1 GMM and K-means

K-means is a simple iterative clustering algorithm which assigns $n$ data points to $k$ clusters. First $k$ random cluster centroids are chosen. Then in each iteration

- each of the $n$ data points is assigned to the nearest cluster
- each of the $k$ cluster centroids is updated to be the mean of all the points assigned to this cluster.

The algorithm terminates when cluster centroids stop changing.

Consider a mixture of Gaussians model defined by $K$ means $\mu_1, \ldots, \mu_K$, variance $\sigma^2$, and proportions (mixture coefficients) $\pi = \pi_1, \ldots, \pi_K$. In such a model, each (real-valued) $X_n$ is generated as follows: First, one of the mixture components $Z_n \in \{1, \ldots, K\}$ is chosen at random according to $\pi$ (so that $Z_n = z$ with probability $\pi_z$). Then, given that $Z_n = z$, $X_n$ is chosen according to a Gaussian distribution with mean $\mu_z$ and variance $\sigma^2$. Note that only $X_n$ is visible; $Z_n$ is hidden. We assume that $\sigma > 0$ is known and fixed. Given data $X_1 : N$, recall the EM algorithm for estimating $\mu_1, \ldots, \mu_K$ and $\pi$.

By answering these questions show how is the K-means algorithm related to the EM algorithm for Gaussian Mixtures.

a. Argue, from the algorithmic perspective, that as $\sigma^2 \to 0$, this algorithm approaches the K-means algorithm.

b. Argue now that the EM objective approaches the K-means objective.

2 Clustering using GMM

In this part of the exercise we will use the EM algorithm for GM to cluster the iris data: [iris.txt]. Each Gaussian will have two dimensions corresponding to the petal width and petal length. The class labels will only be used for evaluation.

Initialization Set the number of Gaussians in the mixture to $k = 2$. Set the mixture components to be $\frac{1}{k}$. Initialize the the covariance matrices to identity matrices. For each Gaussian choose 5 random points from the data and set the initial mean to be the mean of those points.

Model evaluation Run the algorithm until convergence and plot the model likelihood as it changes. Once the algorithm terminates, evaluate the cluster assignments against the class labels (versicolor or virginica). For this purpose map cluster ids to class labels such that the classification error is minimized.

Report classification errors for 10 different random initializations of the algorithm, as explained above.
Please send your solutions by Tuesday Jan 15 to psr-tutorial@lsv.uni-saarland.de

Important: Please use PDF as a document format. If you need to compress files, use ZIP or GZIP.