Gaussian Mixture Models

David Rosenberg, Brett Bernstein

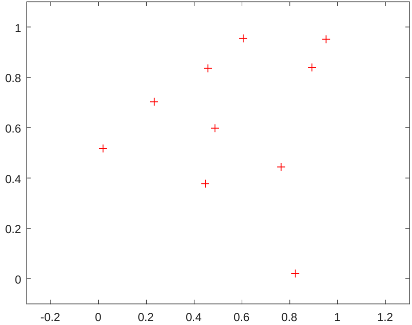
New York University

April 26, 2017

Intro Question

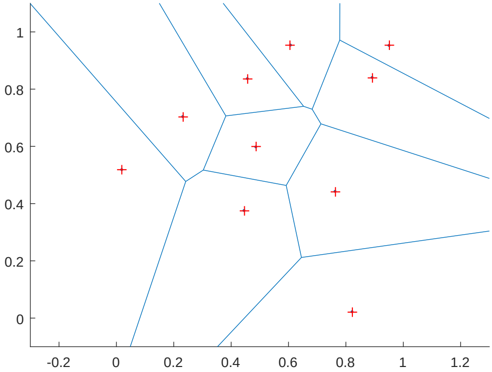
Intro Question

Suppose we begin with a dataset $\mathcal{D} = \{x_1, \ldots, x_n\} \subseteq \mathbf{R}^2$ and we run k-means (or k-means++) to obtain k cluster centers. Below we have drawn the cluster centers. If we are given a new $x \in \mathbf{R}^2$, we can assign it a label based on which cluster center is closest. What regions of the plane below correspond to each possible labeling?



Intro Solution

- Note that each cell is disjoint (except for the boarders), and convex.
- This can be thought of as a limitation of k-means: neither will be true for GMMs.



Gaussian Mixture Models

Yesterday's Intro Question

Consider the following probability model for generating data.

1. Roll a weighted k-sided die to choose a label $z \in \{1, \ldots, k\}$. Let π denote the PMF for the die.
2. Draw $x \in \mathbf{R}^d$ randomly from the multivariate normal distribution $\mathcal{N}(\mu_z, \Sigma_z)$.

Solve the following questions.

1. What is the joint distribution of x, z given π and the μ_z, Σ_z values?
2. Suppose you were given the dataset $\mathcal{D} = \{(x_1, z_1), \ldots, (x_n, z_n)\}$. How would you estimate the die weightings, and the μ_z, Σ_z values?
3. How would you determine the label for a new datapoint x?

Yesterday's Intro Solution

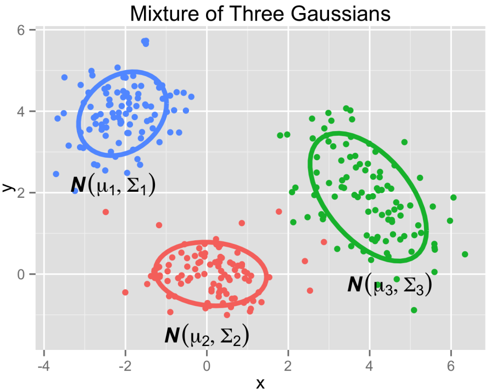
1. The joint PDF/PMF is given by
$$p(x,z) = \pi(z) f(x;\mu_z,\Sigma_z)$$
where
$$f(x;\mu_z,\Sigma_z) = \frac{1}{\sqrt{|2\pi\Sigma_z|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$
2. We could use maximum likelihood estimation. Our estimates are
$$\begin{aligned} n_z &= \textstyle\sum_{i=1}^n \mathbf{1}(z_i = z) \\ \hat{\pi}(z) &= \tfrac{n_z}{n} \\ \hat{\mu}_z &= \tfrac{1}{n_z} \textstyle\sum_{i:z_i=z} x_i \\ \hat{\Sigma}_z &= \tfrac{1}{n_z} \textstyle\sum_{i:z_i=z} (x_i - \hat{\mu}_z)(x_i - \hat{\mu}_z)^T. \end{aligned}$$
3. $\arg\max_z p(x,z)$

Probabilistic Model for Clustering

- Let's consider a **generative model** for the data.
- Suppose
  1. There are k clusters.
  2. We have a probability density for each cluster.
- Generate a point as follows
  1. Choose a random cluster $z \in \{1, 2, \dots, k\}$.
  2. Choose a point from the distribution for cluster Z.
- The clustering algorithm is then:
  1. Use training data to fit the parameters of the generative model.
  2. For each point, choose the cluster with the highest likelihood based on model.

Gaussian Mixture Model ($k = 3$)

1. Choose $z \in \{1, 2, 3\}$
2. Choose $x \mid z \sim \mathcal{N}(X \mid \mu_z, \Sigma_z)$.



Gaussian Mixture Model Parameters (k Components)

$$
\begin{aligned}
\text{Cluster probabilities}: &\quad \pi = (\pi_1, \ldots, \pi_k) \\
\text{Cluster means}: &\quad \mu = (\mu_1, \ldots, \mu_k) \\
\text{Cluster covariance matrices:} &\quad \Sigma = (\Sigma_1, \ldots \Sigma_k)
\end{aligned}
$$

- What if one cluster had many more points than another cluster?

Gaussian Mixture Model: Joint Distribution

- Factorize the joint distribution:

$$\begin{aligned} p(x,z) &= p(z)p(x \mid z) \\ &= \pi_z \mathcal{N}(x \mid \mu_z, \Sigma_z) \end{aligned}$$

  - π_z is probability of choosing cluster z.
  - $x \mid z$ has distribution $\mathcal{N}(\mu_z, \Sigma_z)$.
  - z corresponding to x is the true cluster assignment.

- Suppose we know all the parameters of the model.
- Then we can easily compute the joint $p(x,z)$, and the conditional $p(z \mid x)$.

Latent Variable Model

- We observe x.
- In the intro problem we had labeled data, but here we don't observe z, the cluster assignment.
- Cluster assignment z is called a **hidden variable** or **latent variable**.

Definition

A **latent variable model** is a probability model for which certain variables are never observed.

e.g. The Gaussian mixture model is a latent variable model.

The GMM "Inference" Problem

- We observe x. We want to know z.
- The conditional distribution of the cluster z given x is

$$p(z \mid x) = p(x, z)/p(x)$$

- The conditional distribution is a **soft assignment** to clusters.
- A **hard assignment** is

$$z^* = \underset{z \in \{1, \ldots, k\}}{\arg\max}\, p(z \mid x).$$

- So if we have the model, clustering is trivial.

Mixture Models

Gaussian Mixture Model: Marginal Distribution

- The **marginal distribution** for a single observation x is

$$\begin{aligned} p(x) &= \sum_{z=1}^{k} p(x,z) \\ &= \sum_{z=1}^{k} \pi_z \mathcal{N}(x \mid \mu_z, \Sigma_z) \end{aligned}$$

- Note that $p(x)$ is a convex combination of probability densities.
- This is a common form for a probability model...

Mixture Distributions (or Mixture Models)

Definition

A probability density $p(x)$ represents a **mixture distribution** or **mixture model**, if we can write it as a **convex combination** of probability densities. That is,

$$p(x) = \sum_{i=1}^{k} w_i p_i(x),$$

where $w_i \geqslant 0$, $\sum_{i=1}^{k} w_i = 1$, and each p_i is a probability density.

- In our Gaussian mixture model, x has a **mixture distribution**.
- More constructively, let S be a set of probability distributions:
  1. Choose a distribution randomly from S.
  2. Sample x from the chosen distribution.
- Then x has a mixture distribution.

Learning in Gaussian Mixture Models

The GMM "Learning" Problem

- Given data $x_1, \ldots, x_n$ drawn from a GMM,
- Estimate the parameters:

$$\begin{aligned} \text{Cluster probabilities:} \quad & \pi = (\pi_1, \ldots, \pi_k) \\ \text{Cluster means:} \quad & \mu = (\mu_1, \ldots, \mu_k) \\ \text{Cluster covariance matrices:} \quad & \Sigma = (\Sigma_1, \ldots \Sigma_k) \end{aligned}$$

- Once we have the parameters, we're done.
- Just do "inference" to get cluster assignments.

Estimating/Learning the Gaussian Mixture Model

- One approach to learning is **maximum likelihood**
  - find parameter values that give **observed data** the **highest likelihood**.
- The model likelihood for $\mathcal{D} = \{x_1, \ldots, x_n\}$ is

$$\begin{aligned} L(\pi, \mu, \Sigma) &= \prod_{i=1}^{n} p(x_i) \\ &= \prod_{i=1}^{n} \sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z). \end{aligned}$$

- As usual, we'll take our objective function to be the log of this:

$$J(\pi, \mu, \Sigma) = \sum_{i=1}^{n} \log \left\{ \sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z) \right\}$$

Properties of the GMM Log-Likelihood

- GMM log-likelihood:

$$J(\pi,\mu,\Sigma) \quad = \quad \sum_{i=1}^{n} \log\left\{\sum_{z=1}^{k} \pi_z \mathcal{N}\left(x_i \mid \mu_z, \Sigma_z\right)\right\}$$

- Let's compare to the log-likelihood for a single Gaussian:

$$\begin{aligned} & \sum_{i=1}^{n} \log \mathcal{N}\left(x_i \mid \mu, \Sigma\right) \\ = \quad & -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu) \end{aligned}$$

- For a single Gaussian, the log cancels the exp in the Gaussian density.
    - $\Longrightarrow$ Things simplify a lot.
- For the GMM, the sum inside the log prevents this cancellation.
    - $\Longrightarrow$ Expression more complicated. No closed form expression for MLE.

Issues with MLE for GMM

Identifiability Issues for GMM

- Suppose we have found parameters

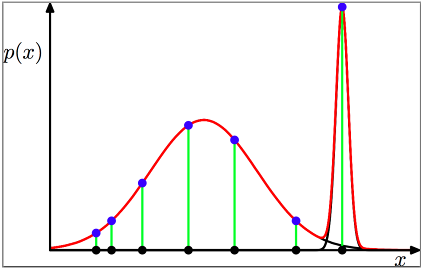
$$\begin{aligned}\text{Cluster probabilities}: &\quad \pi = (\pi_1, \ldots, \pi_k) \\ \text{Cluster means}: &\quad \mu = (\mu_1, \ldots, \mu_k) \\ \text{Cluster covariance matrices:} &\quad \Sigma = (\Sigma_1, \ldots \Sigma_k)\end{aligned}$$

  that are at a local minimum.

- What happens if we shuffle the clusters? e.g. Switch the labels for clusters 1 and 2.
- We'll get the same likelihood. How many such equivalent settings are there?
- Assuming all clusters are distinct, there are $k!$ equivalent solutions.
- Not a problem per se, but something to be aware of.

Singularities for GMM

- Consider the following GMM for 7 data points:



- Let σ^2 be the variance of the skinny component.
- What happens to the likelihood as $\sigma^2 \to 0$?
- In practice, we end up in local minima that do not have this problem.
  - Or keep restarting optimization until we do.
- Bayesian approach or regularization will also solve the problem.

From Bishop's *Pattern recognition and machine learning*, Figure 9.7.

Gradient Descent / SGD for GMM

- What about running gradient descent or SGD on

$$J(\pi, \mu, \Sigma) \quad = \quad -\sum_{i=1}^{n} \log \left\{ \sum_{z=1}^{k} \pi_z \mathcal{N}(x_i \mid \mu_z, \Sigma_z) \right\}?$$

- Can be done – but need to be clever about it.
- Each matrix $\Sigma_1, \ldots, \Sigma_k$ has to be positive semidefinite.
- How to maintain that constraint?
    - Rewrite $\Sigma_i = M_i M_i^T$, where M_i is an unconstrained matrix.
    - Then Σ_i is positive semidefinite.

The EM Algorithm for GMM

MLE for GMM

- From yesterday's intro questions, we know that we can solve the MLE problem if the cluster assignments z_i are known

$$
\begin{aligned}
n_z &= \sum_{i=1}^{n} \mathbf{1}(z_i = z) \\
\hat{\pi}(z) &= \frac{n_z}{n} \\
\hat{\mu}_z &= \frac{1}{n_z} \sum_{i:z_i=z} x_i \\
\hat{\Sigma}_z &= \frac{1}{n_z} \sum_{i:z_i=z} (x_i - \hat{\mu}_z)(x_i - \hat{\mu}_z)^T.
\end{aligned}
$$

- In the EM algorithm we will modify the equations to handle our evolving **soft assignments**, which we will call **responsibilities**.

Cluster Responsibilities: Some New Notation

- Denote the probability that observed value x_i comes from cluster j by

$$\gamma_i^j = \mathbb{P}(Z = j \mid X = x_i).$$

- The **responsibility** that cluster j takes for observation x_i.
- Computationally,

$$\begin{aligned}
\gamma_i^j &= \mathbb{P}(Z = j \mid X = x_i). \\
&= p(Z = j, X = x_i)/p(x) \\
&= \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)}
\end{aligned}$$

- The vector $\left(\gamma_i^1, \ldots, \gamma_i^k\right)$ is exactly the **soft assignment** for x_i.
- Let $n_c = \sum_{i=1}^{n} \gamma_i^c$ be the "number" of points **soft assigned** to cluster c.

EM Algorithm for GMM: Overview

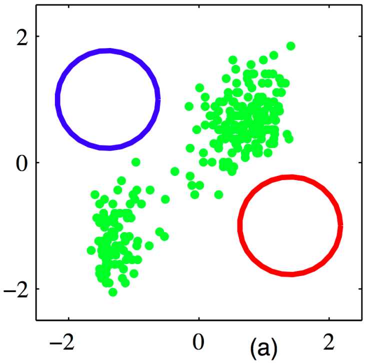
- If we know π and μ_j, Σ_j for all j then we can easily find $\gamma_i^j = \mathbb{P}(Z = j \mid X = x_i)$.
- If we know the (soft) assignments, we can easily find estimates for π, μ_j, Σ_j for all j.
- Repeatedly alternate the previous 2 steps.

EM Algorithm for GMM: Overview

1. Initialize parameters μ, Σ, π.
2. "E step". Evaluate the responsibilities using current parameters:
$$\gamma_i^j = \frac{\pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}{\sum_{c=1}^k \pi_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c)}, \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, k.$$
3. "M step". Re-estimate the parameters using responsibilities. [Compare with intro question.]
$$\begin{aligned} \mu_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^n \gamma_i^c x_i \\ \Sigma_c^{\text{new}} &= \frac{1}{n_c} \sum_{i=1}^n \gamma_i^c \left(x_i - \mu_c^{\text{new}}\right)\left(x_i - \mu_c^{\text{new}}\right)^T \\ \pi_c^{\text{new}} &= \frac{n_c}{n}, \end{aligned}$$
4. Repeat from Step 2, until log-likelihood converges.

EM for GMM

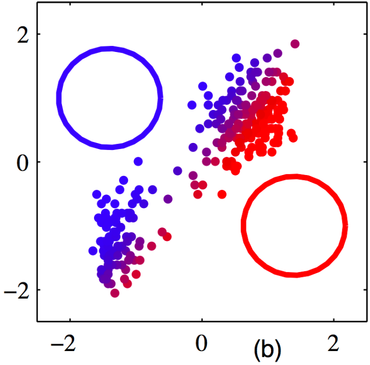
- Initialization



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

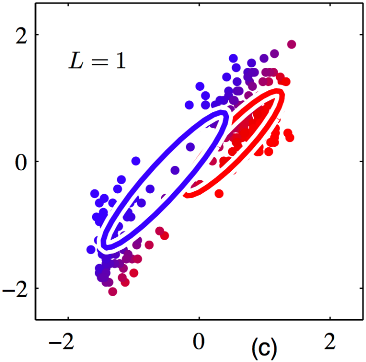
- First soft assignment:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

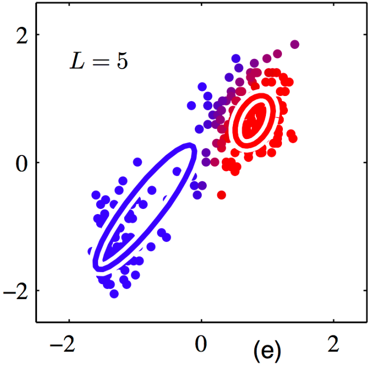
- First soft assignment:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

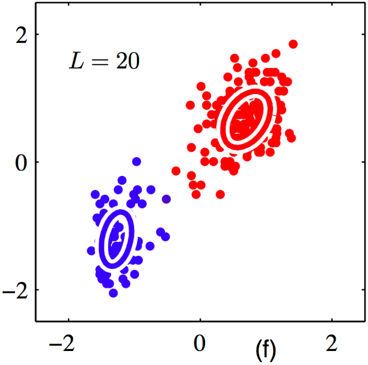
- After 5 rounds of EM:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

EM for GMM

- After 20 rounds of EM:



From Bishop's *Pattern recognition and machine learning*, Figure 9.8.

Relation to K-Means

- EM for GMM seems a little like k-means.
- In fact, there is a precise correspondence.
- First, fix each cluster covariance matrix to be $\sigma^2 I$.
- Then the density for each Gausian only depends on distance to the mean.
- As we take $\sigma^2 \to 0$, the update equations converge to doing k-means.
- If you do a quick experiment yourself, you'll find
    - Soft assignments converge to hard assignments.
    - Has to do with the tail behavior (exponential decay) of Gaussian.
- Can use k-means++ to initialize parameters of EM algorithm.

Math Prerequisites for General EM Algorithm

Jensen's Inequality

- Which is larger: $\mathbb{E}[X^2]$ or $\mathbb{E}[X]^2$?

Jensen's Inequality

- Which is larger: $\mathbb{E}[X^2]$ or $\mathbb{E}[X]^2$?
- Must be $\mathbb{E}[X^2]$ since $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geqslant 0$.
- More general result is true:

Theorem

Jensen's Inequality If $f : \mathbf{R} \to \mathbf{R}$ is convex and X is a random variable then $\mathbb{E}[f(X)] \geqslant f(\mathbb{E}[X])$. If f is strictly convex then we have equality iff $X = \mathbb{E}[X]$ with probability 1 (i.e., X is constant).

Proof of Jensen

Exercise

Suppose X can take exactly two value: x_1 with probability π_1 and x_2 with probability π_2. Then prove Jensen's inequality.

Proof of Jensen

Exercise

Suppose X can take exactly two value: x_1 with probability π_1 and x_2 with probability π_2. Then prove Jensen's inequality.

- Let's compute $\mathbb{E}[f(X)]$:

$$\mathbb{E}[f(X)] = \pi_1 f(x_1) + \pi_2 f(x_2) \leqslant f(\pi_1 x_1 + \pi_2 x_2) = f(\mathbb{E}[X]).$$

- For the general proof, what do we know is true about all convex functions $f : \mathbf{R} \to \mathbf{R}$?

Proof of Jensen

1. Let $e = \mathbb{E}[X]$. (Remember e is just a number.)
2. Since f has a subgradient at e, there is an underestimating line $g(x) = ax + b$ that passes through the point $(e, f(e))$.
3. Then we have

$$\begin{aligned}
\mathbb{E}[f(X)] &\geqslant \mathbb{E}[g(X)] \\
&= \mathbb{E}[aX + b] \\
&= a\mathbb{E}[X] + b \\
&= ae + b \\
&= f(e) \\
&= f(\mathbb{E}[X]).
\end{aligned}$$

4. If f is strictly convex then $f = g$ at exactly 1 point, so equality iff X is constant.

KL-Divergence

- Let $p(x)$ and $q(x)$ be probability mass functions (PMFs) on $\mathcal{X}$.
- We want to measure how different they are.
- The **Kullback-Leibler** or **"KL" Divergence** is define by

$$\mathrm{KL}(p\|q)=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)}.$$

(Assumes *absolute continuity*: $q(x)=0$ implies $p(x)=0$.)

- Can also write

$$\mathrm{KL}(p\|q)=\mathbb{E}_{x\sim p}\log\frac{p(x)}{q(x)}.$$

- Note, the KL-divergence is not symmetric and doesn't satisfy the triangle inequality.

Gibbs' Inequality

Theorem

Gibbs' Inequality Let $p(x)$ and $q(x)$ be PMFs on $\mathcal{X}$. Then

$$KL(p\|q) \geqslant 0,$$

with equality iff $p(x) = q(x)$ for all $x \in \mathcal{X}$.

- Since

$$\mathrm{KL}(p\|q) = \mathbb{E}_p\left[-\log\left(\frac{q(x)}{p(x)}\right)\right],$$

this is screaming for Jensen's inequality.

Gibbs' Inequality: Proof

$$
\begin{aligned}
\mathrm{KL}(p\|q) &= \mathbb{E}_p\left[-\log\left(\frac{q(x)}{p(x)}\right)\right] \\
&\geqslant -\log\left(\mathbb{E}_p\left[\frac{q(x)}{p(x)}\right]\right) \\
&= -\log\left(\sum_{x:p(x)>0} p(x)\frac{q(x)}{p(x)}\right) \\
&= -\log\left(\sum_x q(x)\right) \\
&= -\log 1 = 0.
\end{aligned}
$$

- Since $-\log$ is strictly convex, we have equality iff q/p is constant, i.e., $q=p$.