Probabilistic clustering and the EM algorithm

Guillaume Obozinski

Ecole des Ponts - ParisTech

Master MVA 2014-2015

Outline

Clustering

Supervised, unsupervised and semi-supervised learning

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

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 $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

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

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

K-means++, (Arthur and Vassilvitskii, 2007)

Algorithm

- Choose first center μ_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.

See Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms.

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

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

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

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

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

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

⚠ 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. $y \mapsto y \log y$ 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

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

2. $D(p \parallel q) = 0$ iff there is equality in Jensen's inequality
   - $\Rightarrow$ $p(x) = cq(x)$ q-a.s.,
   - $\Rightarrow$ but summing this last equality over x implies that $c = 1$,
   - $\Rightarrow$ 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 μ.

Differential entropy:

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

Differential Kullback Leibler Divergence

$$\begin{aligned} D_{\text{diff}}(p \parallel q) &= \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x) \\ &= E_{X\sim p}\left[\log \frac{p(X)}{q(X)}\right] \end{aligned}$$

⚠

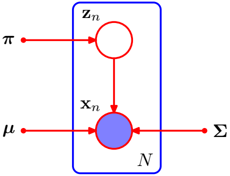
- $H_{\text{diff}}(p) \not\geq 0$
- $H_{\text{diff}}(p)$ depends on the reference measure μ.

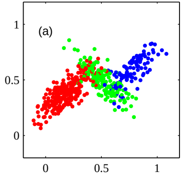
$\Rightarrow$ $H_{\text{diff}}(p)$ does not capture intrinsic properties of P.

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

Gaussian mixture model

- K components
- $\boldsymbol{z}$ component indicator
- $\boldsymbol{z} = (z_1, \ldots, z_K)^\top \in \{0,1\}^K$
- $\boldsymbol{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$
- $p(\mathbf{x}|\boldsymbol{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- 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 $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

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

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:

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

Applying ML to the multinomial mixture

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

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

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

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

Principle of the Expectation-Maximization Algorithm

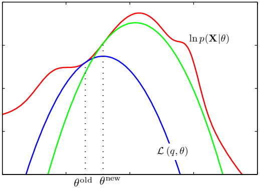
$$
\begin{aligned}
\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta}) = \log \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})} \\
&\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z})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

- 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

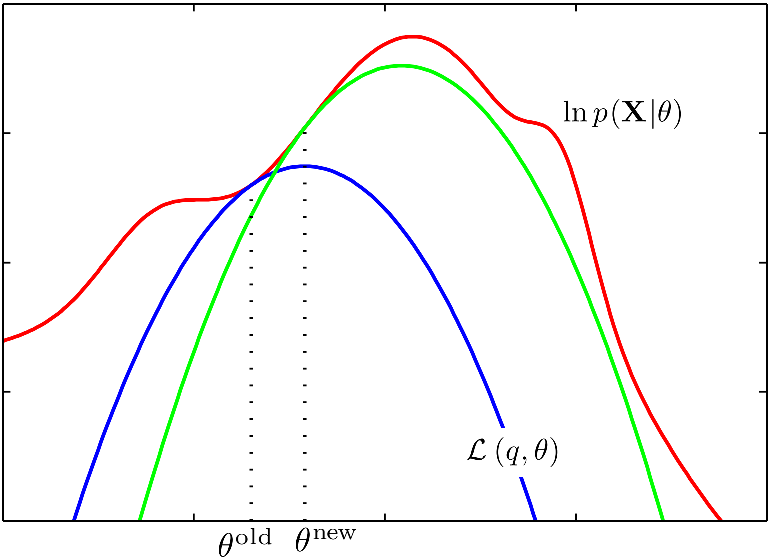
$$\mathcal{L}(q,\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - KL(q||p(\cdot|\mathbf{x};\boldsymbol{\theta}))$$

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

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



A graphical idea of the EM algorithm



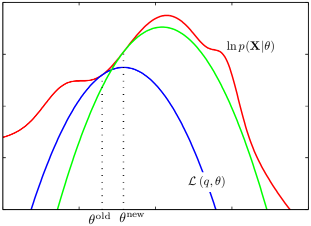
Expectation Maximization algorithm

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

WHILE (Not converged)

Expectation step

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



Maximization step

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

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

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

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

Expectation step for the Gaussian mixture

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

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

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

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

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

Maximization step for the Gaussian mixture

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

This yields the updates:

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

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

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

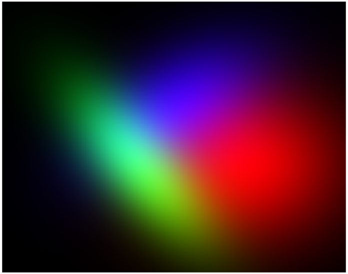
$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbf{x}^{(i)} q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}, \quad \boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_i \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}$$

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

ENDWHILE

EM Algorithm for the Gaussian mixture model III

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



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