CS 475 Machine Learning: Lecture 11
Gaussian Mixture Models[*]

Prof. Mark Dredze

1 Goal

Our goal is to create a clustering algorithm based on Gaussians. We will replace each cluster with a Gaussian distribution generating that cluster. This favors points closer to the center of the Gaussian (higher probability).
Draw a picture
The generative process works as follows. We are given a set of Gaussians. For each Gaussian, we know the probability of selecting that Gaussian, and the parameters of the Gaussian (μ, Σ). To generate a point, we select a Gaussian, then generate a point according to the distribution for that Gaussian.

Why is this better? Each point can be represented using a probability instead of a hard assignment.

2 Mixtures of Gaussians

Let's recall some basic definitions.

The probability of a Gaussian distribution generating an example $\mathbf{x}$ is written as:

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$$

where μ is the mean and $\mathbf{\Sigma}$ is the co-variance matrix.

Let's assume that we have a Gaussian Mixture Model, which means we have k Gaussians from which we can generate examples $\mathbf{x}$. The probability of an example is given by adding the probability that each Gaussian generated this example:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

The weight π_k is called the mixture coefficient. It indicates how likely each Gaussian is to generate an example. Since this encodes a distribution over the Gaussians, we have:

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

and

$$0 \leq \pi_k \leq 1$$

[*] You will notice that the notation I use differs from the Bishop book. I have chosen this alternate notation to remain consistent with the majority of the literature on machine learning. This means that when you read papers in the literature you will be familiar with their notation.

2.1 Mixture Model Formulation

We have shown how to write the probability of an example $\mathbf{x}$ according to our model. We also want to write the probability of an example and a cluster assignment, meaning an assignment to a specific Gaussian.

To do this, we introduce $\mathbf{z}$, a k dimensional binary random variable, where $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$.

We now ask what is the marginal distribution over $\mathbf{z}$:

$$p(z_k = 1) = \pi_k$$

Since the probability of cluster k is π_k, then the probability of the kth position of $\mathbf{z}$ is π_k.

We can then write

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

Since z_k is either 1 or 0, this will just select the π_k that corresponds to the active elements in $\mathbf{z}$.

Let's return to writing the probability of an example $\mathbf{x}$ according to our model. Using marginalization and rewriting the joint we get:

$$p(\mathbf{x}) = \sum_z p(\mathbf{x}, \mathbf{z}) = \sum_z p(\mathbf{x}|\mathbf{z}) p(\mathbf{z})$$

We already have an expression for $p(\mathbf{z})$. We need an expression for the conditional $p(\mathbf{x}|\mathbf{z})$. We know that $p(\mathbf{x})$ is given by a Gaussian, and conditioning on $\mathbf{z}$ indicates which Gaussian to use. We can write this as:

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)^{z_k}$$

As before, since z_k is either 0 or 1 it simply selects the active Gaussian.

Substituting this back into $p(\mathbf{x})$ we obtain:

$$p(\mathbf{x}) = \sum_z p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

We will also make use of the conditional probability of $\mathbf{z}$ given $\mathbf{x}$. We represent this as $\gamma(z_k)$. We call this the responsibility of k for the example $\mathbf{x}$.

We can obtain the definition of $\gamma(z_k)$ using Bayes rule:

$$\begin{aligned}
\gamma(z_k) = p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1) p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1) p(\mathbf{x}|z_j = 1)} \\
&= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \mathbf{\Sigma}_j)}
\end{aligned}$$

2.2 Maximum Likelihood Updates

Now that we have a probabilistic model, we can write the likelihood of our data given our model parameters π, μ and $\mathbf{\Sigma}$.

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

How should we set our model parameters? Let's use maximum likelihood!

First, we write the log-likelihood function:

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

Note: there are some problems with maximum likelihood in this setting, which you can read about in the book.

Let's begin by taking the derivative with respect to μ_k. This involves the derivative of the Gaussian, which we encountered in linear regression. With some algebra we can set the derivative equal to 0 and write:

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \mathbf{\Sigma}_j)} \Sigma_k(\mathbf{x}_n - \mu_k)$$

Notice that the first term (fraction) is $\gamma(z_{nk})$, which is the responsibility for cluster k generating point n. We can replace this with $\gamma(z_{nk})$. Multiplying by $\mathbf{\Sigma}_k^{-1}$ and solving for μ_k we get:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

N_k is interpreted as the number of points in cluster k, or the total responsibility of k for the data. Note that the mean of the kth Gaussian is a weighted mean of all of the examples, weighted by the responsibility of cluster k for each example.

Following the same procedure we can solve for the covariance $\mathbf{\Sigma}_k$:

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

This has the same form as a single Gaussian's variance, except each example is weighted by the probability of that cluster generating that example.

Finally, we solve for π_k, for which we need to include the constraint that $\sum_k \pi_k = 1$. Doing so yields:

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \mathbf{\Sigma}_j)} - N$$

Notice again the first term is $\gamma(z_{nk})$, so solving for π_k we get:

$$\pi_k = \frac{N_k}{N}$$

This means that the responsibility for the kth Gaussian component is the average responsibility that component takes for explaining the data.

2.3 Maximizing Objective

Note that we do not have a closed-form solution. The parameters $\gamma(z_{nk})$ depend on the other parameters.

As with K-means, we take an iterative approach to maximizing the objective. First, choose some initial values for the means, covariances and mixing coefficients. Alternate between two steps: first estimate the responsibilities $\gamma(z_{nk})$ and then re-estimate the means (μ_k), covariances ($\mathbf{\Sigma}_k$) and mixing coefficients (π_k).