DM825

Introduction to Machine Learning

Lecture 13

Unsupervised Learning

Marco Chiarandini

Department of Mathematics & Computer Science
University of Southern Denmark

Outline

1. k-means

2. Expectation Maximization Algorithm

Outline

1. k-means

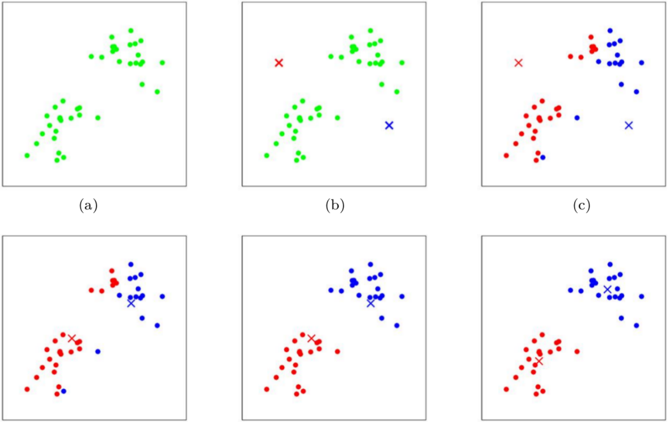
2. Expectation Maximization Algorithm

k-means

Given $\{\vec{x}_1, \ldots, \vec{x}_m\}$ and no y^i we want to cluster the data

Initialize cluster centroids randomly $\mu_1, \ldots, \mu_k \in \mathbb{R}^n$
repeat
  for $i = 1 \ldots m$ **do**
    $c^i \leftarrow \arg\min_l \parallel x^i - \mu_l \parallel^2$; // assign
  for $l = 1 \ldots k$ **do**
    $\mu_l \leftarrow \frac{\sum_{i=1}^{m} I\{c^i=l\}x^i}{\sum_{i=1}^{m} I\{c^i=l\}}$; // move
until convergence ;

k is a parameter
Optimization of the distortion function $J(\vec{c}, \vec{\mu}) = \sum_{i=1}^{m} \parallel x^i - \mu_{c^i} \parallel^2$
k-means $\equiv$ coordinate descent on J: solve in $\vec{c}, \vec{\mu}$ by changing one variable while keeping the others fixed. Each probability solved optimally.
$J(\vec{c}, \vec{\mu})$ is non convex hence local optimality issues
Convergence guaranteed by decreasing J.

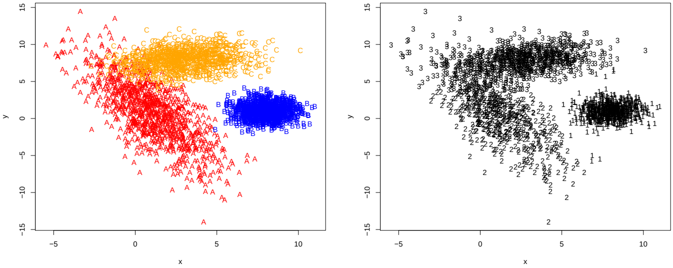


(a) (b) (c)

(d) (e) (f)

In R

```
k <- kmeans(train[,1:2], 3)
> k$centers
      x y
1 8.0123 1.0406
2 1.5735 -0.7285
3 2.1856 7.5940
plot(train[,1:2], type='n')
text(train[,1:2], as.character(k$cluster))
```



Outline

1. k-means

2. Expectation Maximization Algorithm

Mixture Models

We can simplfy complicated distributions $p(\vec{x})$ by introducing latent variables. Then:

$$p(\vec{x}) = \sum_{z} p(\vec{x}, \vec{z}) = \sum_{z} p(\vec{x} \mid \vec{z}) p(\vec{z})$$

$p(\vec{x} \mid \vec{z})$ may be more tractable to express.

Expectation Maximization Algorithm

Given $\{\vec{x}_1, \dots, \vec{x}_m\}$ and no y^i we want to cluster the points.
we wish to model the joint prob. distribution $p(\mathbf{x}^i, \mathbf{z}^i) = p(\mathbf{x}^i \mid \mathbf{z}^i)p(\mathbf{z}^i)$
$\mathbf{z}^i$ are latent random variables

- $z^i \sim \text{Multinomial}(\vec{\phi}), \phi_l \geq 0, \sum_{l=1}^{k} \phi_l = 1$ ($p(z^i = l) = \phi^l$)
- $\mathbf{x}^i \mid \mathbf{z}^i = l \sim N(\mu_l, \Sigma_l)$

Estimation of ϕ, μ, σ (learning)

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^i, \phi, \mu, \Sigma) = \sum_{i=1}^{m} \log \sum_{z^i=l}^{k} p(x^i \mid z^i, \mu, \Sigma) p(z^i, \phi)$$

If z^i known (supervised learning): $\rightsquigarrow$ Gaussian discriminant analysis generalized to $k > 2$ and different variance

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^i \mid z^i, \mu, \Sigma) + \log p(z^i, \phi)$$

$$\phi_l = \frac{1}{m} \sum_{i=1}^{m} I\{z^i = l\}$$

$$\mu_l = \frac{\sum_{i=1}^{m} I\{z^i = l\} x^i}{\sum_{i=1}^{m} I\{z^i = l\}}$$

$$\Sigma_l = \frac{\sum_{i=1}^{m} I\{z^i = l\}(\mathbf{x}^i - \vec{\mu}^i)(\mathbf{x}^i - \vec{\mu}^i)^T}{\sum_{i=1}^{m} I\{z^i = l\}}$$

If z^i not known (unsupervised learning):

repeat

  for i=1...m, l=1...k **do**

    $w_j \leftarrow p(z^i = l \mid x^i, \phi, \mu, \Sigma)$; // (E-step)

  for l=1...k **do**

$$\phi_l = \frac{1}{m}\sum_{i=1}^{m} w_l^i$$

$$\mu_l = \frac{\sum_{i=1}^{m} w_l^i x^i}{\sum_{i=1}^{m} w_l^i} \qquad \text{(M-step)}$$

$$\Sigma_l = \frac{\sum_{i=1}^{m} w_l^i (\mathbf{x}^i - \vec{\mu}^i)(\mathbf{x}^i - \vec{\mu}^i)^T}{\sum_{i=1}^{m} w_l^i}$$

until convergence ;

$$w_j \leftarrow p(z^i = l \mid x^i, \phi, \mu, \Sigma) = \frac{p(\mathbf{x}^i = l \mid z^i = l, \phi, \mu, \Sigma) p(z^i = l, \phi)}{\sum_{l=1}^{k} p(\mathbf{x}^i = l \mid z^i = l, \phi, \mu, \Sigma) p(z^i = l, \phi)}$$

Analysis of EM algorithm

Definition (Convex functions)

$$f : \mathbb{R} \to \mathbb{R} \text{ is convex } \iff f'' \geq 0 \qquad \forall x \in \mathbb{R}$$
$$f : \mathbb{R}^n \to \mathbb{R} \text{ is convex } \iff H \geq 0 \qquad \forall x \in \mathbb{R}^n$$

Theorem (Jensen's inequality)

f convex, x random variable $\Rightarrow E[f(x)] \geq f(E[x])$
(if f strictly convex $\implies E[f(x)] = f(E[x])$ iff $x = E[x]$, ie. $x = c$)

We wish to fit the parameters of a model $p(\mathbf{x}, z)$

$$\ell(\vec{\theta}) = \sum_{i=1}^{m} \log p(\mathbf{x}^i, \vec{\theta}) = \sum_{i=1}^{m} \log \sum_{z}^{i} p(\mathbf{x}^i, z^i, \vec{\theta})$$

z^i not observed $\rightsquigarrow$ opt problem not easy
EM does max likelihood estimation:

- E-step construct lower bound for $\ell(\vec{\theta})$
- M-step optimize the LB

Q_j distribution over z^i ($Q_i(z) \geq 0), \sum_z Q_i(z) = 1$)

$$\begin{aligned}
\ell(\theta) &= \sum_i \log \sum_{z^i} p(x^i, z^i, \theta) \\
&= \sum_i \log \sum_{z^i} Q_i(z^i) \frac{p(x^i, z^i, \theta)}{Q_i(z^i)} \qquad \text{Jensen's ineq. for concave functions} \\
&\geq \sum_i \sum_{z^i} Q_i(z^i) \log \frac{p(x^i, z^i, \theta)}{Q_i(z^i)} \qquad (*)
\end{aligned}$$

(*) gives a LB for $\ell(\theta) \forall Q_i$.
Which Q_i should we choose? Given some parameters θ, try to make q_i highest possible. It holds with equality, i.e.:

$$\begin{aligned}
\frac{p(x^i, z^i, \theta)}{Q_i(z^i)} &= c \\
Q_i(z^i) &\propto p(x^i, z^i, \theta) \\
Q_i(z^i) &= \frac{p(x^i, z^i, \theta)}{\sum_{z^i} p(x^i, z^i, \theta)} \\
&= \frac{p(x^i, z^i, \theta)}{p(x^i, \theta)} = \\
&= p(z^i \mid x^i, \theta)
\end{aligned}$$

Then maximize (*) wrt θ

repeat

  for each i **do**

    $Q_i(z^i) \leftarrow p(z^i \mid x^i, \theta)$; // E-step

  $\theta \leftarrow \arg\max_\theta \sum_i \sum_{z^i} Q_i(z^i) \log \frac{p(x^i, z^i, \theta)}{Q_i(z^i)}$; // M-step

until convergence ;

Convergence: we want to show that $\ell(\theta^t) \leq \ell(\theta^{t+1})$

$$
\begin{aligned}
Q_i^t(z^i) &= p(z^i \mid x^i, \theta^t) \\
\ell(\theta^t) &= \sum_i \sum_{z^i} Q_i^t(z^i) \log \frac{p(x^i, z^i, \theta^t)}{Q_i^t(z^i)} \\
\ell(\theta^{t+1}) &\geq \sum_i \sum_{z^i} Q_i^t(z^i) \log \frac{p(x^i, z^i, \theta^{t+1})}{Q_i^t(z^i)} \qquad \text{because Jensen } \forall \theta \\
&\geq \sum_i \sum_{z^i} Q_i^t(z^i) \log \frac{p(x^i, z^i, \theta^t)}{Q_i^t(z^i)} \qquad \text{because } \theta^{t+1} \text{ max's } \ell(\theta) \\
&= \ell(\theta^t)
\end{aligned}
$$

Thus monotonic convergence. Stop if improvement smaller than a tollerance. EM-algorithm as a coordinate descent on

$$
J(Q, \vec{\theta}) = \sum_i \sum_{z^i} Q_i(z^i) \log \frac{p(x^i, z^i, \theta)}{Q_i(z^i)}
$$

Mixture of Gaussian revisited

E-step:

$$w_i^l = Q_i(z^i = l) = p(z^i = l \mid x^i, \phi, \mu, \Sigma)$$

(prob. of z^i taking l under $Q_i(z^i = l)$)

M-step:

maximize w.r.t. ϕ, μ, Σ:

$$\begin{aligned}
\ell(\theta) &= \sum_i \sum_{z^i} Q_i(z^i) \log \frac{p(x^i, z^i, \theta^t)}{Q_i^t(z^i)} \\
&= \sum_i \sum_{l=1}^{k} Q_i(z^i) \log \frac{p(x^i \mid z^i = l, \theta^t) p(z^i = l, \phi)}{Q_i^t(z^i)} \\
&= \sum_i \sum_{l=1}^{k} w_l^i \log \frac{\frac{1}{2\pi^{n/2}\|\Sigma\|^{1/2}} \exp(-1/2(x^i - \mu_l)\Sigma_l^{-1}(x^i - \mu_l))}{w_l^i}
\end{aligned}$$