Probabilistic clustering and the EM algorithm

Guillaume Obozinski

Ecole des Ponts - ParisTech

Master MVA
Outline

1. Clustering

2. The EM algorithm for the Gaussian mixture model
Clustering
Supervised, unsupervised and semi-supervised learning

Supervised learning

Training set composed of pairs \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \).
→ Learn to classify new points in the classes

Unsupervised learning

Training set composed of pairs \( \{x_1, \ldots, x_n\} \).
→ Partition the data in a number of classes.
→ Possibly produce a decision rule for new points.

Transductive learning

Data available at train time composed of
train data \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) + test data \( \{x_{n+1}, \ldots, x_n\} \)
→ Classify all the test data

Semi-supervised learning

Data available at train time composed of
labelled data \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) + unlabelled data
\( \{x_{n+1}, \ldots, x_n\} \)
→ Produce a classification rule for future points
Clustering

- Clustering is word usually used for **unsupervised classification**
- Clustering techniques can be useful to solve semi-supervised classification problem.

Clustering is not a well-specified problem

- Classes might be impossible to infer from the distribution of $X$ alone
- Several goals possible:
  - Find the modes of the distribution
  - Find a set of denser **connected** regions supporting most of the density
  - Find a set of denser **convex** regions supporting most of the density
  - Find a set of denser **ellipsoidal** regions supporting most of the density
  - Find a set of denser **round** regions supporting most of the density
K-means

**Key assumption:** Data composed of $K$ “roundish” clusters of similar sizes with centroids $(\mu_1, \cdots, \mu_K)$.

Problem can be formulated as: \[
\min_{\mu_1, \cdots, \mu_K} \frac{1}{n} \sum_{i=1}^{n} \min_k ||x_i - \mu_k||^2.
\]

Difficult (NP-hard) nonconvex problem.

**K-means algorithm**

1. Draw centroids at random
2. Assign each point to the closest centroid
   \[C_k \leftarrow \{ i \mid ||x_i - \mu_k||^2 = \min_j ||x_i - \mu_j||^2 \}\]
3. Recompute centroid as center of mass of the cluster
   \[\mu_k \leftarrow \frac{1}{|C_k|} \sum_{i \in C_k} x_i\]
4. Go to 2
K-means properties

Three remarks:

- K-means is greedy algorithm
- It can be shown that K-means converges in a finite number of steps.
- The algorithm however typically get stuck in local minima and it practice it is necessary to try several restarts of the algorithm with a random initialization to have chances to obtain a better solution.
- Will fail if the clusters are not round
**K-means++**, (Arthur and Vassilvitskii, 2007)

**Algorithm**

- Choose first center $\mu_1$ uniformly among data points

**For** $k = 2 ... K$

- Let $D_i^2 = \min_{j<k} \|x_i - \mu_k\|_2^2$
- Choose the next center among $\{x_1, \ldots, x_n\}$ with probability $\propto D_i^2$.

**endFor**

$\rightarrow$ Solution is $\log(K)$ optimal.

The Gaussian mixture model
and
the EM algorithm
Jensen’s Inequality

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$

1. if $f$ is **convex** and if $X$ is a random variable, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

2. if $f$ is **strictly convex**, we have equality in the previous inequality if and only if $X$ is constant almost surely.
Entropy

Let $X$ a r.v. with values in the finite set $\mathcal{X}$ and $p(x) = P(X = x)$.

Quantity of information of the observation $x$

$$I(x) := \log \frac{1}{p(x)}$$

Definition of entropy

$$H(X) := E[I(X)] = - \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

Remarks:

- Convention: $0 \log 0 = 0$
- $H$ defined either with natural log or the log in base 2 (i.e. $\log_2$).
- $\log_2$ is better for coding interpretations
- In this course we will use the natural logarithm.
Kullback-Leibler divergence

Definition

Let $p$ and $q$ be two finite distributions on $\mathcal{X}$ finite. The Kullback-Leibler divergence is defined by

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_{X \sim p} \left[ \log \frac{p(X)}{q(X)} \right]$$

$$= \sum_{x \in \mathcal{X}} \frac{p(x)}{q(x)} \left( \log \frac{p(x)}{q(x)} \right) q(x) = E_{X \sim q} \left[ \frac{p(X)}{q(X)} \log \frac{p(X)}{q(X)} \right]$$

⚠️ The KL divergence is not a distance: it is not symmetric. If $\exists x \in \mathcal{X}$ with $q(x) = 0$ and $p(x) \neq 0$ then $D(p \parallel q) = +\infty$. 
Kullback-Leibler divergence

Proposition

\[ D(p \parallel q) \geq 0 \] and equality holds if and only if \( p = q \).

Proof.

W.l.o.g assume \( q(x) > 0 \) everywhere.

1. \( y \mapsto \log y \) is convex so by Jensen’s inequality, we have

\[
D(p \parallel q) = E_q \left[ \frac{p(X)}{q(X)} \log \left( \frac{p(X)}{q(X)} \right) \right] \geq E_q \left[ \frac{p(X)}{q(X)} \right] \log E_q \left[ \frac{p(X)}{q(X)} \right] = 0
\]

since

\[
E_q \left[ \frac{p(X)}{q(X)} \right] = \sum_{x \in X} \frac{p(x)}{q(x)} q(x) = \sum_{x \in X} p(x) = 1.
\]

2. \( D(p \parallel q) = 0 \) iff there is equality in Jensen’s inequality

\Rightarrow \quad p(x) = cq(x) \text{ q-a.s.},

\Rightarrow \quad \text{but summing this last equality over } x \text{ implies that } c = 1,

\Rightarrow \quad \text{in turn implies that } p = q.
Differential entropy and KL

Let $X$ be a r.v. with distribution $P$ and density $p$ w.r.t. a measure $\mu$.

Differential entropy:

$$H_{\text{diff}}(p) = - \int_X p(x) \log(p(x)) d\mu(x)$$

Differential Kullback Leibler Divergence

$$D_{\text{diff}}(p \parallel q) = \int_X p(x) \log \frac{p(x)}{q(x)} d\mu(x)$$

$$= \mathbb{E}_{X \sim p} \left[ \log \frac{p(X)}{q(X)} \right]$$

- $H_{\text{diff}}(p) \geq 0$
- $H_{\text{diff}}(p)$ depends on the reference measure $\mu$.

$\Rightarrow$ $H_{\text{diff}}(p)$ does not capture intrinsic properties of $P$.

- However, $D_{\text{diff}}(p \parallel q)$ does not depend on $\mu$. 
Gaussian mixture model

- \( K \) components
- \( z \) component indicator
- \( z = (z_1, \ldots, z_K)^\top \in \{0, 1\}^K \)
- \( z \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K)) \)
- \( p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \)
- \( p(x|z; (\mu_k, \Sigma_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(x; \mu_k, \Sigma_k) \)
- \( p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k) \)
- Estimation: \( \arg\max_{\mu_k, \Sigma_k} \log \left[ \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k) \right] \)
Applying maximum likelihood to the Gaussian mixture

Let \( Z = \{ z \in \{0, 1\}^K \mid \sum_{k=1}^{K} z_k = 1 \} \)

\[
    p(x) = \sum_{z \in Z} p(x, z) = \sum_{z \in Z} \prod_{k=1}^{K} \left[ \pi_k N(x; \mu_k, \Sigma_k) \right]^{z_k} = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k)
\]

**Issue**

- The marginal log-likelihood \( \tilde{\ell}(\theta) = \sum_i \log(p(x^{(i)})) \) with \( \theta = (\pi, (\mu_k, \Sigma_k)_{1 \leq k \leq K}) \) is now complicated
- No hope to find a simple solution to the maximum likelihood problem
- By contrast the complete log-likelihood has a rather simple form:

\[
    \tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(x^{(i)}, z^{(i)}) = \sum_{i, k} z_k^{(i)} \log N(x^{(i)}; \mu_k, \Sigma_k) + \sum_{i, k} z_k^{(i)} \log(\pi_k),
\]
Applying ML to the multinomial mixture

\[ \tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(x^{(i)}, z^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k), \]

- If we knew \( z^{(i)} \) we could maximize \( \tilde{\ell}(\theta) \).
- If we knew \( \theta = (\pi, (\mu_k, \Sigma_k)_{1 \leq k \leq K}) \), we could find the best \( z^{(i)} \) since we could compute the true a posteriori on \( z^{(i)} \) given \( x^{(i)} \):

\[
p(z_k^{(i)} = 1 \mid x; \theta) = \frac{\pi_k \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x^{(i)}; \mu_j, \Sigma_j)}
\]

→ Seems a chicken and egg problem...

- In addition, we want to solve

\[
\max_{\theta} \sum_{i} \log \left( \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}) \right) \quad \text{and not} \quad \max_{\theta, \{z^{(1)}, \ldots, z^{(M)}\}} \sum_{i} \log p(x^{(i)}, z^{(i)})
\]

- Can we still use the intuitions above to construct an algorithm maximizing the marginal likelihood?
Principle of the Expectation-Maximization Algorithm

\[
\log p(x; \theta) = \log \sum_z p(x, z; \theta) = \log \sum_z q(z) \frac{p(x, z; \theta)}{q(z)} \\
\geq \sum_z q(z) \log \frac{p(x, z; \theta)}{q(z)} \\
= \mathbb{E}_q[\log p(x, z; \theta)] + H(q) =: \mathcal{L}(q, \theta)
\]

- This shows that \( \mathcal{L}(q, \theta) \leq \log p(x; \theta) \)
- Moreover: \( \theta \mapsto \mathcal{L}(q, \theta) \) is a concave function.
- Finally it is possible to show that

\[
\mathcal{L}(q, \theta) = \log p(x; \theta) - KL(q \| p(\cdot | x; \theta))
\]

So that if we set \( q(z) = p(z \mid x; \theta^{(t)}) \) then

\[
L(q, \theta^{(t)}) = p(x; \theta^{(t)}).
\]
A graphical idea of the EM algorithm

\[ L(q, \theta) = \ln p(X|\theta) \]

\[ \theta_{old} \quad \theta_{new} \]
Expectation Maximization algorithm

Initialize \( \theta = \theta_0 \)

WHILE (Not converged)

**Expectation step**
1. \( q(z) = p(z \mid x; \theta^{(t-1)}) \)
2. \( \mathcal{L}(q, \theta) = \mathbb{E}_q \left[ \log p(x, z; \theta) \right] + H(q) \)

**Maximization step**
1. \( \theta^{(t)} = \arg\max_{\theta} \mathbb{E}_q \left[ \log p(x, z; \theta) \right] \)

ENDWHILE
Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i}(z_k^{(i)} = 1) = \mathbb{E}_{q_i}[z_k^{(i)}]$, we have

$$
\mathbb{E}_{q(t)}[\tilde{\ell}(\theta)] = \mathbb{E}_{q(t)}[\log p(X, Z; \theta)]
= \mathbb{E}_{q(t)} \left[ \sum_{i=1}^{M} \log p(x^{(i)}, z^{(i)}; \theta) \right]
= \mathbb{E}_{q(t)} \left[ \sum_{i,k} z_k^{(i)} \log \mathcal{N}(x^{(i)}, \mu_k, \Sigma_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k) \right]
= \sum_{i,k} \mathbb{E}_{q_i^{(t)}}[z_k^{(i)}] \log \mathcal{N}(x^{(i)}, \mu_k, \Sigma_k) + \sum_{i,k} \mathbb{E}_{q_i^{(t)}}[z_k^{(i)}] \log(\pi_k)
= \sum_{i,k} q_{ik}^{(t)} \log \mathcal{N}(x^{(i)}, \mu_k, \Sigma_k) + \sum_{i,k} q_{ik}^{(t)} \log(\pi_k)
$$
Expectation step for the Gaussian mixture

We computed previously $q_i^{(t)}(z^{(i)})$, which is a multinomial distribution defined by

$$q_i^{(t)}(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta^{(t-1)})$$

Abusing notation we will denote $(q_{i1}^{(t)}, \ldots, q_{iK}^{(t)})$ the corresponding vector of probabilities defined by

$$q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}} [z_k^{(i)}]$$

$$q_{ik}^{(t)} = p(z_k^{(i)} = 1 | x^{(i)}; \theta^{(t-1)}) = \frac{\pi_k^{(t-1)} \mathcal{N}(x^{(i)}, \mu_k^{(t-1)}, \Sigma_k^{(t-1)})}{\sum_{j=1}^{K} \pi_j^{(t-1)} \mathcal{N}(x^{(i)}, \mu_j^{(t-1)}, \Sigma_j^{(t-1)})}$$
Maximization step for the Gaussian mixture

\[
(\pi^t, (\mu_k^t, \Sigma_k^t)_{1 \leq k \leq K}) = \arg\max_{\theta} \mathbb{E}_{q(t)} [\tilde{\ell}(\theta)]
\]

This yields the updates:

\[
\mu_k^{(t)} = \frac{\sum_i x^{(i)} q^{(t)}_{ik}}{\sum_i q^{(t)}_{ik}}, \quad \Sigma_k^{(t)} = \frac{\sum_i (x^{(i)} - \mu_k^{(t)}) (x^{(i)} - \mu_k^{(t)})^\top q^{(t)}_{ik}}{\sum_i q^{(t)}_{ik}}
\]

\[
\pi_k^{(t)} = \frac{\sum_i q^{(t)}_{ik}}{\sum_i q^{(t)}_{ik}}
\]

and
Final EM algorithm for the Multinomial mixture model

Initialize $\theta = \theta_0$

WHILE (Not converged)

Expectation step

$$q_{ik}^{(t)} \leftarrow \frac{\pi_k^{(t-1)} \mathcal{N}(x^{(i)}, \mu_k^{(t-1)}, \Sigma_k^{(t-1)})}{\sum_{j=1}^K \pi_j^{(t-1)} \mathcal{N}(x^{(i)}, \mu_j^{(t-1)}, \Sigma_j^{(t-1)})}$$

Maximization step

$$\mu_k^{(t)} = \frac{\sum_i x^{(i)} q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}, \quad \Sigma_k^{(t)} = \frac{\sum_i (x^{(i)} - \mu_k^{(t)}) (x^{(i)} - \mu_k^{(t)})^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}$$

and

$$\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{\sum_i q_{ik'}^{(t)}}$$

ENDWHILE
EM Algorithm for the Gaussian mixture model III

\[ p(x|z) \quad p(z|x) \]