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 $\pi_k = \dfrac{|C_k|}{N}$ 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. In EM, we are doing alternate coordinate ascent on $\mathcal{L}_{ELBO}(q, \theta)$ for some posterior distribution $q(z)$, i.e. two steps:

1. Optimization w.r.t. q:

$$\arg\max_q \mathcal{L}_{ELBO}(q, \theta) = \arg\max_q \mathbb{E}_{q(z)}[\ln p(x, z|\theta)] + \mathrm{H}(q)$$

   With

$$\mathbb{E}_{q(z)}[\ln p(x, z|\theta)] = \mathbb{E}_{q(z)}[\ln p(z|x, \theta)] + \mathbb{E}_{q(z)}[\ln p(x|\theta)]$$

   one arrives at

$$\arg\max_q \mathcal{L}_{ELBO}(q, \theta) = \arg\max_q \left(\mathbb{E}_{q(z)}[\ln p(z|x, \theta)] + \mathbb{E}_{q(z)}[\ln p(x|\theta)] + \mathrm{H}(q)\right)$$

2. Optimization w.r.t. parameters θ:

$$\begin{aligned}\arg\max_\theta \mathcal{L}_{ELBO}(q, \theta) &= \arg\max_\theta \left(\mathbb{E}_{q(z)}[\ln p(x, z|\theta)] + \mathrm{H}(q)\right) \\ &= \arg\max_\theta \left(\mathbb{E}_{q(z)}[\ln p(x, z|\theta)]\right) \\ &= \arg\max_\theta \left(\mathbb{E}_{q(z)}[\ln p(z|x, \theta)] + \mathbb{E}_{q(z)}[\ln p(x|\theta)]\right)\end{aligned}$$

The *K-Means algorithm* is an instance of the EM-algorithm. In the first step (E-Step), using

$$p(z|x,\theta) = r_{nk}(\theta) \sim \mathcal{N}(d(x_n,\mu_k)|\mu_k, \sigma^2 \boldsymbol{I}) \qquad \text{with} \quad d(x_n,\mu_k) = ||x_n - \mu_k||_2$$

and taking σ towards 0, one can see that $\mathcal{L}_{ELBO}(q,\theta)$ is maximized by picking a j for every x_n such that $d(x_n,\mu_j)$ is minimized, which corresponds to the step of "calculating clusters" in *K-Means*. [Euclidean distance puts equidistant points on a circle (2D) resp. sphere (3D), so does a Gaussian with $\Sigma = \sigma^2 \boldsymbol{I}$]

In the second step (M-step) the optimization of the arguments θ is done, namely the μ_k are updated as follows:

$$\mu_k \leftarrow \frac{1}{|C_k|} \sum_{x \in C_k} x$$

and thus maximizing $\mathcal{L}_{ELBO}(q,\theta)$, which corresponds to the step "recalibrating the cluster mean".

Problem 2

?

Problem 3

The KL divergence for two Gaussian distributions of dimension k is defined as follows:

$$KL(\mathcal{N}_1||\mathcal{N}_2) = \frac{1}{2}\Big(\text{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_1^{-1}(\mu_2 - \mu_1) - k + \ln\big(\frac{\det \Sigma_2}{\det \Sigma_1}\big)\Big)$$

With Σ_1 and Σ_2 being diagonal, this amounts to

$$KL(\mathcal{N}_1||\mathcal{N}_2) = \frac{1}{2}\Big(\sum_k \frac{\sigma_{1_k}}{\sigma_{2_k}} + (\mu_2 - \mu_1)^T \Sigma_1^{-1}(\mu_2 - \mu_1) - k + \ln\big(\prod_k \sigma_{2_k}\big) - \ln\big(\prod_k \sigma_{1_k}\big)\Big)$$