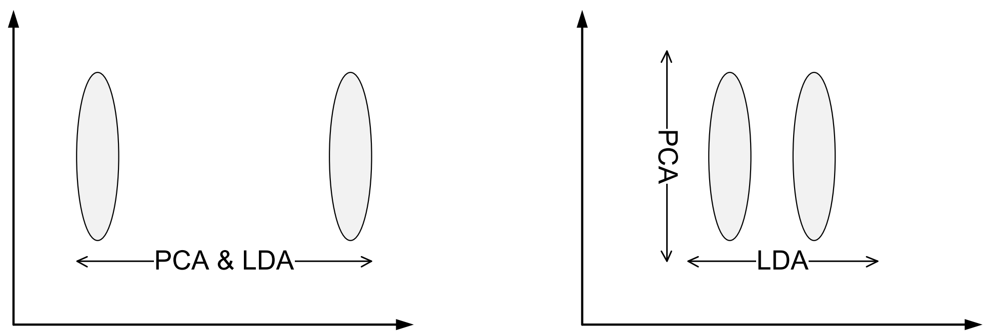
Solution of Assignment 3

November 24, 2015

1. Suppose $d = 2$. Draw data of two Gaussian classes such that a) PCA and LDA find the same direction; b) PCA and LDA find orthogonal directions.

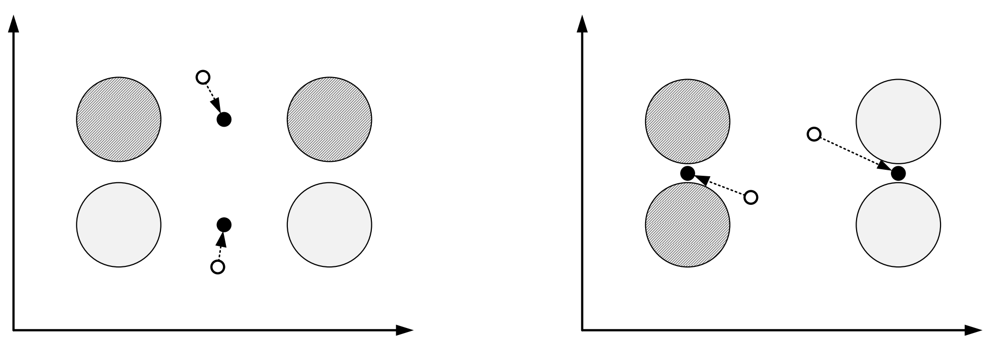
   Answer:



(a) (b)

2. Suppose $d = 2$ and $K = 2$. a) Draw data and locations of initial prototypes such that the groups found by the K-means algorithm is obviously wrong; b) Given the same data, draw locations of initial prototypes that allow the K-means to find the correct answer.

   Answer:



(a) (b)

3. In hierarchical clustering, we may decide to ignore the trees (or branches) that have a small number of descendents after cutting the dendrogram at a certain level, and report only those with sufficient descendents as groups. Why should we do this?

Answer:

We do this because a tree/branch in the dendrogram having a small number of descendents means that the corresponding group will be small and far away from other (big) groups. This implies that the group may contain outliers and is ignorable.

4. Given a set $\mathcal{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^N$ of i.i.d. instances. Suppose the attributes $x_i^{(t)}$, $1 \le i \le d$, of each instance are binary and independent with each other. Describe an EM algorithm that finds K clusters based on the multivariate Bernoulli mixture model where $P[\boldsymbol{x}^{(t)}|z_i^{(t)}, \theta_i] = \prod_{j=1}^d \rho_{i,j}^{x_j^{(t)}} (1-\rho_{i,j})^{1-x_j^{(t)}}$ and $\theta_i = (\rho_{i,1}, \cdots, \rho_{i,d})$ for $i = 1, \cdots, K$.

Answer:
Given mixtures of i.i.d. samples, we have $\mathcal{Q}(\Theta; \Theta^{old})$
$= \sum_{t=1}^N \sum_{i=1}^K \ln(\pi_i)\, P[z_i^{(t)}|\boldsymbol{x}^{(t)}, \Theta^{old}]$
$+ \sum_{t=1}^N \sum_{i=1}^K \ln\left(P[\boldsymbol{x}^{(t)}|z_i^{(t)}, \theta_i]\right) P[z_i^{(t)}|\boldsymbol{x}^{(t)}, \Theta^{old}]$.

E-step: We need to evaluate $h_i^{(t)} = P[z_i^{(t)}|\boldsymbol{x}^{(t)}, \Theta^{old}]$
$= \frac{P[\boldsymbol{x}^{(t)}|z_i^{(t)}, \theta_i^{old}]\pi_i^{old}}{\sum_{j=1}^K P[\boldsymbol{x}^{(t)}|z_j^{(t)}, \theta_j^{old}]\pi_j^{old}}$
$= \frac{\left(\prod_{k=1}^d (\rho_{i,k}^{old})^{x_k^{(t)}} (1-\rho_{i,k}^{old})^{1-x_k^{(t)}}\right)\pi_i^{old}}{\sum_{j=1}^K \left(\prod_{k=1}^d (\rho_{j,k}^{old})^{x_k^{(t)}} (1-\rho_{j,k}^{old})^{1-x_k^{(t)}}\right)\pi_j^{old}}$ for all $1 \le i \le K$ and $1 \le t \le T$.

M-step: The optimization problem for $\{\pi_i\}_{i=1}^K$ is the same as the case of Gaussian mixtures. The objective for $\{\theta_i\}_{i=1}^K$, on the other hand, now becomes $\arg_{\theta_1,\cdots,\theta_K} \max Q$, where $Q = \sum_{t=1}^N \sum_{i=1}^K \ln\left(P[\boldsymbol{x}^{(t)}|z_i^{(t)}, \theta_i]\right) h_i^{(t)}$
$= \sum_{t=1}^N \sum_{i=1}^K \sum_{j=1}^d \left(x_j^{(t)} \ln \rho_{i,j} + (1 - x_j^{(t)}) \ln(1-\rho_{i,j})\right) h_i^{(t)}$. Taking partial derivatives of Q with respect to $\rho_{i,j}$ and setting it to zero we have
$\sum_{t=1}^N \left(x_j^{(t)} \frac{1}{\rho_{i,j}} - (1 - x_j^{(t)}) \frac{1}{1-\rho_{i,j}}\right) h_i^{(t)} = 0 \Rightarrow \rho_{i,j} = \frac{\sum_{t=1}^N x_j^{(t)} h_i^{(t)}}{\sum_{t=1}^N h_i^{(t)}}$.