Clustering

Gaussian mixture model-based clustering

Overview

- Hard vs. soft clustering
- A generative model for clustering:
  - Gaussian mixture model
- Clustering as parameter estimation
- Parameter estimation via Expectation–Maximization

Hard vs. soft clustering

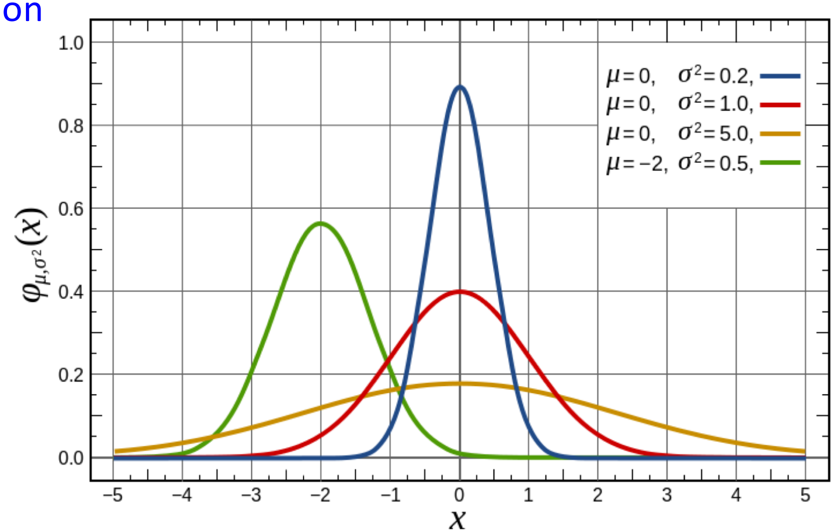
- K-means is hard clustering
  - At each iteration, a data point is assigned to **one and only one** cluster
- We can do soft clustering based on Gaussian mixture models
  - Each cluster is represented by a distribution (in our case a Gaussian)
  - The **probability** that a point belongs to a particular cluster is proportional to the cluster's Gaussian density at that point
    - A point has a non-zero probability of belonging to each cluster

Gaussian distribution

- Gaussian distribution

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

μ : Mean
σ : Standard deviation



Representation of Clusters

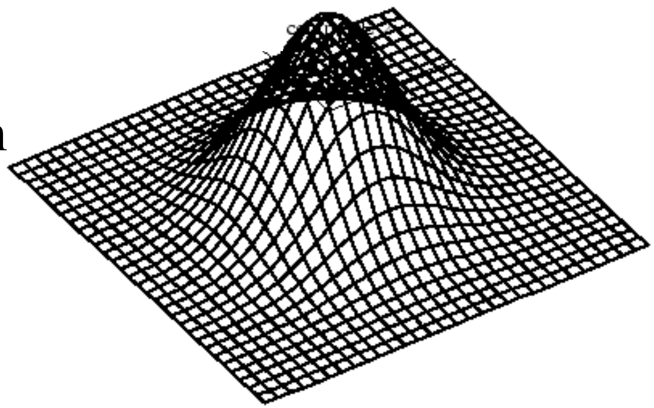
- in the EM approach, we'll represent each cluster using a *p*-dimensional multivariate Gaussian

$$N_j(\vec{x}_i) = \frac{1}{\sqrt{(2\pi)^p \,|\, \Sigma_j \,|}} \exp\left[-\frac{1}{2}(\vec{x}_i - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}_i - \vec{\mu}_j)\right]$$

where

$\vec{\mu}_j$ is the mean of the Gaussian

Σ_j is the covariance matrix



this is a representation of a Gaussian in a 2-D space

Cluster generation

- We model our data points with a generative process
- We assume the data is generated by a mixture of the Gaussians
- Each point is generated by:
  - Choosing a cluster k (where k is one of 1,2,…,K) by sampling from probability distribution over the clusters
    - Prob(cluster k) = P_k
  - Sampling a point from the Gaussian distribution N_k

Clustering as parameter estimation

- Given parameter values for a Gaussian mixture model, we can compute the probability of a point belonging to a particular cluster
- But how do we get the parameter values?
  - Easy if we knew the true assignment of points to clusters
  - But cluster assignments are **hidden** random variables
  - We can use the Expectation–Maximization (EM) algorithm
    - Computes maximum likelihood parameters when some variables are hidden

EM maximizes the log likelihood

- the EM algorithm will try to set the parameters of the Gaussians, Θ , to maximize the log likelihood of the data, X

$$\text{log likelihood}(X \mid \Theta) = \log \prod_{i=1}^{n} \Pr(\vec{x}_i)$$

$$= \log \prod_{i=1}^{n} \sum_{k=1}^{K} P_k N_k(\vec{x}_i)$$

$$= \sum_{i=1}^{n} \log \sum_{k=1}^{K} P_k N_k(\vec{x}_i)$$

Parameters of the Gaussian mixture model

- the parameters of the model, Θ , include the means, the covariance matrix and sometimes prior weights for each Gaussian
- here, we'll assume that the covariance matrix is fixed; we'll focus just on setting the means and the prior weights

EM Clustering: Hidden Variables

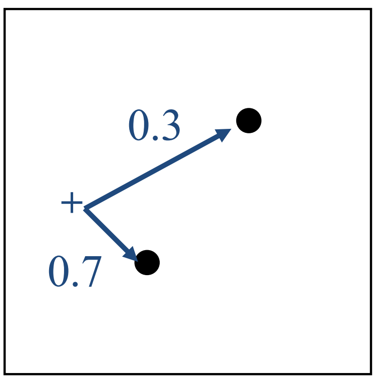
- on each iteration of *K*-means clustering, we had to assign each instance to a cluster
- in the EM approach, we'll use hidden variables to represent this idea
- for each instance $\vec{x}_i$ we have a set of hidden variables $Z_{i1}, \ldots, Z_{iK}$
- we can think of Z_{ij} as being 1 if $\vec{x}_i$ is a member of cluster *j* and 0 otherwise (it is an **indicator** random variable)

EM Clustering: the E-step

- recall that Z_{ij} is a hidden variable which is 1 if N_j generated $\vec{x}_i$ and 0 otherwise
- in the E-step, we compute h_{ij}, the expected value of this hidden variable

$$h_{ij} = E(Z_{ij} \mid \vec{x}_i) = \Pr(Z_{ij} = 1 \mid \vec{x}_i) = \frac{P_j N_j(\vec{x}_i)}{\sum_{l=1}^{K} P_l N_l(\vec{x}_i)}$$

$$= \frac{\mathbb{P}(\vec{x}_i, Z_{ij} = 1)}{\mathbb{P}(\vec{x}_i)}$$



assignment

EM Clustering: the M-step

- given the expected values h_{ij} , we re-estimate the means of the Gaussians and the cluster probabilities

$$\vec{\mu}_j = \frac{\sum_{i=1}^{n} h_{ij}\vec{x}_i}{\sum_{i=1}^{n} h_{ij}} \qquad P_j = \frac{\sum_{i=1}^{n} h_{ij}}{n}$$

- can also re-estimate the covariance matrix if we're not treating it as fixed

EM Clustering – Overall algorithm

- Initialize parameters (e.g., means)
- Loop until convergence
  - E-step: Compute expected values of Z_{ij} values given current parameters
  - M-step: Update parameters using $E[Z_{ij}]$ values
    - Means
    - Cluster probabilities

Comparing K-means and GMMs

- K-means
  - Hard clustering
  - Optimizes within cluster scatter
  - Requires estimation of means

- GMMs
  - Soft clustering
  - Optimizes likelihood of the data
  - Requires estimation of mean (and possibly covariance) and prior cluster probabilities

Summary

- Gaussian mixture model-based clustering is a probabilistic extension of K-means clustering
- Soft clustering instead of hard clustering
- With Gaussian mixture models, clustering = parameter estimation
- Parameter estimation can be performed by the EM algorithm