Machine Learning Worksheet 11

Tomas Ladek, Michael Kratzer
3602673, 3612903
tom.ladek@tum.de, mkratzer@mytum.de

Problem 1

Consider a Gaussian Mixture Model that describes the data points $\boldsymbol{x}$

$$p(x|\theta) = \sum_{k=0}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma^2 \boldsymbol{I})$$

with some $\sigma \in \mathbb{R}$. Let us define $\pi_k = \frac{|C_k|}{N}$ with $|C_k|$ being the number of data point belonging to a cluster k and N the total number of data points.

Since we cannot easily optimize

$$\arg\max_{\theta} \ln p(\mathcal{D}|\theta) = \sum_{n=1}^{N} \ln \sum_{k=0}^{K} \frac{|C_k|}{N} \mathcal{N}(x_n|\mu_k, \sigma^2 \boldsymbol{I})$$

we use the EM algorithm, which consists of

1. Fixing the parameters of some posterior distribution and calculating responsibilities for a data point x.
2. Fixing responsibilities for a data point x and optimizing the parameters of the underlying distribution.

Because the Euclidean metric puts equidistant points on a sphere (in 3D; on a circle in 2D, etc.), one can easily see how this metric can be implemented by employing a spherical Gaussian, i.e. one with a diagonal covariance matrix. The first step in the EM algorithm now is choosing some Gaussian means μ_k and calculating responsibilities by evaluating the Gaussian at $||x_n - \mu_k||_2$ for every data point x_n - corresponds to the step "Calculate clusters". The second step of finding optimal parameters for the Gaussians then corresponds to "recalibrating the cluster means" with $\mu_k \leftarrow \frac{1}{|C_k|} \sum_{x \in C_k} x$.

More formally: