14. Unsupervised Learning

Target values not available.

14.1 K-means

Computing K means of data generated from K Gaussian distributions.

1. Choose k = number of clusters.

2. Put initial partition that classifies the data into k clusters. Randomly or systematically.

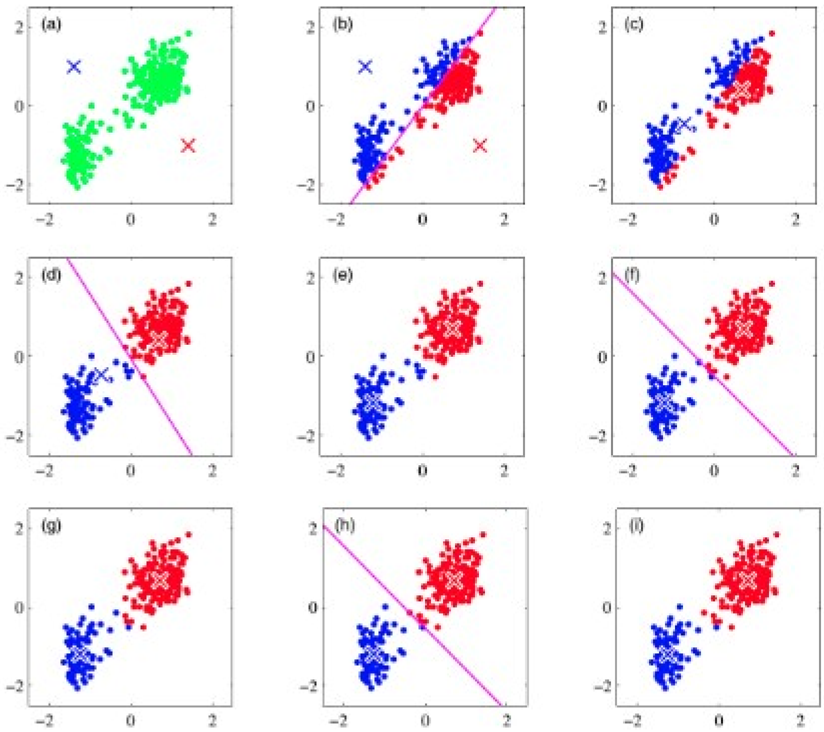
3. Take each sample in sequence and compute its distance from centroid, if it is nearer to another cluster switch and update.

4. Repeat until convergence.

Convergence occur if:

- the sum of distances from each training sample is decreasing

- there are finite partitions



Not robust to outlirs, very far data can influece centroid.

Improvements:

- use K-means clustering only if there are many data available

- use median instead of mean

- define better distance functions

14.2 Gaussian Mixture Model

Mixed probability distribution P formed by k different Gaussian distributions.

Each instance xn generated by

1 Choosing Gaussian k according to prior probabilities [π1,...,πK]

2 Generating an instance at random according to that Gaussian, thus using μk, Σk

Introduce new variables $z_k \in \{0, 1\}$, with $\mathbf{z} = (z_1, \ldots, z_K)^T$ using a 1-out-of-K encoding (only one component is 1, all the others are 0).

Let's define

$$P(z_k = 1) = \pi_k$$

thus

$$P(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

For a given value of $\mathbf{z}$:

$$P(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Thus

$$P(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Joint distribution: $P(\mathbf{x}, \mathbf{z}) = P(\mathbf{x}|\mathbf{z})P(\mathbf{z})$ (chain rule).

When $\mathbf{z}$ are variables with 1-out-of-K encoding and $P(z_k = 1) = \pi_k$

$$P(\mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{z})P(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

GMM distribution $P(\mathbf{x})$ can be seen as the marginalization of a distribution $P(\mathbf{x}, \mathbf{z})$ over variables $\mathbf{z}$.

z_{nk}= 1 denotes x_n sampled from Gaussian k

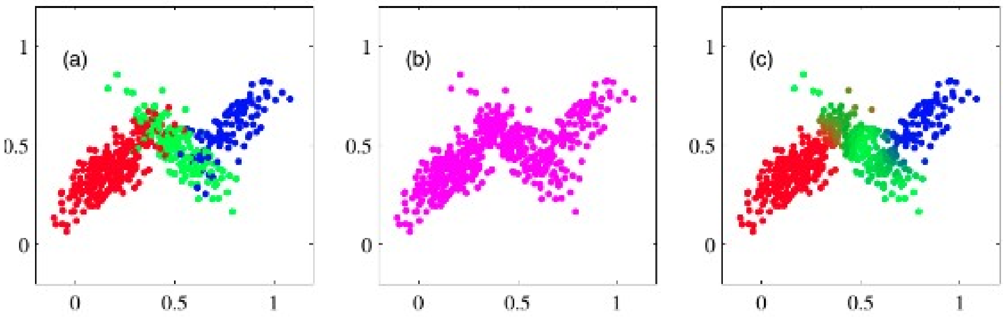
z_n are called **latent variables.**

Let's define the posterior

$$\gamma(z_k) \equiv P(z_k = 1|\mathbf{x}) = \frac{P(z_k = 1)\, P(\mathbf{x}|z_k = 1)}{P(\mathbf{x})}$$

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Note:
π_k: prior probability of z_k
$\gamma(z_k)$: posterior probability after observation of $\mathbf{x}$.



a) $P(\mathbf{x}, \mathbf{z})$ with 3 latent variables $\mathbf{z}$ (red, green, blue)
b) $P(\mathbf{x})$ marginalized distribution
c) $\gamma(z_{n,k})$ posterior distribution

14.3 Expectation Maximization (EM)

Note: generalization of K-means algorithm

Maximum likelihood

$$\underset{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmax}} \ln P(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

At maximum:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

$$\pi_k = \frac{N_k}{N}, \quad \text{with} \;\; N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

- Initialize $\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}$
- Repeat until termination condition $t = 0, \ldots, T$
  - **E step**

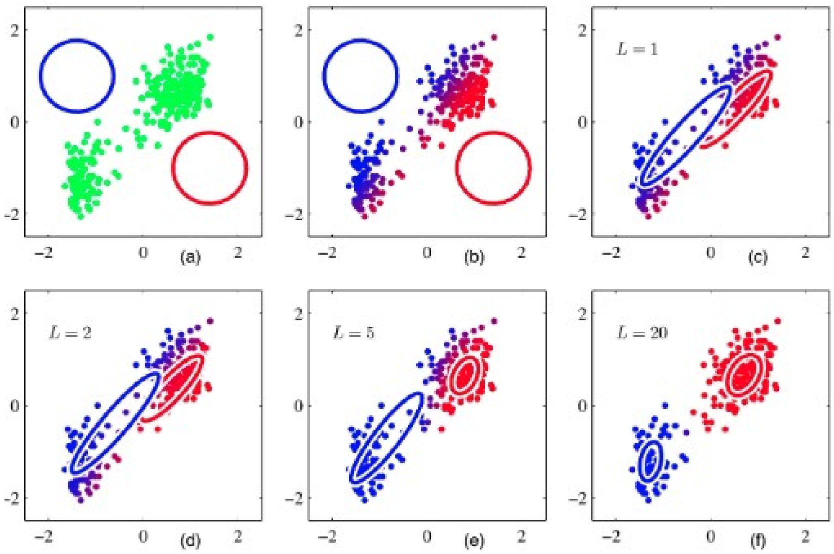
$$\gamma(z_{nk})^{(t+1)} = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}(\mathbf{x}_k; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}$$

  - **M step**

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})^{(t+1)} \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})^{(t+1)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})^T$$

$$\pi_k^{(t+1)} = \frac{N_k}{N}, \quad \textit{with } N_k = \sum_{n=1}^{N} \gamma(z_{nk})^{(t+1)}$$



Converges to local maximum likelihood; Provides estimates of the latent variables variables z_{nk} ;Not only gaussian

Define likelihood function $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ defined on variables $\mathbf{Y} = \mathbf{X} \cup \mathbf{Z}$, using observed $\mathbf{X}$ and current parameters $\boldsymbol{\theta}$ to estimate $\mathbf{Z}$

EM Algorithm:

Estimation (E) step: Calculate $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ using current hypothesis $\boldsymbol{\theta}$ and observed data $\mathbf{X}$ to estimate probability distribution over $\mathbf{Y}$

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \leftarrow E[\ln P(\mathbf{Y}|\boldsymbol{\theta}')|\boldsymbol{\theta}, \mathbf{X}]$$

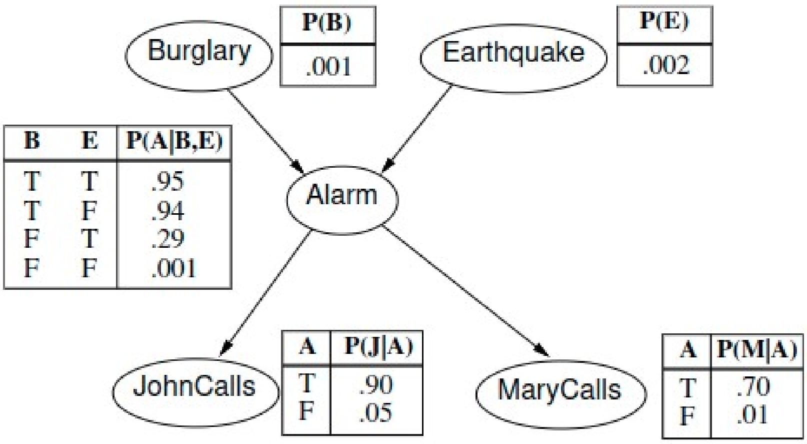
Maximization (M) step: Replace hypothesis $\boldsymbol{\theta}$ by the hypothesis $\boldsymbol{\theta}'$ that maximizes this Q function

$$\boldsymbol{\theta} \leftarrow \underset{\boldsymbol{\theta}'}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$$

14.4 Bayesian Network

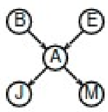
Examples of cavity, tooth etc…

Example:



All joint probabilities computed with the chain rule:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{Parents}(X_i))$$



e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

A CPT for Boolean variable X_i with k Boolean parents has 2^k rows for the combinations of parent values.

If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers.

When structure known and all variables observable conditional probabilities can be estimated with maximum likelihood.

Unsupervised learning can be seen as learning with a BN with one hidden variable (the class of the instances). This can be generalized to general BN with multiple hidden variables.

Example:

Consider three random variables $X \in \{0,1\}$, $A \in \{a_1, a_2\}$, $B \in \{b_1, b_2\}$, with X unobservable.

How can we learn BN parameters for $P(X)$, $P(A|X)$, and $P(B|X)$ from instances $D = \{d_1, ..., d_n\}$, with $d_k = \langle a_k, b_k \rangle$?

Define:
$P(X=0) = \theta_0$, $P(A=a_1|X=0) = \theta_1$, $P(A=a_1|X=1) = \theta_2$,
$P(B=b_1|X=0) = \theta_3$, $P(B=b_1|X=1) = \theta_4$

$\boldsymbol{\theta} = \langle \theta_0, \theta_1, \theta_2, \theta_3, \theta_4 \rangle$

Apply EM method to find maximum likelihood wrt $\boldsymbol{\theta}$ from D.

Estimation of BN parameters:

$$\begin{aligned} P(X=x_j) &= \frac{1}{n} E[\hat{N}(X=x_j)] \\ P(A=a_i|X=x_j) &= \frac{E[\hat{N}(A=a_i, X=x_j)]}{E[\hat{N}(X=x_j)]} \end{aligned}$$

Note that

$$E[\hat{N}(\cdot)] = E[\sum_k I(\cdot|d_k)] = \sum_k P(\cdot|d_k)$$

Estimation of BN parameters:

$$\begin{aligned} P(X=x_j) &= \frac{1}{n} \sum_{k=1}^{n} P(X=x_j|d_k) \\ P(A=a_i|X=x_j) &= \frac{\sum_{k=1}^{n} P(A=a_i, X=x_j|d_k)}{\sum_{k=1}^{n} P(X=x_j|d_k)} \\ P(B=b_l|X=x_j) &= \ldots \end{aligned}$$

Apply Bayes rule

$$\begin{aligned} P(x_j|d_k) &= P(x_j|\langle a_k, b_k \rangle) = \frac{P(a_k|x_j)P(b_k|x_j)}{\sum_i P(a_k|x_i)P(b_k|x_i)P(x_i)} = \phi_1(\boldsymbol{\theta}) \\ P(a_i, x_j|d_k) &= P(a_i|x_j, d_k)P(x_j|d_k) = \phi_2(\boldsymbol{\theta}) \\ P(b_l, x_j|d_k) &= P(b_l|x_j, d_k)P(x_j|d_k) = \phi_3(\boldsymbol{\theta}) \end{aligned}$$

... to define $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$

Unsupervised learning useful to deal with unknown variables

EM algorithm is a general method to estimate likelihood for mixed distributions

Concepts to be extended to continuous latent variables