Assignment 3

Due: 2015/11/24 23:59:59

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.

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.

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?

4. Given a set $\mathcal{X} = \{\boldsymbol{x}^{(t)}\}_{t=1}^{N}$ of i.i.d. instances. Suppose the attributes $x_i^{(t)}$, $1 \leq i \leq d$, 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 $i = 1, \cdots, K$.