Lecture 7: Clustering

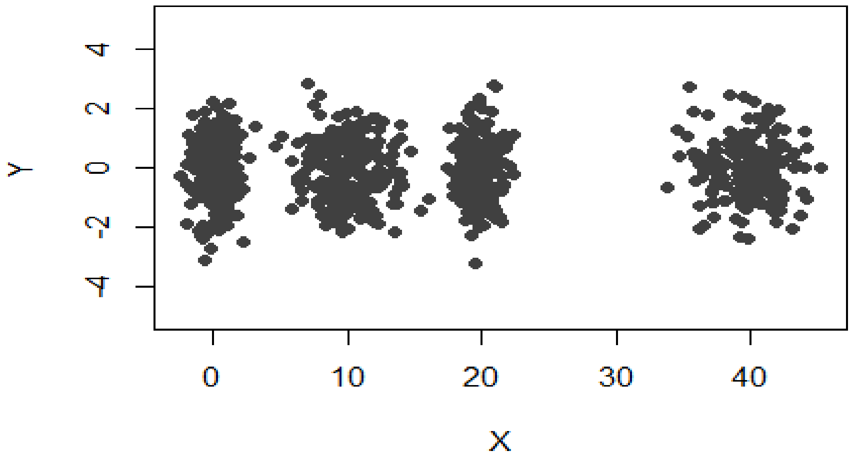
Instructor: Prof. Shuai Huang

Industrial and Systems Engineering

University of Washington

What does Clustering do

- Let’s start with the Gaussian mixture model (GMM), that has been one of the most popular clustering model.
- GMM assumes that the data come from not just one distribution but a few.



Formulation of GMM

- Suppose that there are M distributions mixed together.
- For each data point $\boldsymbol{x}_n$, the probability that it comes from the mth distribution is denoted as π_m, while $\sum_{m=1}^{M} \pi_m = 1$.
- In GMM, the mth distribution is $N(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$.
- The task is to learn the unknown parameters $\{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, m = 1,2, \ldots, M\}$ and the probability vector $\boldsymbol{\pi}$: $\{\pi_m, m = 1,2, \ldots, M\}$.
- For simplicity in the presentation, use $\boldsymbol{\Theta}$ to denote all these parameters.

Log-likelihood function of GMM

The complete log-likelihood function is:

$$
\begin{aligned}
l(\boldsymbol{\Theta}) &= \log \textstyle\prod_{n=1}^{N} p(\boldsymbol{x}_n | z_{nm} = 1; \boldsymbol{\Theta}), \\
&= \log \textstyle\prod_{n=1}^{N} p(\boldsymbol{x}_n, z_{nm} | \boldsymbol{\Theta}), \\
&= \log \textstyle\prod_{n=1}^{N} \prod_{m=1}^{M} [p(\boldsymbol{x}_n | z_{nm} = 1, \boldsymbol{\Theta}) p(z_{nm} = 1)]^{z_{nm}}, \\
&= \textstyle\sum_{n=1}^{N} \sum_{m=1}^{M} [z_{nm} \log p(\boldsymbol{x}_n | z_{nm} = 1, \boldsymbol{\Theta}) + z_{nm} \log \pi_m].
\end{aligned}
$$

Log-likelihood function of GMM (cont'd)

Meanwhile, we can derive that

$$p(\boldsymbol{x}_n | z_{nm} = 1; \boldsymbol{\Theta}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}_m|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_m)\right\}.$$

Thus,

$$l(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \left[z_{nm} \log\left((2\pi)^{-p/2} |\boldsymbol{\Sigma}_m|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_m)\right\}\right) + z_{nm} \log \pi_m \right].$$

A two-step iterative procedure

To optimize for $\mathbf{\Theta}$, we need to overcome the challenge that z_{nm}s are latent and unknown. Here, an intuitive proposal could be:

- Even we don't know z_{nm}, but we can estimate it if we have known $\mathbf{\Theta}$. For instance, it is easy to know that

$$p(z_{nm} = 1|\mathbf{X}, \mathbf{\Theta}) = \frac{p(\boldsymbol{x}_n|z_{nm}=1,\mathbf{\Theta})\pi_m}{\sum_{k=1}^{M} p(\boldsymbol{x}_n|z_{nk}=1,\mathbf{\Theta})\pi_k}.$$

  Thus, given $\mathbf{\Theta}$, the best estimate of z_{nm} could be the expectation of z_{nm} as

$$\langle z_{nm} \rangle_{p(z_{nm}|\mathbf{X},\mathbf{\Theta})} = 1 \cdot p(z_{nm} = 1|\mathbf{X}, \mathbf{\Theta}) + 0 \cdot p(z_{nm} = 0|\mathbf{X}, \mathbf{\Theta}) = \frac{p(\boldsymbol{x}_n|z_{nm}=1,\mathbf{\Theta})\pi_m}{\sum_{k=1}^{M} p(\boldsymbol{x}_n|z_{nk}=1,\mathbf{\Theta})\pi_k}.$$

- We can fill in $l(\mathbf{\Theta})$ with the estimated z_{nm} and optimize it to update $\mathbf{\Theta}$. Feed this updated back to Step 1 and repeat the iterations, until all the parameters in the iterations don't change significantly.

The EM-algorithm

This two-step iterative procedure is known as the EM algorithm.

- The E-step: Derive the posterior distribution of $\mathbf{Z}$ as $p(\mathbf{Z}|\mathbf{X}, \mathbf{\Theta})$. Calculate the expectation of $l(\mathbf{\Theta})$ according to this distribution, i.e., denoted as $\langle l(\mathbf{\Theta}) \rangle_{p(\mathbf{Z}|\mathbf{X},\mathbf{\Theta})}$.
- The M-step: obtain $\mathbf{\Theta}$ by maximizing $\langle l(\mathbf{\Theta}) \rangle_{p(\mathbf{Z}|\mathbf{X},\mathbf{\Theta})}$.

The power of the EM algorithm is that it guarantees (under mild conditions) that the objective function won't decrease along the iterations.

R lab

- Download the markdown code from course website
- Conduct the experiments
- Interpret the results
- Repeat the analysis on other datasets