Introduction to Statistical Learning

Omid Safarzadeh

January 19, 2022

Table of contents

1. Clustering
2. Intro to Expectation Maximization
3. K-means clustering
4. K-medoids clustering
5. Application of K-means
6. Gaussians mixture model (GMM)
7. Expectation Maximization
8. Reference

Clustering

- $\{\mathbf{x}_i\}_{i=1}^n$
- $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$
- Group "similar" data points together

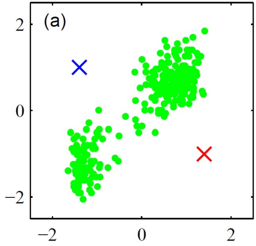

Figure: Bishop, 2013

K-means clustering

- Divide $\{\mathbf{x}_i\}_{i=1}^n$ into K clusters
- Each $\mathbf{x}_i$ belongs to only one of the K clusters (hard assignment)
- Define indicator variables for class membership:

$$r_{ik} = \begin{cases} 1 & \text{if} \quad \mathbf{x}_i \in \text{cluster } k \\ 0 & \text{if} \quad \mathbf{x}_i \notin \text{cluster } k \end{cases}$$

- Loss function (distortion measure)

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} ||\mathbf{x}_i - \mu_k||^2$$

- What is μ_k? Prototype associated with cluster k
- Find clusters such that J is minimized

K-means clustering

An iterative procedure to minimize J:

- **Step 0 (Initialization):** Start with an initial set of prototype vectors $\{\mu_k\}_{k=1}^K$
- **Step 1 (Expectation):** Minimize J with respect to r_{ik}, $i = 1, ..., n$, $k = 1, ..., K$ by fixing $\{\mu_k\}_{k=1}^K$
- **Step 2 (Maximization):** Minimize J with respect to $\{\mu_k\}_{k=1}^K$ by fixing r_{ik}, $i = 1, ..., n$, $k = 1, ..., K$

k-means algorithm:

$$\text{Initialization} \to E \to M \to E \to M \to \cdots \text{ (until convergence)}$$

The expectation step

Problem:
Minimize

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} ||\mathbf{x}_i - \mu_k||^2$$

with respect to r_{ik}, $i = 1, ..., n$, $k = 1, ..., K$
Solution:

$$r_{ik} = \begin{cases} 1 & \text{if} \quad k = arg \min_{j \in \{1,...,K\}} ||\mathbf{x}_i - \mu_j||^2 \\ 0 & \text{otherwise} \end{cases}$$

- Assign $\mathbf{x}_i$ to the cluster with the closest prototype vector!

The maximization step

Problem:
Minimize

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} ||\mathbf{x}_i - \mu_k||^2$$

with respect to $\{\mu_k\}_{k=1}^{K}$
Solution:

$$\mu_k = \frac{\sum_{i=1}^{n} r_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} r_{ik}}$$

- μ_k is the mean (average) of all data points $\mathbf{x}_i$ assigned to cluster k
- This is where the name "K-means" comes from!

Convergence of the K-means algorithm

- Does K-means converge?
- If it converges, how long does it take for it to converge?
- Where does it converge?

A probabilistic analogue of k-means

- For each cluster j we have the following parameters: π_j, μ_j, Σ_j
- Given $\{\mathbf{x}_i\}_{i=1}^n$ find the MLE estimate of (π, μ, Σ)

Solution?

MLE:

$$L(\pi, \mu, \Sigma) = \prod_{i=1}^{n} p(\mathbf{x}_i|\pi, \mu, \Sigma)$$

$$l(\pi, \mu, \Sigma) = \log L(\pi, \mu, \Sigma)$$

$$= \sum_{i=1}^{n} p(\mathbf{x}_i|\pi, \mu, \Sigma)$$

$$= \sum_{i=1}^{n} \log\{\sum_{k=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)\}$$

$$(\hat{\pi}, \hat{\mu}, \hat{\Sigma}) = arg \max l(\pi, \mu, \Sigma)$$

Illustration of k-means

Illustration of k-means

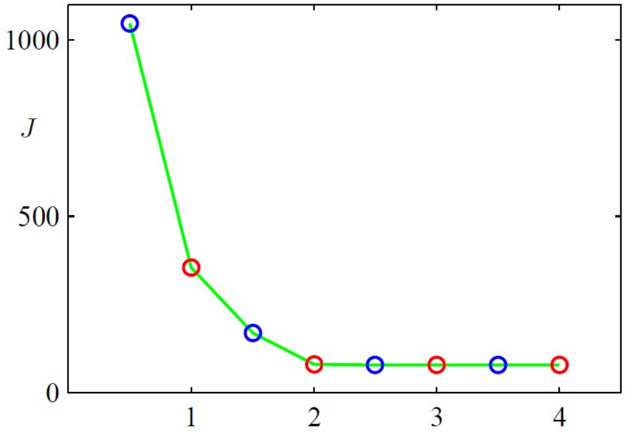

Figure: Bishop, 2013

k-medoids algorithm

- Generalization of K-means
- $||\mathbf{X}_i - \mu_k||^2 \rightarrow \mathscr{V}(\mathbf{x}_i, \mu_k)$ (a general dissimilarity measure)
- New loss function

$$\widetilde{J} = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \mathscr{V}(\mathbf{x}_i, \mu_k)$$

The same procedure applies to minimize $J^{\sim}$:

$$\text{Initialization} \rightarrow E \rightarrow M \rightarrow E \rightarrow M \rightarrow \cdots \text{ (until convergence)}$$

The expectation step

Problem:
Minimize

$$\widetilde{J} = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \mathscr{V}(\mathbf{x}_i, \mu_k)$$

with respect to r_{ik}, $i = 1, ..., n$, $k = 1, ..., K$
Solution:

$$r_{ik} = \begin{cases} 1 & \text{if} \quad k = arg\ \min_{j \in \{1,...,K\}} \mathscr{V}(\mathbf{x}_i, \mu_k) \\ 0 & \text{otherwise} \end{cases}$$

The maximization step

Problem:
Minimize

$$\widetilde{J} = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \mathscr{V}(\mathbf{x}_i, \mu_k)$$

with respect to $\{\mu_k\}_{k=1}^{K}$ such that μ_k is a datapoint that belongs to cluster k
Solution:

- Let $\mathscr{N}_k$ be the set of $\mathbf{x}_i$ that belongs to cluster k
- Choose μ_k from $\mathscr{N}_k$ such that it minimizes

$$\sum_{\mathbf{x}_i \in \mathscr{N}_k} \mathscr{V}(\mathbf{x}_i, \mu_k)$$

Application: image segmentation and compression

- Image segmentation: partition an image into region of similar visual appearance
- Each pixel is R,G,B intensity triplet: $\mathbf{x}_i = \{x_{i1}, x_{i2}, x_{i3}\}$
- Apply K-means, represent each pixel that belongs to cluster k by its cluster center μ_k

Figure: Bishop, 2013

Application: image segmentation and compression

- Assume image has n pixels
- Assume $\mathbf{x}_{ij}$ stored using 8 bits of precision
- Total number of bits to transmit the original image $= 24$n
- What happens if we compress the image by K-means and then transmit?

GMM

$$p(\mathbf{x}) = \sum_{j=1}^{K} \underbrace{\pi_j}_{\text{mixing coefficients}} \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)$$

- What is $\int p(\mathbf{x})$?
- We must have $\sum_{j=1}^{K} \pi_j = 1$ and $0 \leq \pi \leq 1$ for all $j = 1, 2, ..., K$.

Compare with:

$$p(\mathbf{x}) = \sum_{j=1}^{K} p(j)p(\mathbf{x}|j) \quad \text{law of total probability}$$

GMM example

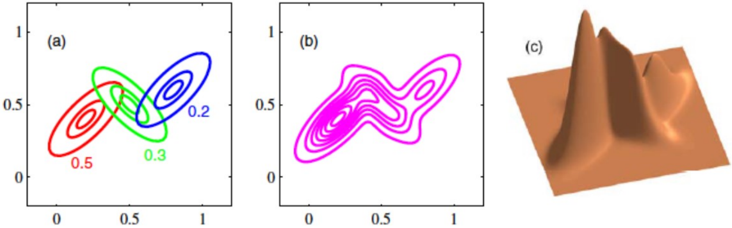

Figure: Bishop, 2013

- (a) 3 different Gaussians with mixture coefficients
- (b) contour plot of the mixture
- (c) 3D plot of the mixture

Latent (hidden) variable

- $\mathbf{z} = (z_1, ..., z_K)$ (latent variable vector)

Properties of $\mathbf{z}$:

1. $z_j \in \{0, 1\}$
2. $\sum_{j=1}^{K} z_j = 1$
3. $\mathbf{z}$ can be viewed as an indicator vector that denotes the cluster membership.

Other properties:

1. $p(z_j = 1) = \pi_j$
2. $p(\mathbf{z}) = \prod_{j=1}^{K} \pi_j^{z_j}$
3. $p(\mathbf{x}|z_j = 1) = \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)$

Conditional and marginal distribution of x

- Conditional distribution of **x**

$$p(\mathbf{x}|\mathbf{z}) = \prod_{j=1}^{K} (\mathcal{N}(\mathbf{x}|\mu_j,]Sigma_j))^{z_j}$$

- Marginal distribution of **x** :

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)$$

- $p(\mathbf{x})$ depends on $\{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{K}$
- Notation:

$$\pi = \{\pi_j\}_{j=1}^{K},$$

$$\mu = \{\mu_j\}_{j=1}^{K},$$

$$\Sigma = \{\Sigma_j\}_{j=1}^{K},$$

$$p(\mathbf{x}) = p(\mathbf{x}|\pi, \mu, \Sigma)$$

Expectation maximization (EM) algorithm

You are given the dataset $\{\mathbf{x}\}_{j=1}^{n}$

1. **Initialize parameters:** $\{\mu_j\}_{j=1}^K, \{\Sigma_j\}_{j=1}^K$ and $\{\pi_j\}_{j=1}^K$
2. **E step:** Calculate responsibilities $\{z_{ij}\}$ using current parameter values:

$$\gamma(z_{ij}) = p(z_{ij} = 1|\mathbf{x}) = \frac{\pi_j \mathcal{N}(\mathbf{x}_i|\mu_j, \Sigma_j)}{\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k)}$$

3. **M step:** Re-calculate parameters using current responsibilities:

$$\mu_j^{new} = \frac{1}{N_j} \sum_{i=1}^n \gamma(z_{ij}) \mathbf{x}_i$$

$$\Sigma_j^{new} = \frac{1}{N_j} \sum_{i=1}^n \gamma(z_{ij})(\mathbf{x}_i - \mu_j^{new})(\mathbf{x}_i - \mu_j^{new})^T$$

$$\pi_j^{new} = \frac{N_j}{n} \quad \text{where} \quad N_j = \sum_{i=1}^n \gamma(z_{ij})$$

4. **Repeat** 2 and 3 until log-likelihood or parameters converge

EM application - probabilistic clustering

- $K = 2$, Initial Gaussian centers are identical with K-means example
- Circles: one standard deviation countour plots of Gaussians
- L: Number of cycles

References

Bishop, C. (2013). *Pattern recognition and machine learning: All "just the facts 101" material*. Springer (India) Private Limited. https://books.google.fr/books?id=HL4HrgEACAAJ