Clustering, Gaussian mixture model and EM

Guillaume Obozinski

Ecole des Ponts - ParisTech

Cours MALAP 2014

Outline

Clustering

Supervised, unsupervised and semi-supervised classification

Supervised learning

Training set composed of pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
$\rightarrow$ Learn to classify new points in the classes

Unsupervised learning

Training set composed of pairs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
$\rightarrow$ Partition the data in a number of classes.
$\rightarrow$ Possibly produce a decision rule for new points.

Transductive learning

Data available at train time composed of
train data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ + test data $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_n\}$
$\rightarrow$ Classify all the test data

Semi-supervised learning

Data available at train time composed of
labelled data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ + unlabelled data $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_n\}$
$\rightarrow$ Produce a classification rule for future points

K-means

Key assumption: Data composed of K "roundish" clusters of similar sizes with centroids $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

Problem can be formulated as: $$\min_{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K} \frac{1}{n} \sum_{i=1}^{n} \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2.$$

Difficult (NP-hard) nonconvex problem.

K-means algorithm

1. Draw centroids at random
2. Assign each point to the closest centroid

$$C_k \leftarrow \left\{ i \mid \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 = \min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \right\}$$

3. Recompute centroid as center of mass of the cluster

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{\mid C_k \mid} \sum_{i \in C_k} \mathbf{x}_i$$

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

The EM algorithm for the Gaussian mixture model

The Kullback-Leibler divergence

Definition Let $\mathcal{X}$ a finite state space and p and q two distributions on $\mathcal{X}$

$$KL(p \,\|\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{X\sim p}\Big[\log \frac{p(X)}{q(X)}\Big]$$

Entropy: $H(p) = -\sum_x p(x) \log p(x)$

So $$KL(p \,\|\, q) = \mathbb{E}_{X\sim p}\big[-\log q(X)\big] - H(p).$$

Property: $\forall p, q,\ KL(p \,\|\, q) \geq 0.$

Proof:

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

Let $Y = \frac{p(X)}{q(X)}$. Then, we have $\mathbb{E}_{X\sim q}[Y] = \sum_x p(x) = 1.$

$$KL(p \,\|\, q) = \mathbb{E}_{X\sim q}\big[Y \log Y\big] = \mathbb{E}_{X\sim q}\big[\phi(Y)\big] \geq \phi(\mathbb{E}_{X\sim q}[Y]) = 0$$

Differential KL and entropies

Let P and Q two probability distributions with densities p and q with respect to a measure μ. Then, we can define

$$KL(p \,\|\, q) = \int_x \left(\log \frac{p(x)}{q(x)}\right) p(x) d\mu(x) = \mathbb{E}_{X\sim P}\Big[\log \frac{p(X)}{q(X)}\Big]$$

Differential entropy

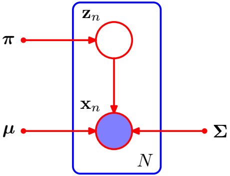
$$\mathcal{H}(p) = -\int_x p(x) \log p(x) d\mu(x)$$

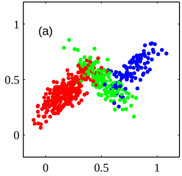
Caveats: the differential entropy is dangerous

- $\mathcal{H}(p) \not\geq 0$
- $\mathcal{H}$ depends on the choice of μ...!

Gaussian mixture model

- K components
- $\mathbf{z}$ component indicator
- $\mathbf{z} = (z_1, \ldots, z_K)^\top \in \{0,1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $p(\mathbf{x}|\mathbf{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Estimation: $\underset{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}{\operatorname{argmax}} \log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$





Applying maximum likelihood to the Gaussian mixture

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

$$p(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \prod_{k=1}^{K} \Big[\pi_k \, \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\Big]^{z_k} = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

Issue

- The marginal log-likelihood $\tilde{\ell}(\theta) = \sum_i \log(p(\mathbf{x}^{(i)}))$ with $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)_{1\le k\le K}\big)$ 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:

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)},\mathbf{z}^{(i)}) = \sum_{i,\,k} z_k^{(i)} \log \mathcal{N}(x^{(i)};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

Applying maximum likelihood to the multinomial mixture

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) = \sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

- If we knew $\mathbf{z}^{(i)}$ we could maximize $\tilde{\ell}(\theta)$.
- If we knew $\theta = \big(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1 \le k \le K}\big)$, we could find the best $\mathbf{z}^{(i)}$ since we could compute the true a posteriori on $\mathbf{z}^{(i)}$ given $\mathbf{x}^{(i)}$:

$$p(z_k^{(i)} = 1 \mid \mathbf{x}; \theta) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}^{(i)}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$\rightarrow$ Seems a chicken and egg problem...

- In addition, we want to solve

$$\max_{\theta} \sum_i \log \Big(\sum_{\mathbf{z}^{(i)}} p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \Big) \quad \text{and not} \quad \max_{\substack{\theta, \\ \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}}} \sum_i \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$$

- Can we still use the intuitions above to construct an algorithm maximizing the marginal likelihood?

Principle of the Expectation-Maximization Algorithm

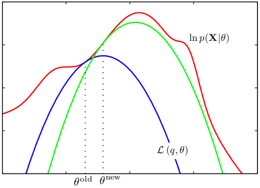
$$\begin{aligned}\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\ &\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})} \\ &\quad = \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})\end{aligned}$$

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

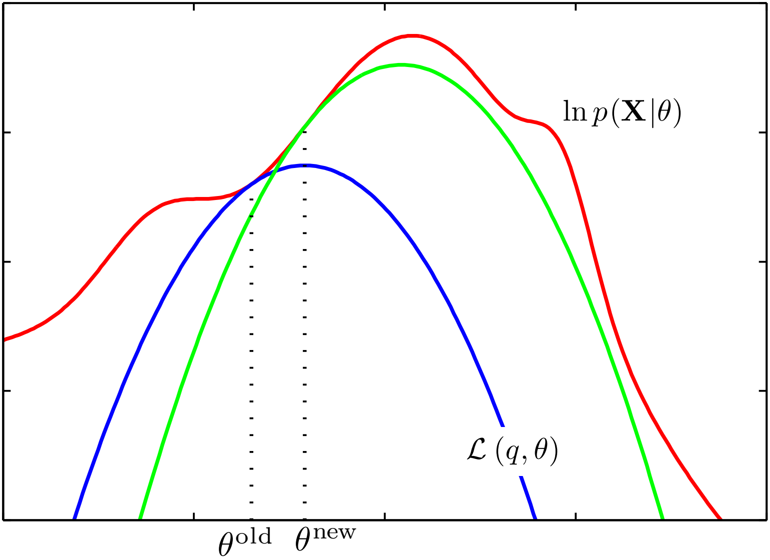
$$\mathcal{L}(q,\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - KL(q \,\|\, p(\cdot|\mathbf{x};\boldsymbol{\theta}))$$

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

$$L(q,\boldsymbol{\theta}^{(t)}) = p(\mathbf{x};\theta^{(t)}).$$



A graphical idea of the EM algorithm



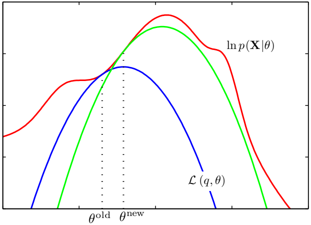
Expectation Maximization algorithm

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

Expectation step

1. $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}^{(t-1)})$
2. $\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q \big[\log p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) \big] + H(q)$



Maximization step

1. $\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_q \big[\log p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) \big]$

$$\begin{aligned} \boldsymbol{\theta}^{\text{old}} &= \boldsymbol{\theta}^{(t-1)} \\ \boldsymbol{\theta}^{\text{new}} &= \boldsymbol{\theta}^{(t)} \end{aligned}$$

ENDWHILE

Expected complete log-likelihood

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

$$
\begin{aligned}
\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] &= \mathbb{E}_{q^{(t)}}\left[\log p(\mathbf{X}, \mathbf{Z}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k)\right] \\
&= \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log(\pi_k) \\
&= \sum_{i,k} q_{ik}^{(t)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} q_{ik}^{(t)} \log(\pi_k)
\end{aligned}
$$

Expectation step for the Gaussian mixture

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

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

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

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

$$q_{ik}^{(t)} = p(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)};\theta^{(t-1)}) = \frac{\pi_k^{(t-1)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^K \pi_j^{(t-1)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\Sigma}_j^{(t-1)})}$$

Maximization step for the Gaussian mixture

$$\left(\boldsymbol{\pi}^t, (\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})_{1\le k\le K}\right) = \operatorname*{argmax}_\theta \mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right]$$

This yields the updates:

$$\boxed{\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbf{x}^{(i)}\, q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}}, \quad \boxed{\boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_i \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}}$$

and $\boxed{\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{\sum_{i,k'} q_{ik'}^{(t)}}}$

Final EM algorithm for the Multinomial mixture model

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

Expectation step

$$q_{ik}^{(t)} \leftarrow \frac{\pi_k^{(t-1)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^{K} \pi_j^{(t-1)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\Sigma}_j^{(t-1)})}$$

Maximization step

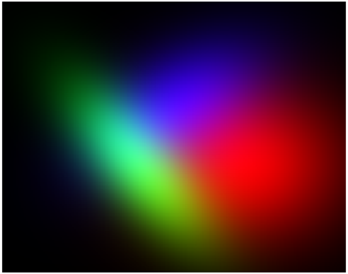
$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbf{x}^{(i)}\, q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}, \quad \boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_i \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}$$

$$\text{and} \qquad \pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{\sum_{i,k'} q_{ik'}^{(t)}}$$

ENDWHILE

EM Algorithm for the Gaussian mixture model III

$p(\mathbf{x}|\mathbf{z})$



$p(\mathbf{z}|\mathbf{x})$