Assignment

SVD, Least Squares, K-means. Study the materials presented in class on Singular Value Decomposition and its applications as Principal Component Analysis and the solution of Least Squares problems, as well as K-means.

Exercises

Exercise 1. Let $K = 2$ be the number of clusters, and let $D$ denote the ambient dimension of $\mathbb{R}^D$. Consider data generated in the following way: with probability $\frac{1}{2}$ we sample from $G_1$, a Gaussian distribution $\mathcal{N}(0, \Sigma)$, and with probability $\frac{1}{2}$ we sample from $G_2$, a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.

- Let $\Sigma = I_D$ (the identity matrix in $D$ dimensions), $\mu = (m, 0, \ldots, 0) \in \mathbb{R}^D$, and generate $n$ points $x_1, \ldots, x_n$ according to the recipe above, keeping track of which point was sampled from $G_1$ and which point was sampled from $G_2$ by storing a label $l_i \in \{1, 2\}$ for each of the $x_i$’s. Run K-means, to obtain labels $\tilde{l}_i \in \{\text{A'}, \text{B'}\}$. Note that the labels returned by K-means know nothing about your original labels (K-means is an unsupervised algorithm, and did not ask for the $l_i$’s as inputs), which is why I called them ’A’ and ’B’. Match the labels ’A’ and ’B’ to the labels 1,2 by assigning ’A’ to the numeric label which has most points labeled as ’A’ by K-means (and similarly for ’B’). After this label-matching, you can compute the error rate $E_{n,m,D}$ of K-means as the number of correctly labelled points ($\# \{i : \tilde{l}_i \neq l_i\}$) divided by $n$. Note that this number $E_{n,m,D}$ is random with the data. Study and plot its mean and standard deviation (computed by doing multiple runs of the above) as a function of $n$, and note it converges to some number $E_{m,D}$ as $n$ goes to infinity.

- Plot the (estimated) $E_{m,D}$ as a function of $m$ and $D$ (you could do multiple Cartesian plots for $m$ fixed and varying $D$, but also a level-set or surface plot in 3-D, with the two independent variables being $m$ and $D$). Try to determine as accurately as you can what the minimal values of $m$ and $D$ (and how they relate to each) in order for $E_{m,D}$ to be below a certain level (say, 5%), and discuss your findings.