T-61.5140 Machine Learning: Advanced Probablistic Methods

Jaakko Hollmén

Department of Information and Computer Science
Helsinki University of Technology, Finland
e-mail: Jaakko.Hollmen@tkk.fi
Web: http://www.cis.hut.fi/Opinnot/T-61.5140/

February 28, 2008

Mixture models and the EM algorithm

Mixture models as (very) simple Bayesian networks

- Observed variables and a hidden variable
- Factorization of the joint probability distribution

Mixture models as probabilistic clustering models

- Similarities with k-means algorithm
- Differences with k-means algorithm
- (k-means is NOT a probabilistic model)

k-means algorithm

Ingredients for the k-means clustering algorithm

- Data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
- Prototypes $\mathbf{c}_1, \ldots, \mathbf{c}_K$, $K < n$
- Distance measure $d(\mathbf{x}_n, \mathbf{c}_k)$, usually Euclidean distance

The goal of the k-means algorithm is to use

- k prototypes to represent n data points
- minimize a distortion $\sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mathbf{c}_k\|^2$
- r_{nk} indicates whether $\mathbf{x}_n$ is closest to $\mathbf{c}_k$, $r_{nk} \in \{0, 1\}$

k-means algorithm

k-means algorithm in brief

- Calculate $d(\mathbf{x}_i, \mathbf{c}_j), \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, K$
- Determine r_{nk}, what does this mean?
- Calculate new $\mathbf{c}_k = \frac{\sum_n r_{nk} x_n}{\sum r_{nk}}$
- repeat until convergence: no apparent changes in $c_1, \ldots, c_K$

Example

Mixture models

Mixture model as a very simple Bayesian network

- Observed d-dimensional variables $x_1, \ldots, x_d$
- Hidden variable S
- Factorization of the joint distribution: $P(X, S) = P(S)P(X|S)$
- $P(X) = \sum_{j=1}^{J} P(S = j)P(X|S = j)$

Parameterization

- $P(S = j) = \pi_j,\ \sum_{j=1}^{J} \pi_j = 1, \pi_j \geq 0$
- Mixing coefficients π_j
- The form of component distribution $P(X|S = j)$ depends on X

Mixture models

Gaussian mixture model

- $P(X) = \sum_{j=1}^{J} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)$
- Parameters π_j, μ_j, Σ_j

Mixture of Bernoulli distributions for 0-1 data

- $P(X) = \sum_{j=1}^{J} \pi_j p(\mathbf{x}|\theta_j)$
- Parameters π_j, θ_j, where $\theta = p(x = 1)$

The whole is the sum of its parts

EM algorithm in general

Parameter estimation in the mixture model

- Framework of maximum likelihood (ML)
- Expectation Maximation algorithm (EM)
- EM algorithm is iterative
- converges to a (local) maximum likelihood estimate

EM algorithm, repeat until convergence

- E-step
- M-step

Mixture modeling, 0-1 data

Probability of an observed data vector $\boldsymbol{x}$:

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} \theta_i^{x_i} (1 - \theta_i)^{1 - x_i}$$

Probability of an observed data vector $\boldsymbol{x}$:

$$p(\boldsymbol{x} | \pi_j, \boldsymbol{\Theta}) = \sum_{j=1}^{J} \pi_j p(\boldsymbol{x} | \boldsymbol{\theta}_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1 - x_i}$$

EM algorithm for the 0-1 mixture model

In the E-step, the expected values of the hidden states are estimated

$$p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \theta^k) = \frac{\pi_j^k p(\boldsymbol{x}_n|\theta_j^k)}{\sum_{j'=1}^{J} \pi_{j'}^k p(\boldsymbol{x}_n|\theta_{j'}^k)}$$

EM algorithm for the 0-1 mixture model

In the M-step, the values of the parameters are updated:

$$\pi_j^{k+1} = \frac{1}{N} \sum_{n=1}^{N} p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \theta^k)$$

$$\boldsymbol{\theta}_j^{k+1} = \frac{1}{N\pi_j^{k+1}} \sum_{n=1}^{N} p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \theta^k)\boldsymbol{x}_n.$$

Clustering with a mixture model

- A cluster is associated with each of the component distributions
- The observations are allocated to the clusters according to the maximum posterior probabilities:

$$j^* = \operatorname*{argmax}_j p(j)p(\boldsymbol{x}|j, \boldsymbol{\theta}_j) = \operatorname*{argmax}_j \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i}(1 - \theta_{ji})^{1-x_i}$$