k',t+1}\right)\right]$$

- Here, $N\left(\mathbf{x}; \boldsymbol{\mu}_{k',t+1}, \boldsymbol{\Sigma}_{k',t+1}\right)$ does not depend on $w_{k'}$ so we can remove that term.
-
$$\frac{\partial L(\theta,\theta_t)}{\partial w_{k'}} = \sum_i \frac{\partial}{\partial w_{k'}} \left[\gamma_{ik'} \log\left(w_{k'}\right)\right] - \lambda$$
$$= \sum_i \gamma_{ik'} \frac{1}{w_{k'}} - \lambda = \frac{1}{w_{k'}} \sum_i \gamma_{ik'} - \lambda$$

- Setting the derivative to 0 , we can now solve for $w_{k',t+1}$:

$$w_{k',t+1} = \frac{\sum_i \gamma_{ik'}}{\lambda}$$

- We know that $\sum_{k'} w_{k',t+1} = 1$, so λ is the normalization factor:

$$\lambda = \sum_k \sum_i \gamma_{ik} = \sum_i \left(\sum_k P\left(Z_i = k \mid \mathbf{x}_i, \theta\right) \right) = \sum_i 1 = n$$

where $\sum_k P\left(Z_i = k \mid \mathbf{x}_i, \theta\right) = 1$ because it is the sum over the whole probability distribution.

- Thus, we finally have our last update for the iterative procedure to optimize the parameters of GMM.

$$w_{k',t+1} = \frac{\sum_i \gamma_{ik'}}{n}.$$