ANLY-601
Advanced Pattern Recognition

Spring 2018

L13 – Clustering and Mixture Models

GEORGETOWN UNIVERSITY

Clustering

Unsupervised classification

Central problem is defining a cluster

Parametric: use a cluster criterion, either a “distance” or “distortion function” (does not have to be a mathematical metric) for closeness, or a parametric model of the distribution

Model-based clustering produces <u>representative values</u> of the data in the cluster. For example, the centroid. Representative values are substitutes for the data in lossy data compression (e.g. scalar and vector quantization). Hence clustering and lossy coding are related.

Model-Based Clustering

General algorithm properties

Data – $x_1, x_2, \ldots, x_N$ each to be assigned to one of L clusters or classes $\omega_1, \ldots, \omega_L$

Assignment of the i^{th} data point to a cluster is denoted

$$\omega_{k_i} \quad k_i \in \{1, \ldots, L\}$$

Classification of entire dataset is denoted

$$\Omega = \left\{ \omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_N} \right\}$$

There's a cluster criterion function $J(\Omega)$

and a best clustering

$$\Omega^* = \arg\min_{\Omega} J(\Omega)$$

Model-Based Clustering

General Properties (cont'd)

There are a set of parameters associated with the kth cluster – denoted θ_k. The parameters may include the mean of the data in the cluster, the covariance of the data in the cluster, and other summary statistics.

The union of all these parameters for all clusters is $\boldsymbol{\Theta}$.

The clustering criterion function is generally a function of the parameters

$$J = J(\Theta, \Omega)$$

and the best clustering model simultaneously minimizes *J* over the cluster assignment and the parameters

$$(\Theta^*, \Omega^*) = \underset{\Theta, \Omega}{\arg\min} \; J(\Theta, \Omega)$$

Example K-means

Classic algorithm. (MacQueen. Some methods for classification and analysis of multivariate observations, in Proce. 5th Berkeley Symposium on Mathematical Stat. and Prob., vol 3, 1967.)

Criterion function is the average squared distance between the data and the cluster “means”

$$J = \sum_{i=1}^{L} \frac{N_i}{N} \; \frac{1}{N_i} \sum_{j=1}^{N_i} \left\| x_j^{(i)} - m_i \right\|^2$$

where $x_j^{(i)}$ is the jth data point assigned to cluster *i* , and N_i is the number of data points assigned to cluster *i* .

The m_i are the “means” associated with the clusters – collectively they comprise the parameters Θ associated with the clusters. (We have yet to show that these are statistical means, hence the quotation marks.)

Note that $\frac{N_i}{N}$ is the fraction of data points in the ith cluster.
It’s an estimator of the cluster prior P_i.

K-Means Iterative Optimization

- Freeze the means m_i , and find the assignment Ω that minimizes *J.* This assigns x_j to the cluster whose mean m_r is the closest in Euclidean distance to x_i.

$$\|x_j - m_r\| < \|x_j - m_k\| \quad \forall k \neq r$$

- Freeze the cluster assignments, move the means to the centroid of each cluster

$$m_r = \frac{1}{N_r}\sum_{i=1}^{N_r} x_i^{(r)}$$

Since the cluster assignments depend on the m_k, and the means depend on the cluster assignment Θ, we must iterate these two steps.

LGB

The k-means algorithm arose in the statistics community. It was independently discovered in the EE community where it’s known as the Lloyd, Gray, Buzo (LGB) algorithm used to design vector quantizers for lossy coding.

The algorithm can be generalized by substituting for the squared Euclidean distance, any distortion function that is bounded below by zero

$$d(x_i\,, \theta_r) \geq 0$$

The criterion function is then

$$J = \sum_{i=1}^{L} \frac{N_i}{N} \; \frac{1}{N_i} \sum_{j=1}^{N_i} d(x_j, \theta_i)$$

LBG Optimization

As before the optimization is iterative and proceeds in two steps

- Freeze the cluster parameters θ_k and assign each datapoint x_i to the cluster with lowest distortion

$$x_i \in \omega_r \quad \text{where} \quad d(x_i, \theta_r) < d(x_i, \theta_j), \quad \forall\, j \neq r$$

- Freeze the cluster assignments and adjust the parameters to minimize the distortion in each cluster. The resulting values of the parameters θ_k are called the *generalized centroids*, they depend on the distortion function (may not be actual centroid)

$$\theta_r = \arg\min_{\theta} \ \tfrac{1}{N_r} \sum_{i=1}^{N_r} d(x_i^{(r)}, \theta)$$

LBG Convergence

Each of the two optimization operations either lowers J, or leaves it unchanged. Since J is bounded below, the algorithm converges to a (local) minimum of J.

For a finite number of data points, J stops changing after a finite number of steps. For L clusters and N datapoints, there are L^N different cluster assignments Ω. Each such assignment, together with its optimal set of parameters Θ^* **(Ω)** produces a particular value of J. Since each step lowers J or leaves it unchanged, the algorithm must arrive at a local minimum of J in L^N steps or less.

Note – the convergence means J comes to rest at a local minimum. It’s conceivable that Ω may continue to change (a finite number of times).

K-means Boundaries

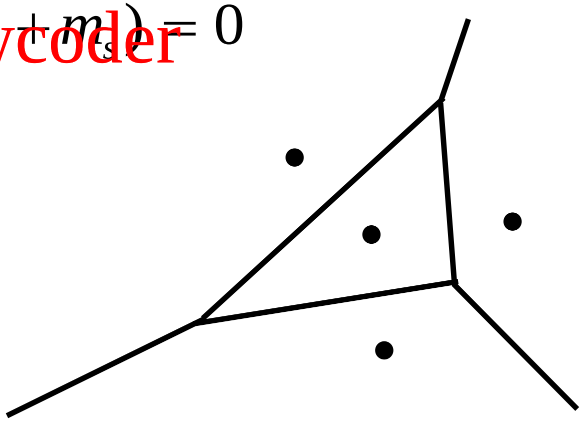
For Euclidean distance distortion, boundaries between clusters are piecewise linear and bisectors perpendicular to the line segment between pairs of means. A point *x* on the boundary between clusters *r* and *s* must satisfy

$$\|x - m_r\|^2 = \|x - m_s\|^2$$

expanding the quadratics gives

$$2\,x^T\,(m_r - m_s) + (m_r - m_s)^T\,(m_r + m_s) = 0$$

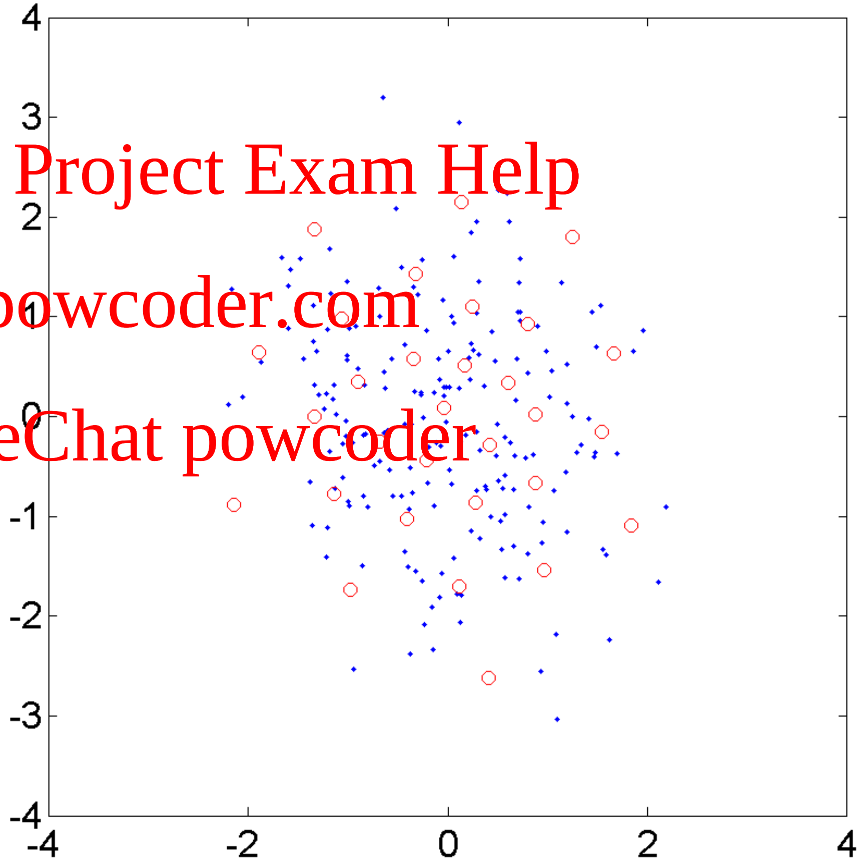
which is linear is *x.* (You can prove the bisector property yourself.) The set of boundaries is called a Voronoi tessellation.



K-means Partition and Mean Location

The algorithm tends to concentrate the means where p(x) is high.

2-D Gaussian data (blue dots) and 30 means (red circles) placed by k-means iterating 65 times over a dataset of 1000 points.



Relation to Gaussian Mixture Models

Take a Gaussian mixture model with spherical components all with the same variance. Don’t fit the component variance, but instead regard it as a ‘knob’. The model is

$$p(x) = \sum_{i=1}^{L} P_i \; p(x|i), \qquad p(x|i) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} \exp(-\tfrac{1}{2\sigma^2} \left|x - m_i\right|^2)$$

the posterior cluster probabilities data point *x* are

$$p(i|x) \;=\; \frac{P_i \; p(x|i)}{\sum_j P_j \; p(x|j)} = \frac{1}{1 + \sum_{j \neq i} \frac{P_j \; p(x|j)}{P_i \; p(x|i)}}$$

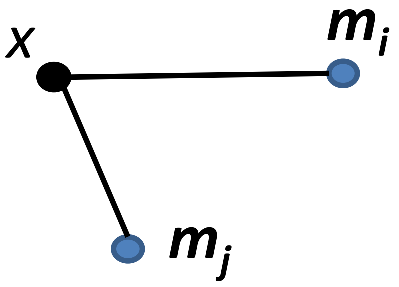
K-means and Gaussian Mixtures

We have the cluster posteriors

$$p(i|x) = \frac{P_i\, p(x|i)}{\sum_j P_j\, p(x|j)} = \frac{1}{1+\sum_{j\neq i} \frac{P_j\, p(x|j)}{P_i\, p(x|i)}}$$

with

$$\frac{P_j\, p(x|j)}{P_i\, p(x|i)} = \frac{P_j}{P_i} \exp -\frac{1}{2\sigma^2}\left(\left|x-m_j\right|^2 - \left|x-m_i\right|^2\right)$$



K-means and Gaussian Mixtures

Take the limit $\sigma^2 \to 0$. It's quick to show that, provided non of the priors P_i are zero,

$$\lim_{\sigma^2 \to 0} p(i \mid x) = \begin{cases} 1 & if \quad |m_i - x| < |m_k - x| \quad \forall k \neq i \\ 0 & otherwise \end{cases}$$

So the cluster posteriors become *hard cluster assignments* (0 or 1) with the same cluster assignment as in k-means!

K-means and Gaussian Mixtures

The optimal position of the mean for the ith cluster is (recall the EM algorithm for mixture model fitting)

$$m_i = \frac{\sum_{l=1}^{N} p(i \mid x_l)\ x_l}{\sum_{k} p(i \mid x_k)}$$

with our limiting values of the posteriors, this becomes

$$\lim_{\sigma^2 \to 0} m_i = \frac{1}{N_i} \sum_{l=1}^{N_i} x_l^{(i)}$$

So the M-step of the EM algorithm becomes the cluster mean adjustment step of the k-means algorithm.

Thus, the EM algorithm for fitting spherical Gaussian mixture models reduces to the k-means clustering algorithm in the limit that the component variances are kept the same and taken to zero.

17 89

GEORGETOWN UNIVERSITY