Data Mining, Practice 5

EM-algorithm

S. Nõmm

[1] Department of Software Science, Tallinn University of Technology

04.10.2018

EM-algorithm

Let us consider K-Means from the probabilistic point of view.

- (E-step) Each data point of the set $\mathcal{D}$ has a probability belonging to cluster j, which is proportional to the scaled and exponentiated Euclidean distance to each representative Y_j . In the k-means algorithm, this is done in a "hard" way, by choosing the smallest Euclidean distance to the representative of Y_j.
- (M-step) The center Y_j is the weighted mean over all the data points where the weight is defined by the probability of assignment to cluster j. The hard version of this is used in k-means, where each data point is either assigned to a cluster or not assigned to a cluster (i.e., 0-1 probabilities).

EM-algorithm

Assumption: the data was generated from a mixture of k distributions with probability distributions $\mathcal{G}_1 \dots \mathcal{G}_k$. Each distribution $\mathcal{G}_i$ represents a cluster and is also referred to as a mixture component.

- (E-Step) Given the current value of the parameters in , estimate the posterior probability $P(\mathcal{G}_i|X_j, \Theta)$ of the component $\mathcal{G}_i$ having been selected in the generative process, given that we have observed data point X_j . The quantity $P(\mathcal{G}_i|X_j, \Theta)$ is also the soft cluster assignment probability that we are trying to estimate. This step is executed for each data point X_j and mixture component G_i.
- (M-Step) Given the current probabilities of assignments of data points to clusters, use the maximum likelihood approach to determine the values of all the parameters in Θ that maximize the log-likelihood fit on the basis of current assignments.

Expectation - Maximization

Expectation - Maximization (EM):

- Let x_i denote the visible observed values in case i, and z_i - hidden or missing variables. The goal is to maximize the $\log$ likelihood of the observed data:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) = \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i \mid \theta) \right]$$

- Way around the problem with the sum under the log. Define the complete data log likelihood as is follows

$$\mathcal{L}_c(\theta) = \sum_{i=1}^{N} \log p(x_i, z_i \mid \theta)$$

Note, that this could not be computed due to the fact that z_i are unknown.

EM

- Define expected complete data log likelihood:

$$Q(\theta, \theta^{t-1}) = \mathbb{E}[l_c(\theta) \mid \mathcal{D}, \theta^{t-1}].$$

  here t is the iteration number. Q will be referred as *auxiliary function*.
- **E** step computes the latent values needed to compute $Q(\theta \mid \theta^{t-1})$.
- **M** step optimizes Q with respect to θ.

$$\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$$

EM -algorithm

- Auxiliary function:

$$Q(\theta, \theta^{t-1}) = \sum_i \sum_k \tau_{i,k} \log \pi_k + \sum_i \sum_k \tau_{i,k} \log p(\mathbf{x_i} \mid \theta_{\mathbf{k}}).$$

- **E step:** compute the responsibilities $\tau_{i,k}$ for each i and k:

$$\tau_{i,k} = \frac{\pi_k p(\mathbf{x_i} \mid \theta_{\mathbf{k}}^{\mathbf{t-1}})}{\sum_{k'} \pi_{k'} p(\mathbf{x_i} \mid \theta_{\mathbf{k'}}^{\mathbf{t-1}})}.$$

EM -algorithm

- Optimize Q with respect to $\pi, \mu_{\mathbf{k}}, \mathbf{\Sigma_k}$.
- $$\pi_k = \frac{1}{N}\sum_i \tau_{i,k} = \frac{\tau_k}{N}$$

  where $\tau_k = \sum_i \tau_{i,k}$
- Derive **M step** for the μ_k and Σ_k

$$\mathcal{L}(\mu_k, \Sigma_k) = -\frac{1}{2}\sum_i \tau_{i,k}[\log \mid \Sigma_k \mid + (x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu k)]$$

$$\begin{aligned} \mu_k &= \frac{\sum_i \tau_{i,k} x_i}{\tau_k} \\ \Sigma_k &= \frac{\sum_i \tau_{i,k} x_i x_i^t}{\tau_k} - \mu_{\mathbf{k}} \mu_{\mathbf{k}}^{\mathbf{T}} \end{aligned}$$