Machine Learning, Lecture 3: K-means & Gaussians

S. Nõmm

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

K-means

The goal is to cluster the data into K clusters, whereas no labeled data are given.

- Case of unsupervised learning.
- K is the hyperparameter.

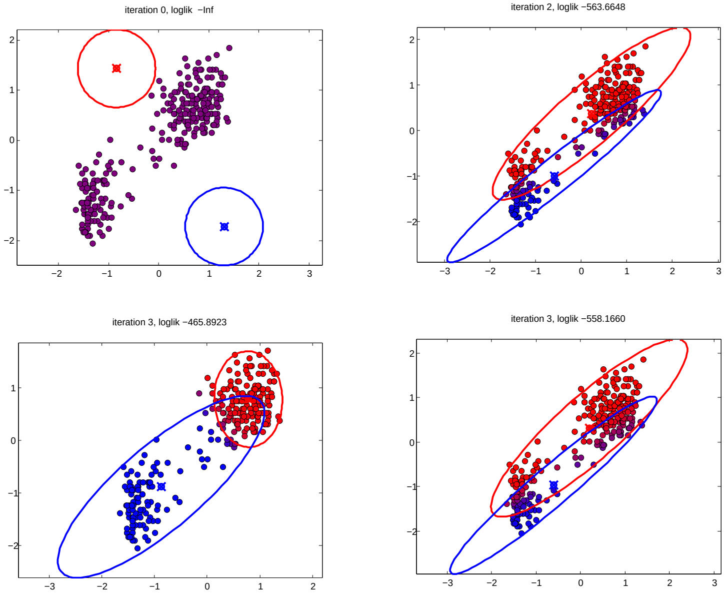
K-means clustering

- Initialization: Generate randomly K points, called *Centroids*. Each centroid represent one of the K classes.
  repeat
  - Associate each point with the cluster represented by the closest centroid. $z_i = \arg\min_k \| x_i - \mu_k \|_2^2$. z_i - is the cluster label.
  - Update centroids for each cluster as

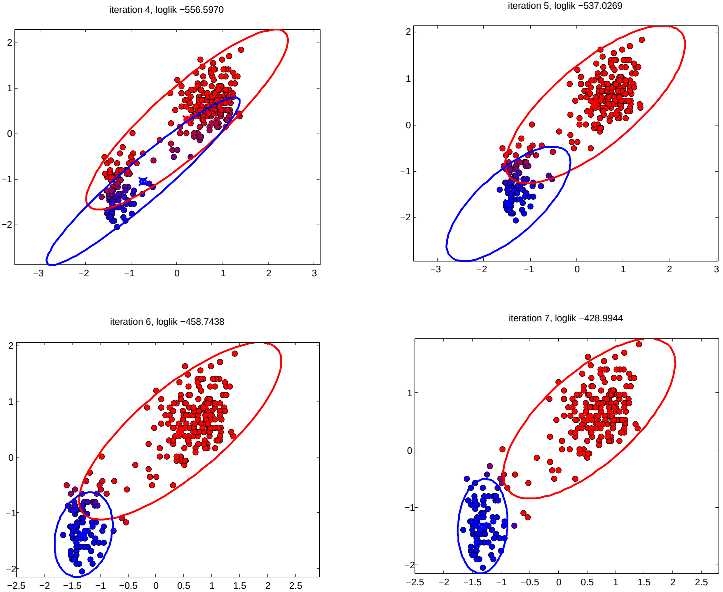
$$\mu_k = \frac{1}{N_k} \sum_{i:z_i=k} x_i$$

  until converged;

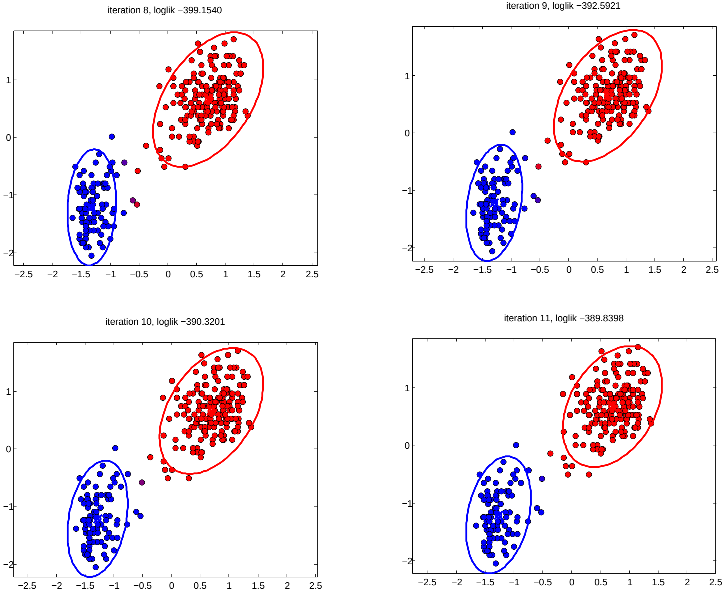
Example 1 of 4



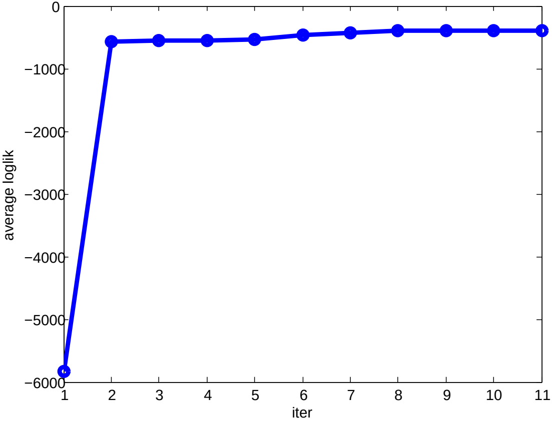
Example 2 of 4



Example 3 of 4



Example 4 of 4, Convergence



K -means algorithm

- K - means algorithm is guaranteed to converge.
- Clustering depend on the particular initialization. Different runs may produce different clusterings. Solution is not global.
- Centroids are the parameters of the model.
- K - means algorithm allows to discover latent structure of the data

K -means algorithm

- K - means algorithm is guaranteed to converge.
- Clustering depend on the particular initialization. Different runs may produce different clusterings. Solution is not global.
- Centroids are the parameters of the model.
- K - means algorithm allows to discover latent structure of the data.
- K - means algorithm works well when the data consists of well-separated Gaussians.
- K - means algorithm performs poorly on the data which does not resemble Gausssian at all.
- Number of classes K should be known or guessed.

K -means implementation in MATLAB environment

```
[idx,C,sumd,D] = kmeans(X,k,Name,Value)
```

- idx - returns cluster indexes for each point.
- C - returns centroids.
- sumd - for each cluster returns the sum of the distances from points to corresponding centroid.
- D - returns distance from each point to every centroid.
- X - initial data to cluster.
- k - number of clusters.
- Name refers to the name of the parameter name to be set.

  `'Distance'`
- Value is the value of the parameter to be set.

  `'cityblock'`

Gaussian

- One-dimensional
  - Do you remember a bell shaped curve?
  - Parameterized by mean μ and variance σ^2
  - Probability density function (pdf):

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

- D-dimensional: Parameterized by mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ.

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \mid \Sigma \mid^{1/2} \exp\Big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\Big]$$

- Derive for the 2- and 3- dimensional cases.

Fitting a Gaussian

Let us suppose, that a sample of n points $\boldsymbol{X} = (x_1, \ldots, x_n)^T$ were independently drawn from some Gaussian.
The goal is to find the mean and the variance of the Gaussian.
(Fitting the Gaussian model to the data.)

- Sample mean is used as the estimate of the mean for the Gaussian
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- sample variance is used as the estimate of the variance of the Gaussian
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Why such estimates are correct?

Probability *versus* Likelihood

- **Data is fixed:** How likely certain set of parameters will result given data set.
- **Parameters are fixed:** What is the probability of drawing given data set with the given set of parameters.

Maximal likelihood estimate

Sometimes referred as maximal likelihood principle.
More formally

- $$\mathcal{L}(\theta \mid x) = P(x \mid \theta)$$
- The goal is to find parameters that maximize the likelihood.
- In many cases natural logarithm of the likelihood function is more easy to deal with. Introduce log-likelihood.

Sufficient statistics

Definition

A statistic $T(X)$ is sufficient for the parameter θ if the conditional probability distribution of the data X, given the statistic $T(x)$ does not depend on the parameter θ

$$P(X = x \mid T(X) = t, \theta) = P(X = x \mid T(X) = t).$$

- A statistic is *sufficient* for a family of probability distributions if the sample from which it was calculated gives no additional information.
- In other words. The value of the *sufficient* statistic (for the parameter) contains all the necessary information to calculate estimate of the parameter.

Example

Consider one dimensional Gaussian: Let us suppose that data points in the sample are drawn independently then the probability of data is:

$$P(\boldsymbol{X} \mid \mu, \sigma^2) = \prod_{i=1}^{n} P(x_i \mid \mu, \sigma^2)$$

$$= \ldots = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2}$$

As a next step: compute log - likelihood

$$\log P(\boldsymbol{X} \mid \mu, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Example

$$\log P(\boldsymbol{X} \mid \mu, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

The last term

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i-1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2$$

Likelihood depends on the sample only through $\sum_{i-1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i$ which are sufficient statistics in this case.

Estimate of the mean μ

Find the partial derivative with respect to μ:

$$\frac{\partial \log P(\boldsymbol{X} \mid \mu\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\Big(\sum_{i=1}^{n} x_i - n\mu\Big)$$

Solve the following equation with respect to μ.

$$\frac{1}{\sigma^2}\Big(\sum_{i=1}^{n} x_i - n\mu\Big) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Estimate of the variance σ^2

Find the partial derivative with respect to σ^2:

$$\frac{\partial P(\boldsymbol{X} \mid \mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2\sigma^2}$$

Solve the following equation with respect to σ^2

$$\frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2\sigma^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Multivariate case

- Mean estimate

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Sample covariance

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T.$$