Machine Learning

Gaussian Mixture Models

Kia Nazarpour

Context

1. Often times we need to analyse data for which we do not have their labels.
2. How can we find any structure in a collection of unlabelled data?
3. Clustering is an established category of methods for organising objects into groups whose members are similar in some way.

Context: K-means Solution

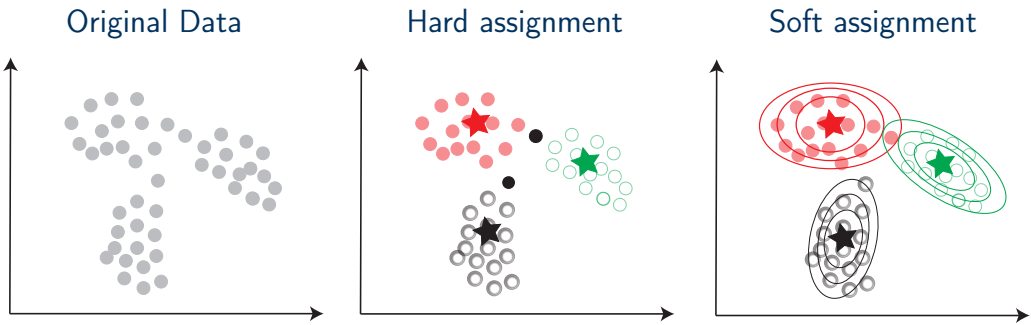
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$

Context: K-means discussion

1. Too crude? Assumes that a cluster can be represented with a single point and a simple distance metric
2. A simple unsupervised method that enables clustering of data with no great computational complexity
3. Hard boundaries!
4. Q: How to generalise it to models that can cluster data of various types and shapes!

Context: Hard assignment vs. Soft assignment



Learning Outcomes

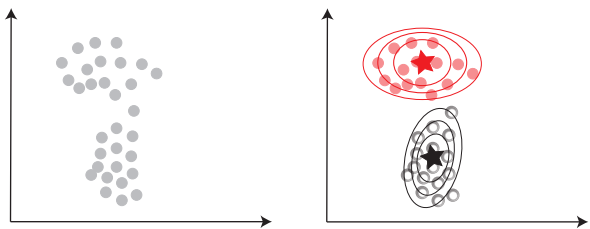
1. Understand the key motivations behind a Guassian Mixture Model (GMM).
2. Understand the formulation of a GMM and the need for the Expectation Maximisation (EM) solver.
3. Analyse the solution to a GMM.

References:

1. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2008. (Section 9.1)
2. Rogers and Girolami, *A First Course in Machine Learning*, CRC Press, 2016. (Section 6.3)

Mixture Models

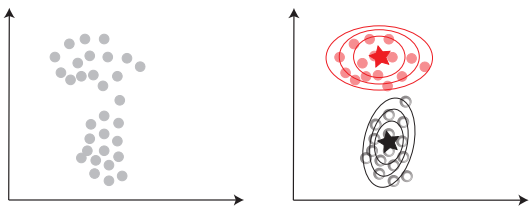
1. Models that can cluster data of various types and shapes!
2. Simple to compute
3. Clustering with statistical mixture models, similar to k-means, but offers richer representation of the data!



Mixture Models - A generative process

1. Let's assume we want to generate the below data with two Gaussians!
2. For data $\mathbf{x}_n$, Select one of the Gaussians (with probability π_k, assuming $\sum_k \pi_k = 1$). Set the parameter $z_{nk} = 1$
3. Sample data $\mathbf{x}_n$ from this Gaussian

$$p(\mathbf{x}_n | z_{nk} = 1, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Mixture Models - A generative process

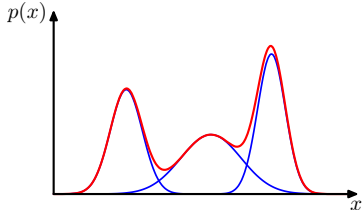
1. We described out data with a generative process
2. In a clustering context all data points with $z_{nk} = 1$ are in cluster k
3. But we need to learn/infer/calculate $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ from the observed data

BUT this is a circular argument

1. Trivial to calculate the component parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ if we knew the assignment rule $z_{nk} = 1$
2. Trivial to work out the assignment rule $z_{nk} = 1$ if we knew the component parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Mixture of Gaussians

Complex probabilities can be approximated with a linear superposition of K Gaussian densities.



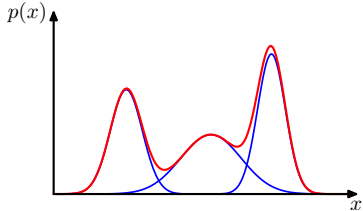
$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

We define $\mathbf{z} = \{z_1, z_2, \cdots, z_k\}$ where $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$.

We know that $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ and $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.

Mixture of Gaussians

Complex probabilities can be approximated with a linear superposition of K Gaussian densities.



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

We define $\mathbf{z} = \{z_1, z_2, \cdots, z_k\}$ where $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$.

We know that $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ and $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.

- $p(z_k = 1) = \pi_k$: $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$.
- $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$

Mixture of Gaussians

Another key quantity is $p(\mathbf{z}|\mathbf{x})$

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

$\gamma(z_k)$ is the *responsibility* that component k takes in explaining the observation $\mathbf{x}$.

A Maximum Likelihood solution to GMM? Not ideal!

Suppose we observe $\mathbf{X}_{N\times D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$. Assuming that the data points are drawn independently, the likelihood function of all N data points is

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

and so the log-likelihood will be

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

We can estimate π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ by differentiating L with respect to these variables and using gradient-based optimisation.

Expectation-Maximisation (EM) for GMMs

- The EM method can be used to overcome challenges of using Maximum Likelihood.
- EM derives a *lower bound* $\mathcal{B}$ on the likelihood L, that is $\mathcal{B} \leq L$.
- Instead of maximising L directly, EM maximises $\mathcal{B}$

- Question: How to determine $\mathcal{B}$? Using Jensen's inequality

$$\log \mathbf{E}_{p(z)}\{f(z)\} \geq \mathbf{E}_{p(z)}\{\log f(z)\}$$

- The logarithm of the expected value of $f(z)$ is always greater than or equal to the expected value of $\log f(z)$

EM - Derivation for GMMs

- Let's define γ_{nk} to be positive and satisfying $\sum_{k=1}^{K} \gamma_{nk} = 1$.
- γ_{nk} is some probability distribution over K components for the n-th data point.

$$
\begin{aligned}
L &= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \frac{\gamma_{nk}}{\gamma_{nk}} \\
&= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \gamma_{nk} \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\gamma_{nk}} \\
&= \sum_{n=1}^{N} \log \mathbf{E}_{\gamma_{nk}} \left\{ \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\gamma_{nk}} \right\}
\end{aligned}
$$

Apply Jensen's inequality

$$\log \mathbf{E}_{p(z)}\{f(z)\} \geq \mathbf{E}_{p(z)}\{\log f(z)\}$$

$$\begin{aligned} L &= \sum_{n=1}^{N} \log \mathbf{E}_{\gamma_{nk}} \left\{ \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\gamma_{nk}} \right\} \\ &\geq \sum_{n=1}^{N} \mathbf{E}_{\gamma_{nk}} \left\{ \log \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\gamma_{nk}} \right\} = \mathcal{B} \end{aligned}$$

Apply Jensen's inequality

$$\log \mathbf{E}_{p(z)}\{f(z)\} \geq \mathbf{E}_{p(z)}\{\log f(z)\}$$

$$\begin{aligned} L &= \sum_{n=1}^{N} \log \mathbf{E}_{\gamma_{nk}} \left\{ \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\gamma_{nk}} \right\} \\ &\geq \sum_{n=1}^{N} \mathbf{E}_{\gamma_{nk}} \left\{ \log \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\gamma_{nk}} \right\} = \mathcal{B} \end{aligned}$$

$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) - \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \gamma_{nk}.$$

EM - Derivation for GMMs

$$\mathcal{B} = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) - \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \gamma_{nk}.$$

- EM is an iterative process, maximising the bound $\mathcal{B}$, until convergence.
- For each update, we take the partial derivative of the bound $\mathcal{B}$ wrt parameters, set it to zero and solve.
- See Rogers and Girolami (2016) [pp.218-222] for full derivations

EM Solution for GMMs

E-step

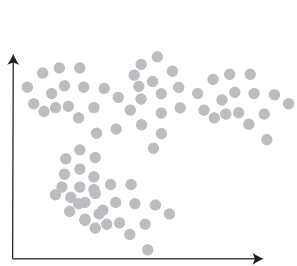
$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

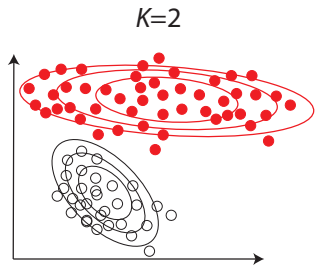
M-step

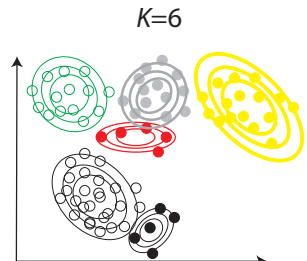
$$\begin{aligned}
\pi_k &= \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk} \\
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}} \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma_{nk}}
\end{aligned}$$

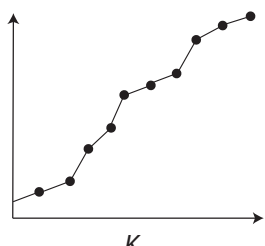
Some intution ...

Choosing the Number of components K for GMMs









GMM: Summary

- Hard boundaries are exchanged for flexible and probabilistic soft boundaries
- Immense flexibility: $p(\mathbf{x}_n|\cdots)$ can take the form of any probability density including Bernoulli distribution (binary data)
- The choice of K remains ad-hoc

Next lecture:

- Delving [a bit] deeper into the EM method