Clustering: Mixture Models

Data Science: Jordan Boyd-Graber
University of Maryland
SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

Mixture Models

K-means associates data with cluster centers.

What if we actually modeled the data?

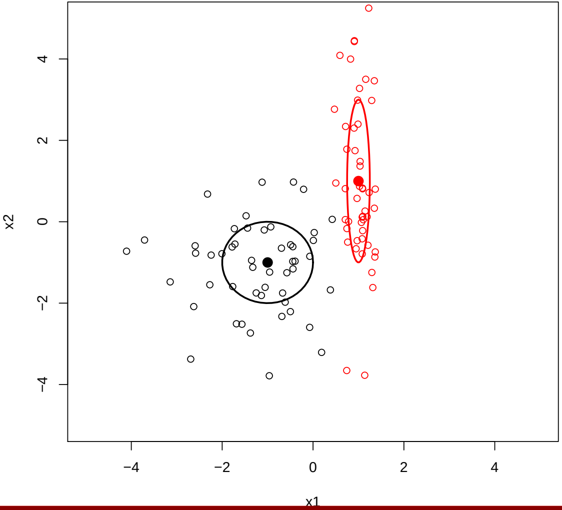
- real-valued data
- observation $\mathbf{x}_i$ in cluster c_i
- have K clusters
- model each cluster with a Gaussian distribution

$$\mathbf{x}_i \,|\, c_i = k \sim N(\mu_k, \Sigma_k)$$

- μ_k is mean vector, Σ_k is covariance matrix

Mixture Models

Gaussian mixture model ($K = 2$):



Mixture Models

Why mixture models?

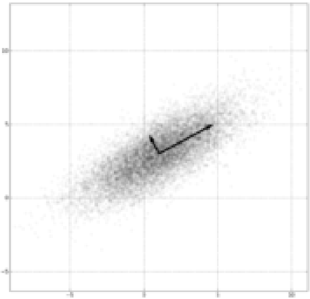
- more flexible: can account for clusters with different shapes
- have data model (will be useful for choosing K)
- less sensitive to data scaling

Multivariate Gaussian

Multivariate Gaussian distribution for $\mathbf{x} \in \mathbb{R}^d$:

$$p(\mathbf{x} \,|\, \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} \, |\Sigma|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

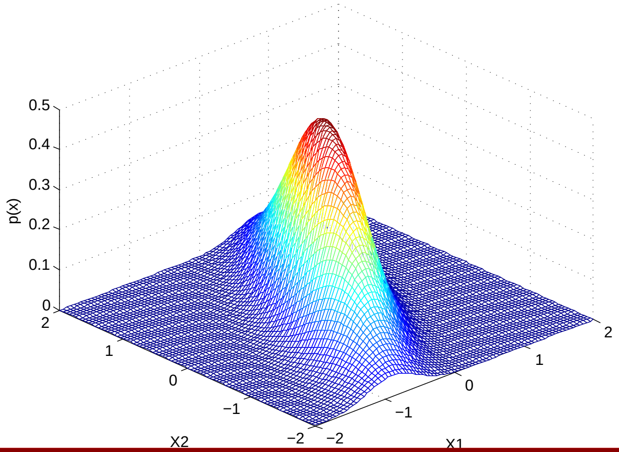
- μ is vector of means
- Σ is covariance matrix



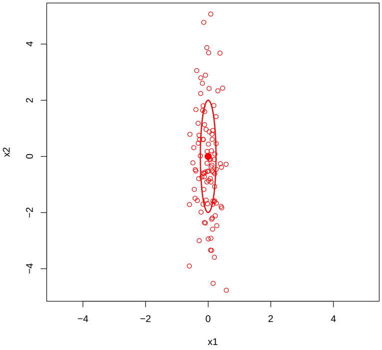
Credit: Wikipedia

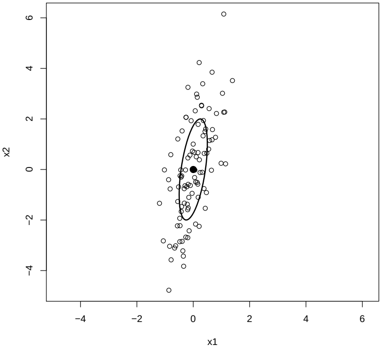
Multivariate Gaussian

pdf when $\mu = [0, 0]$ and $\Sigma = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.3 \end{bmatrix}$:



Multivariate Gaussian





Fitting a Mixture Model

Mixture model:

- observation $\mathbf{x}_i$ in cluster c_i with K clusters
- model each cluster with a Gaussian distribution

$$\mathbf{x}_i \,|\, c_i = k \sim N(\mu_k, \Sigma_k)$$

How do we find $c_1, \ldots, c_n$ (clusters) and $(\mu_1, \Sigma_1), \ldots, (\mu_K, \Sigma_K)$ (cluster centers)?

Fitting a Mixture Model

First, let's simplify the model:

- covariance matrices have only diagonal elements,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & \sigma_K^2 \end{bmatrix}$$

- set $\sigma_1^2 = \cdots = \sigma_K^2$, suppose known

Fitting a Mixture Model

Next, use a method similar to K-means:

- start with random cluster centers
- associate observations to clusters by (log-)likelihood,

$$\ell(\mathbf{x}_i \mid c_i = k) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(\prod_{j=1}^{d}\sigma_{k,j}^2\right) - \frac{1}{2}\sum_{j=1}^{d}(x_{i,j} - \mu_{k,j})^2/\sigma_{k,j}^2$$

$$\propto -d\log(\sigma_k) - \frac{1}{2\sigma_k^2}\sum_{j=1}^{d}(x_{i,j} - \mu_{k,j})^2$$

$$\propto -\sum_{j=1}^{d}(x_{i,j} - \mu_{k,j})^2$$

- refit centers $\mu_1, \ldots, \mu_K$ given clusters by

$$\mu_{k,j} = \frac{1}{n_k}\sum_{c_i = k} x_{i,j}$$

- recluster observations...

Fitting a Mixture Model

clustering with K-means

minimize distance

$$d(\mathbf{x}_i, \mu_k) = \sqrt{\sum_{j=1}^{d} (x_{i,j} - \mu_{k,j})^2}$$

clustering with GMM

maximize likelihood

$$\ell(\mathbf{x}_i \mid c_i = k) \propto -\sum_{j=1}^{d} (x_{i,j} - \mu_{k,j})^2$$

update means with K-means

use average

$$\mu_{k,j} = \frac{1}{n_k} \sum_{c_i = k} x_{i,j}$$

update means with GMM

use average

$$\mu_{k,j} = \frac{1}{n_k} \sum_{c_i = k} x_{i,j}$$

Fitting a Mixture Model

OK, now what if

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & \sigma_K^2 \end{bmatrix}$$

and $\sigma_1^2, \dots, \sigma_K^2$ can take different values?

- use same algorithm
- update μ_k and σ_k^2 with maximum likelihood estimator,

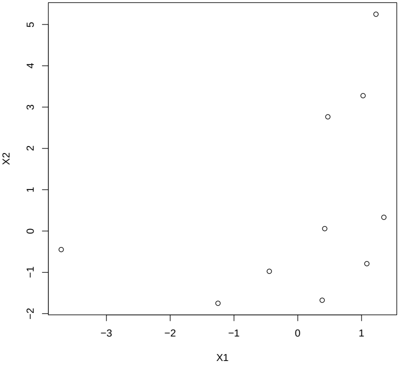
$$\mu_{k,j} = \frac{1}{n_k} \sum_{c_i = k} x_{i,j}$$

$$\sigma_{k,j}^2 = \frac{1}{n_k} \sum_{c_i = k} (x_{i,j} - \mu_{k,j})^2$$

Fitting a Mixture Model

Data:

x_1	x_2
-3.7	-0.4
0.4	0.1
0.4	-1.7
-0.4	-1.0
-1.3	-1.7
1.0	3.3
1.2	5.2
1.3	0.3
1.1	-0.8
0.5	2.8



Fitting a Mixture Model

- pick centers and variances, $\mu_1 = [-1, -1]$, $\sigma_1^2 = [1, 1]$, $\mu_1 = [1, 1]$, $\sigma_1^2 = [1, 1]$
- compute (proportional) log likelihoods,

$$\ell(\mathbf{x}_i \,|\, c_i = k) = -\sum_{j=1}^{d} \log(\sigma_j) - \frac{1}{2} \sum_{j=1}^{d} (x_{i,j} - \mu_{k,j})^2 / \sigma_{k,j}^2$$

x_1	x_2	$k=1$	$k=2$
-3.7	-0.4	-3.8	-12.1
0.4	0.1	-1.6	-0.6
0.4	-1.7	-1.2	-3.8
-0.4	-1.0	-0.2	-3.0
-1.3	-1.7	-0.3	-6.3
1.0	3.3	-11.2	-2.6
1.2	5.2	-22.0	-9.0
1.3	0.3	-3.6	-0.3
1.1	-0.8	-2.2	-1.6
0.5	2.8	- 8.2	-1.7

Fitting a Mixture Model

- fit new means and variances:

$$\mu_1 = [-1.3, -1.2]$$
$$\sigma_1^2 = [3.1, 0.4]$$
$$\mu_2 = [0.9, 1.8]$$
$$\sigma_2^2 = [0.2, 5.4]$$

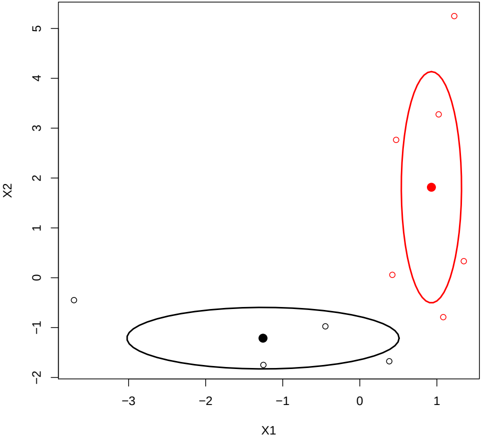
- compute new distances...

Fitting a Mixture Model

x_1	x_2	$k=1$	$k=2$
-3.7	-0.4	-1.8	-70.8
0.4	0.1	-2.7	-1.0
0.4	-1.7	-0.8	-2.0
-0.4	-1.0	-0.3	-6.8
-1.3	-1.7	-0.5	-16.6
1.0	3.3	-27.4	-0.1
1.2	5.2	-55.9	-1.3
1.3	0.3	-4.3	-0.7
1.1	-0.8	-1.2	-0.6
0.5	2.8	-21.3	-0.7

No change, so clusters are final

Fitting a Mixture Model



Limitations of *k*-means / mixture models

k-means is fast and simple, but . . .

- What if your data are discrete?
- What if each data point has more than one cluster? (digits vs. documents)
- What if you don't know the number of clusters?

Wrapup

- Clustering helps discover patterns
- k-means is a simple approach
- Gaussian mixture models more probabilistic foundation