# Data Mining: Lecture 4

## Cluster Analysis

S. Nõmm

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

20.09.2022

# 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 $\mathcal{G}_1 \ldots \mathcal{G}_k$. 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 $\Theta$ that maximize the log-likelihood fit on the basis of current assignments.

# Gaussian

- Gaussian or normal distribution. Its probability density function is given by
$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big)$$
where $\mu$ is the mean,$\sigma$ is the variance and $\sqrt{2\pi\sigma^2}$ is the normalization constant.
- Multivariate Gaussian or Multivariate Normal (MVN). Probability density function is given by.
$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big]$$
where $\mu$ is the mean vector, $\Sigma$ is covariance matrix of the data set, $d$ is the dimensionality of the data set.

# Geometric interpretation

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big]$$

- Expression under the exponent is Mahalanobis distance between point $x$ and mean.
- Perform an eigendecomposition of $\Sigma$.

$$\Sigma^{-1} = U^{-T}\Lambda U^{-1} = U\Lambda^{-1}U^T = \sum_{i=1}^{D}\frac{1}{\lambda_i}u_i u_i^T$$

  where $u_i$ is the i'th colum of $U$ ($i$th eigenvector).
- Rewrite Mahalanobis distance and denote $y_i = u_i^T(x-\mu)$

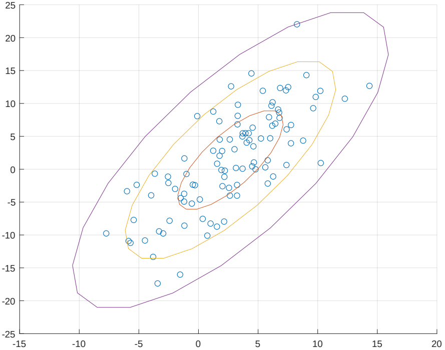
$$(x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)^T\sum_{i=1}^{D}\frac{1}{\lambda_i}u_i u_i^T(x-\mu)$$

$$= \sum_{i=1}^{D}\frac{1}{\lambda_i}(x-\mu)^T u_i u_i^T(x-\mu) = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}.$$

# Geometric interpretation: example

$$(x-\mu)^T\Sigma^{-1}(x-\mu)=\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}.$$

Contours of equal probability density of a gaussian lie along ellipses.



# Likelihood

- **Likelihood:** Roles of parameters and outcomes distinguish likelihood from probability. Probability describes how possible the outcome before data is available, given the values of parameter. Likelihood describe possibility of parameter values given available data.
  - *Discrete:* Let $X$ be a discrete random variable and $p$ its probability mass function then
    $$\mathcal{L}(\theta|x) = p_\theta(x),$$
    is called likelihood function of $\theta$ given the outcome $x$.
  - *Continuous:* Let $X$ be a continuous random variable and $f$ its density function.
    $$\mathcal{L}(\theta|x) = f_\theta(x).$$
    is called likelihood function of $\theta$ given the outcome $x$.

NB! Note the difference with conditional probabilities.

# Prior and posterior

It is presumed that *new* data is expected during the process.

- **Prior** Prior probability is the probability of the event (before collection of a *new* data).
- **Posterior** Posterior probability of the event is the probability of the event (after collection of a *new* data). Easy to memorize: Posterior probability is proportional to likelihood multiplied by prior probability.

# Maximal Likelihood Estimate for MVN

## Theorem

*If one have $N$ samples $x_i \backsim \mathcal{N}(\mu, \Sigma)$ then the maximal likelihood estimate (MLE) for the parameters is given by*

$$\begin{aligned}
\hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} x_i \triangleq \bar{x} \\
\hat{\Sigma} &= \frac{1}{N}\left(\sum_{i=1}^{N} x_i x_i^T\right)
\end{aligned}$$

# Gaussian Mixture Model

- **LVM** - latent variable models
- **Mixture of Gaussians**

$$p(x_i|\theta) = \sum_{k=1}^{K} \tau_k \mathcal{N}(x_i|\mu_k, \Sigma_k).$$

  where $\tau_k$ are the mixing weights, $\mu_k$ are the means and $\Sigma_k$ are the covariance matrices for each base distribution of the mixture.

- Applications:
  - Black box density model to be used in data compression,outlier detection etc.
  - Clustering. Fit the mixture model and then compute $p(z_i = k|x, \theta)$ - The posterior probability that point $i$ belongs to cluster $k$.

# reminder: Bayes rule

- NB! This is short reminder of Bayes theorem.
- We will return to Bayesian theory in the next chapter.
- Let $A$ and $B$ are two events, whereas $P(B) \neq 0$. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Computational example.

# Mixture models for clustering

- The posterior probability that point $i$ belongs to cluster $k$ is referred as the *responsibility of cluster* $k$ *for point* $i$. According to Bayes rule:

$$r_{i,k} = p(z_i = k|x_i, \theta) = \frac{p(z_i = k|\theta)p(x_i|z_i = k, \theta)}{\sum_{k'=1}^{K} p(z_i = k'|\theta)p(x_i|z_i = k', \theta)}$$

- This procedure is referred as *soft clustering*. NB! In the mixture case we never observe variables $z_i$.
- Link to *hard clustering* using MAP estimate

$$z_i^* = \arg\max_k r_{i,k} = \arg\max_k \log p(x_i|z_i = k, \theta) + \log p(z_i = k|\theta).$$

- Presence of latent variables makes complicated to compute ML estimates. Introduce negative log likelihood function.

$$NLL(\theta) = -\frac{1}{N}\log p(\mathcal{D}|\theta).$$

- Let $x$ be the observed variables and $z_i$ be the hidden or missing variables. The goal is to maximize the log likelihood of the observed data.

$$\ell(\theta) = \sum_{i=1}^{N}\log p(x_i|\theta) = \sum_{i=1}^{N}\log\Big[p(x_i, z_i|\theta)\Big].$$

- Complete data log likelihood could not be computed because $z_i$ is unknown.

$$\ell_C(\theta) = \sum_{i=1}^{N}\log p(x_i, z_i|\theta).$$

- Expected complete data log likelihood

$$\begin{aligned} Q(\theta, \theta^{t-1}) &= \mathbb{E}[\ell_c(\theta)|\mathcal{D}, \theta^{t-1}] \\ &= \sum_i\sum_k r_{i,k}\log\tau_k + \sum_i\sum_k r_{r,k}\log p(x_i|\theta_k). \end{aligned}$$

# EM for GMM

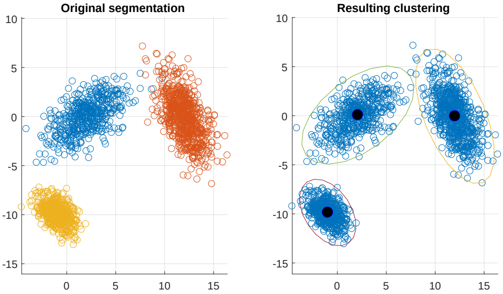
- E step:

$$r_{i,k} = \frac{\tau_k p(x_i|\theta_k^{(t-1)})}{\sum_{k'} \tau_{k'} p\left(x_i|\theta_{k'}^{(t-1)}\right)}$$

- M step: Optimize $Q$ with respect to the $\theta$ and $\tau$.

$$\begin{aligned}
\tau_k &= \frac{\sum_i r_{i,k}}{N} \\
\mu_k &= \frac{\sum_i r_{i,k} x_i}{r_k} \\
\Sigma_k &= \frac{\sum_i r_{i,k}(x_i - \mu_k)(x_i - \mu_k)^T}{r_k} = \frac{\Sigma_i r_{i,k} x_i x_i^T}{r_k} - \mu_k \mu_k^T
\end{aligned}$$

# Example



Original segmentation
Resulting clustering
10
5
0
-5
-10
-15
0
5
10
15

# Clustering overview

- EM estimates the parameters of mixture.
- EM may be referred as parametric method. Model is described by the parameters of clusters.
- How model is described for other clustering techniques? Representative? Hierarchical? Density-based?
- What is clustering model?

# Related topics

- Self organizing maps, will be discussed later (together with Neural Networks).
- Outlier analysis.

# Exercises for self practice

Please note this is not a mandatory Home Assignment, nevertheless some or all of the exercises may be included into Home Assignments.

## Exercises

- Implement EM algorithm.
- Compare performance of your implementation of EM algorithm to the performance of k-means.
- Could you formally verify if given set is gaussian? (question to refresh yor knowledge of probability and statistics)

# Cluster Purity. NB! Not unsupervised any more!!!

- Let $m_{ij}$ represent the number of data points from class (ground-truth cluster) i that are mapped to (algorithm determined) cluster $j$.
- Denote number of data points in true cluster $i$ are by $N_i$,the number of data points in algorithm-determined cluster $j$ by $M_j$.

$$N_i = \sum_{j=1}^{k_d} m_{ij}; \qquad M_j = \sum_{i=1}^{k_t} m_{ij};$$

- For a given algorithm-determined cluster $j$, the number of data points $P_j$ in its dominant class is: $P_j = \max_i m_{ij}$.
- Purity index is defined

$$P_a = \frac{\sum_{j=1}^{k_d} P_j}{\sum_{j=1}^{k_d} M_j}.$$

# Gini index

- Gini index for algorithm determined cluster $j$ is defined:

$$G_j = 1 - \sum_{i=1}^{k_t} \left( \frac{m_{ij}}{M_j} \right)^2 .$$

- Average Gini index is defined as follows:

$$G = \frac{\sum_{j=1}^{k_d} G_j M_j}{\sum_{j=1}^{k_d} M_j} .$$

# Computational exercise

We will consider implementation of EM during the practice.