K-means clustering, Gaussian Mixture Model

Kairit Sirts

21.02.2014

K-means clustering

- Begin by initializing randomly K points.
- These will be the cluster **centroids**.
- Attach each point to the closest centroid.

$$z_i = \arg\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2$$

- z_i is the cluster label for point $\mathbf{x}_i$.
- Proceed until no changes made or certain number of iterations done:
  - Recompute the mean of each cluster - these will be the new centroids.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:x_i=k} \mathbf{x}_i$$

  - Reattach each point to the closest centroid.

Example

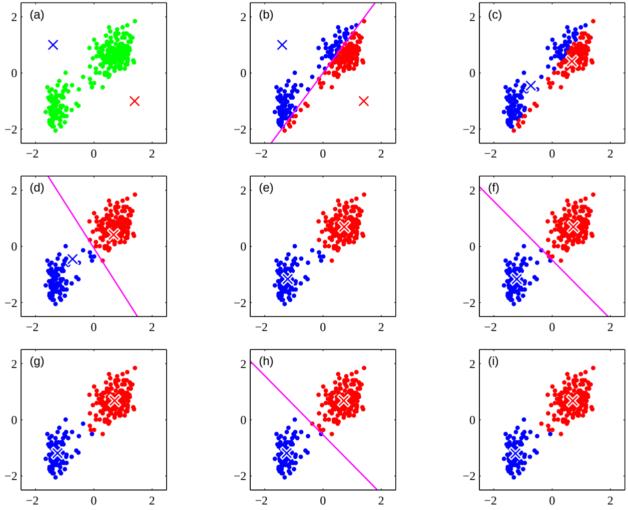


Figure 9.1 from Pattern Recognition and Machine Learning (Bishop).

K-means algorithm

- It is an **unsupervised** learning method - no labelled data is needed.
- It is used to solve **clustering** problems where we want to discover latent structure from unlabelled data.
- K-means algorithm is guaranteed to **converge** - it will find a stable solution.
- This solution is **not** guaranteed to be globally optimal - different runs may produce different clusterings, depending on the particular initialization.
- K is the hyperparameter defining how many clusters will be found.
- Centroids are the parameters of the model learned during training.

Some remarks

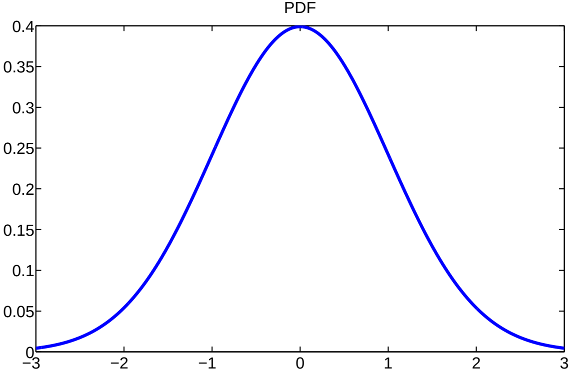
- It is a very well-known and widely used clustering algorithm.
- K-means works well when the data consists of well-separated Gaussians.
- It works pretty poorly when the data does not resemble Gaussian at all.
- We have to know or guess the number of clusters K.

Probabilistic approach

One-dimensional Gaussian

- Parameterized by mean μ and variance σ^2
- Probability density function (pdf):

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

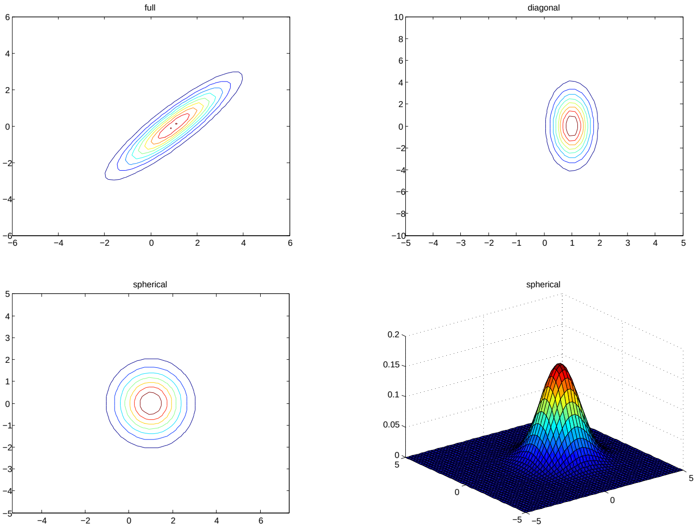


D-dimensional Gaussian

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

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

2-dimensional Gaussian example



Fitting a Gaussian

- Assume we have a dataset with n points $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$.
- Assume these points were drawn **independently** from some Gaussian.
- Finding the mean and variance of this Gaussian is **fitting the model to the data**.
- The model in this context is **probabilistic** - a Gaussian distribution.
- How do we find the mean and variance?

Estimated Gaussian parameters

- Sample mean:

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

- Sample variance:

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

Where do these estimates come from?

- We can derive them using **maximum likelihood** (**ML**) principle.
- ML approach gives us the mean and variance that **maximize** the probability of the sample points.
- This gives us an **estimate** of the parameter, not the true value.
- ML principle is widely used in machine learning for deriving formulas for learning model parameters.

General recipe for applying ML principle

- Take the formula of data probability according to the model.
- Take the (natural) logarithm of it.
- Drop the constant terms.
- Take the partial derivative with respect to the parameter.
- Set the derivative to zero.
- Solve for parameter value.

Probability of data

- If the data points are drawn **independently** as we assumed then the total probability of the data is the product of point probabilities:
- Let's take one-dimensional data for now:

$$
\begin{aligned}
P(\mathbf{X}|\mu,\sigma^2) &= \prod_{i=1}^{n} P(x_i|\mu,\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}
\end{aligned}
$$

Probability and likelihood

- In the context of ML parameter estimation we call this probability **data likelihood**.
- Probability and likelihood are essentially the same thing.
- The subtle difference lies in the assumption of **what is being fixed**.
- When talking about likelihood the **data is fixed** and the probability formula is a **function of parameters**:
  - We can compute how likely a certain set of parameters gave rise to this data.
- When talking about probability the **parameters are fixed**:
  - We can compute the probability of drawing this data using the given parameters.

Computing the log-likelihood

- We do it because this replaces the product with summation and thus makes the derivative computation easier.
- We can do it because the logarithm is a monotonically increasing function having the extremums at the same points where the probability density function.

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

Sufficient statistics

- The last term with summation can be expanded:

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

- Likelihood depends on data set only through two quantities: $\sum_{i=1}^{n}x_i^2$ and $\sum_{i=1}^{n}x_i$.
- These are called **sufficient statistics**.
- When we know sufficient statistics then we know all the information that is possible to obtain from the data to make parameter estimates.

Estimate for mean μ

- Take the partial derivative from log-likelihood with respect to μ:

$$\frac{\partial \log P(\mathbf{X}|\mu, \sigma^2)}{\partial \mu} = -\frac{1}{2\sigma^2}\left(-2\sum_{i=1}^{n} x_i + 2n\mu\right)$$
$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

- Set it two 0:

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

Estimate for variance σ^2

- Take the partial derivative from log-likelihood with respect to σ^2:

$$\frac{\partial \log P(\mathbf{X}|\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\left(-\frac{1}{\sigma^4}\right)$$

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

- Set it to 0:

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

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

Unbiased estimators

- It is possible to show that $E\left[\hat{\mu}\right] = \mu$
- Thus $\hat{\mu}$ is the **unbiased** estimator for true mean μ.
- However, the expected value of the MLE variance is:

$$E\left[\hat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2$$

- Thus, this estimate is **biased** - MLE underestimates the variance.
- It can be shown that with a small modification the variance estimator becomes unbiased:

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

Multivariate case

- For deriving estimates for multivariate data we need to use matrix algebra.
- Otherwise the principles are similar to the univariate case.
- If you are interested in the derivations, I can give you pointers.
- Mean estimate:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

- Sample covariance:

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