(1)

Spring 2013

Clustering; K-means, and EM

Note Title 11/6/2006

Task: set of unlabelled data

$D = \{x_1, \ldots, x_n\}$,

- Decompose into classes $\omega_1, \ldots, \omega_M$ where M is unknown..
- Learn class models $p(x|\omega)$..
- Discovery of new classes (concepts)

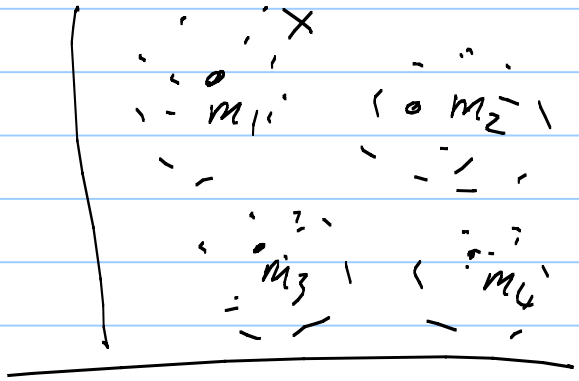
This is a form of unsupervised learning.

How do people learn grammar? Is it innate? Coded in DNA? Or is it learn in an unsupervised manner?

Practical Motivations for unsupervised learning - labelling large datasets is hard, costly, and time consuming.

Spring 2013

(2) K-means algorithm.



Basic Assumption:
the data D is clustered around (unknown) mean values

$m_1, \ldots, m_K$

Seek to associate the data to these means, and to simultaneously estimate the means.

For now, we assume that the number of means is known. K fixed.

Idea: decompose $D = \bigcup_{a=1}^{K} D_a$, where

D_a is associated with mean m_a.

Measure of Fit $F(\{D_a\}) = \sum_{a=1}^{K} \sum_{x \in D_a} (x - m_a)^2$

Want to select $\{m_a\}$, and assignment $x \in D_a$ to minimize fit $F(\cdot)$.

Spring 2013

(3.) Assignment Variable:

$V_{ia} = 1$, if data X_i is assigned to M_a
$= 0$, otherwise.

Constraint $\sum_{a=1}^{K} V_{ia} = 1, \forall K,$
(i.e. each datapoint is assigned to a class).

Deterministic K-means:

1. Initialize a partition $\{D_a^0 : a = 1 \text{ to } K\}$ of the data.
(E.G. randomly choose points x and put them into sets $D_1^0, D_2^0 \ldots D_K^0$ — so that all datapoints are in exactly one set)

2. Compute the mean of each cluster D_a,

$$m_a = \frac{1}{|D_a|} \sum_{x \in D_a} x$$

3. For $i = 1 \text{ to } N$, compute $d_a(x_i) = |x_i - m_a|^2$
Assign x_i to cluster D_{a^*}
s.t. $a^* = \arg\min \{d_1(x_i), \ldots d_K(x_i)\}$

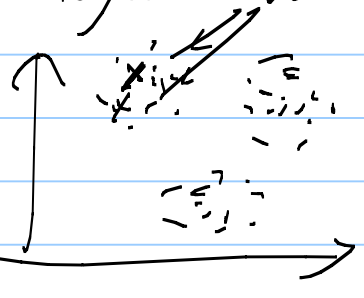
4. Repeat steps 2 & 3 until convergence

Spring 2013

(4.) Comment: The K-mean algorithm is not guaranteed to converge to the best fit of $F(D)$,
It will almost always converge to a fixed point.

Typically, you can improve K-means by giving it different initial conditions. Then select the solution which has best fit.

Evaluate solutions with different values of K. If the number of means (clusters) is too large, then you will usually find some of the mean will be close together. i.e. do postprocessing to eliminate means which are too close to each other.



Spring 2013

(5) A "softer" version of EM. Assign datapoint X_i to each cluster with with probability $(p_1, \ldots p_k)$

1. Initialize a partition. E.G. randomly chose k points as centres $m_1, m_2 \ldots m_k$

2. For $j = 1$ to N,

Compute distances $d_a(x_j) = |X_j - m_a|^2$

Compute the probability that X_j belongs to a

$$P_a(X_j) = \frac{1}{(2\pi\sigma_a^2)^{d/2}} e^{-\frac{1}{2}\sigma_a^2 (X_j - m_a)^2}$$

dimension $\longrightarrow$

3. Compute the mean and variance for each cluster:

$$m_a = \frac{1}{|W_a|} \sum_{x \in W_a} x \, P_a(x)$$

$$\sigma_a^2 = \frac{1}{|W_a|} \sum_{x \in W_a} (x - m_a)^2 P_a(x)$$

Repeat steps 2 & 3 until convergence.

Spring 2013

(6) Deeper Understanding

Modelling data with hidden variables

$$P(x, h, \theta)$$

x: observed data; h: hidden variables; θ: model parameters.

Example: data is generated by a mixture of Gaussian distributions with unknown means and covariance. The variable h indicates which Gaussian generated the data.

$$P(\underline{x} \mid h=a, \theta_a) = \frac{1}{(2\pi)\left|\underline{\underline{\Sigma}}_a\right|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_a)\,\underline{\underline{\Sigma}}_a^{-1}(\underline{x}-\underline{\mu}_a)}$$

$$\theta_a = (\underline{\mu}_a, \underline{\underline{\Sigma}}_a)$$

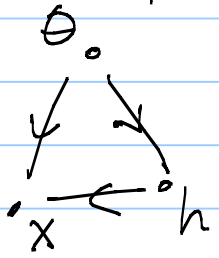
Assume $P(h=a \mid \theta) = \dfrac{e^{\phi_a}}{\sum_c e^{\phi_c}}$.

And a prior probability $P(\theta)$ on $\{\phi_a, \theta_a\}$,

Spring 2013

(7) Mixture of Gaussian example

$$p(x, h, \theta) = p(x|h,\theta)\, p(h|\theta)\, p(\theta)$$



Now perform MAP estimation of θ from data examples $\mathcal{D} = \{x_1, \ldots x_N\}$

$$p(\mathcal{D}|h,\theta) = p(\{x_1 \ldots x_N\}|h,\theta) = \prod_{i=1}^{N} p(x_i|h,\theta)$$

i.i.d. assumption.

$$p(\theta|\mathcal{D}) = \sum_h p(\theta, h|\mathcal{D}).$$

How to estimate $\hat{\theta}$ from $\mathcal{D}$?

Answer (most popular), the Expectation Maximization (EM) algorithm.

Comment: Standard k-means algorithm can be thought of as a limiting case of EM for mixtures of Gaussians — where the covariance is fixed to be the identity matrix.

Spring 2013

(8) The EM Algorithm

$$p(\theta \mid \mathcal{D}) = \sum_h p(\theta, h \mid \mathcal{D}).$$

Define a new distribution $q(h)$

Minimize: $\mathcal{F}(\theta, q) = -\log p(\theta \mid \mathcal{D}) + \sum_h q(h) \log \frac{q(h)}{p(h \mid \theta, \mathcal{D})}$

Kullback-Leibler divergence

Note, the minimum occurs at $\hat{\theta} = \underset{\theta}{\text{ARG MIN}} \{-\log p(\theta \mid \mathcal{D})\}$

$= \underset{\theta}{\text{ARG MAX}}\ p(\theta \mid \mathcal{D})$

and at $\hat{q}(h) = p(h \mid \hat{\theta}, \mathcal{D})$

(Because the Kullback-Leibler divergence attains its minimum at $\hat{q}(h) = p(h \mid \hat{\theta}, \mathcal{D})$.)

Spring 2013

(9) . We can re-express the Free Energy

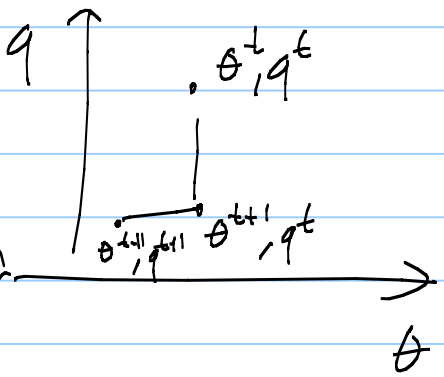
$$\mathcal{F}(\theta, q) = -P(\theta|\mathcal{D}) + \sum_h q(h) \log \frac{q(h)}{p(h|\theta,\mathcal{D})}$$

$$= -\sum_h q(h) \log P(\theta|\mathcal{D}) + \sum_h q(h) \log q(h) - \sum_h q(h) \log p(h|\theta,\mathcal{D})$$

$$= -\sum_h q(h) \log \{P(\theta|\mathcal{D})\, p(h|\theta,\mathcal{D})\} + \sum_h q(h) \log q(h).$$

$$\mathcal{F}(\theta, q) = -\sum_h q(h) \log p(h, \theta|\mathcal{D}) + \sum_h q(h) \log q(h)$$

EM minimizes $\mathcal{F}(\theta, q)$ w.r.t. θ & q alternately



Step 1: Fix q^t,

Set $\theta^{t+1} = \arg\min_\theta \{-\sum_h q(h) \log p(h, \theta|\mathcal{D})\}$

Step 2: Fix θ^t,

set $q^{t+1}(h) = p(h|\theta^t, \mathcal{D})$

Iterate steps 1 & 2 until convergence.

(16) Exponential EM

Spring 2013

Note Title 11/8/2006

EM takes a simple form for exponential distributions.

$$p(\underline{x}, \underline{h} \mid \underline{\lambda}) = \frac{e^{\underline{\lambda} \cdot \underline{\phi}(\underline{x}, \underline{h})}}{Z[\underline{\lambda}]}$$

General exponential ~~form~~, where the $\underline{h}$ are hidden variables / latent variables

$$p(\underline{x} \mid \underline{\lambda}) = \sum_{\underline{h}} p(\underline{x}, \underline{h} \mid \underline{\lambda}).$$

Two steps of EM

Step 1: $$q^{t+1}(\underline{h}) = p(\underline{h} \mid \underline{x}, \underline{\lambda}^t) = \frac{p(\underline{x}, \underline{h} \mid \underline{\lambda}^t)}{\sum_{\underline{h}} p(\underline{x}, \underline{h} \mid \underline{\lambda}^t)}$$

Step 2: $$\underline{\lambda}^{t+1} = \arg\min_{\underline{\lambda}} - \sum_{\underline{h}} q^{t+1}(\underline{h}) \{ \log p(\underline{x}, \underline{h} \mid \underline{\lambda}),$$

$$\underline{\lambda}^{t+1} = \arg\min_{\underline{\lambda}} - \left\{ \sum_{\underline{h}} q^{t+1}(\underline{h}) \underline{\phi}(\underline{x}, \underline{h}) \cdot \underline{\lambda} + \log Z[\underline{\lambda}] \right\},$$

Differentiating gives:

$$\sum_{\underline{x}, \underline{h}} p(\underline{x}, \underline{h} \mid \underline{\lambda}^{t+1}) \underline{\phi}(\underline{x}, \underline{h}) = \sum_{\underline{h}} q^{t+1}(\underline{h}) \underline{\phi}(\underline{x}, \underline{h}).$$

(11)

Spring 2013

Note Title 11/6/2006

- Mixture of Gaussian Example:

$$p(y_i \mid v_{ia}=1, \theta) = \frac{1}{(2\pi)^{d/2} \|\Sigma_a\|^{1/2}} e^{-\frac{1}{2}(y_i-\mu_a)^T \Sigma_a^{-1} (y_i-\mu_a)}$$

datapoint y_i is generated by a^{th} model.

We can write:

$$p(y_i \mid \{v_{ia}\}, \theta) = \frac{1}{(2\pi)^{d/2}} e^{-\sum_a v_{ia} \left\{ \frac{1}{2}(y_i-\mu_a)^T \Sigma_a^{-1} (y_i-\mu_a) + \frac{1}{2}\log \|\Sigma_a\|^{1/2} \right\}}$$

By i.i.d assumption

$$p(\{y_i\} \mid \{v_{ia}\}, \theta) = \prod_i p(y_i \mid \{v_{ia}\}, \theta)$$

$$= \frac{1}{(2\pi)^{nd/2}} e^{-\sum_{ia} v_{ia} \left\{ \frac{1}{2}(y_i-\mu_a)^T \Sigma_a^{-1} (y_i-\mu_a) + \frac{1}{2}\log \|\Sigma_a\|^{1/2} \right\}}$$

Let the prior probability on the $\{v_{ia}\}$ be uniform

Hence $p(\{v_{ia}\}) = \text{Constant}$.

Then

$$p(\{y_i\}, \{v_{ia}\} \mid \theta) = \frac{e}{Z} \times e^{-\sum_{ia} v_{ia} \left\{ \frac{1}{2}(y_i-\mu_a)^T \Sigma_a^{-1} (y_i-\mu_a) + \frac{1}{2}\log \|\Sigma_a\|^{1/2} \right\}}$$

normalization constant.

(13)

Spring 2013

To apply EM to this problem,

E-step: $Q(\theta|\theta^{(t)}) = \sum_V \{\log p(y,V|\theta)\}_{p(V|\theta^t)}$

M-step: Determine $\theta^{(t+1)}$ by maximizing $Q(\theta|\theta^{(t)})$. //

$$p(\{V_{ia}\}|\theta) = \prod_i P_i(V_{ia}|\theta).$$

$$P_i(V_{ia}|\theta) = e^{V_{ia}\{-\frac{1}{2}(\underline{y}_i-\underline{\mu}_a)^T \underline{\underline{\Sigma}}_a^{-1}(\underline{y}_i-\underline{\mu}_a) - \frac{1}{2}\log|\underline{\underline{\Sigma}}_a|\}} \Big/ \sum_b e^{V_{ib}\{-\frac{1}{2}(\underline{y}_i-\underline{\mu}_b)^T \underline{\underline{\Sigma}}_b^{-1}(\underline{y}_i-\underline{\mu}_b) - \frac{1}{2}\log|\underline{\underline{\Sigma}}_b|\}}$$

note: $\sum_a P_i(V_{ia}=1|\theta)=1$

Also $P_i(V_{ia}|\theta)$ is biggest for the model $\underline{\mu}_a, \underline{\underline{\Sigma}}_a$ closest to data point.

Let $\sum_{V_{ia}} V_{ia} P_i(V_{ia}|\theta) = q_{ia}$

E-step:

$$Q(\theta|\theta^{(t)}) = -\sum q_{ia}\left\{\frac{1}{2}(\underline{y}_i-\underline{\mu}_a)^T \underline{\underline{\Sigma}}_a^{-1}(\underline{y}_i-\underline{\mu}_a) + \frac{1}{2}\log|\underline{\underline{\Sigma}}_a|\right\},$$

Spring 2023

M-step: minimize w.r.t. μ_a & Σ_a gives

$$\mu_a = \frac{1}{\sum_i q_{ia}} \left(\sum_i q_{ia} y_i \right) \leftarrow \text{weighted sum}$$

$$\Sigma_a = \frac{1}{\sum_i q_{ia}} \sum_i q_{ia} (y_i - \mu_a)(y_i - \mu_a)^T.$$

Hence, for this case (mixture of Gaussians) we can perform the E and M steps analytically.

Extend the probability model. Let the number of classes be a random variable K

$$P(K) = \frac{e^{-\lambda K}}{Z(\lambda)} \cdot P(y | V, \mu, \Sigma, K) P(K).$$