Machine Learning Worksheet 11

Inference in latent variable models

1 Expectation Maximisation

The exercises on EM might be easier after the tutorial.

Problem 1: The *K-Means algorithm* is an intuitive clustering algorithm. Initialize K cluster centres μ_k so that they are distinct. Then iterate the following procedure until convergence:

1. Calculate clusters

$$C_k = \left\{ \mathbf{x}_i \in \mathcal{D} : k = \arg\min_j d(\mathbf{x}_i, \boldsymbol{\mu}_j) \right\},$$

   i.e., C_k gathers all data points for which $\boldsymbol{\mu}_k$ is the closest centre in some metric $d(\cdot, \cdot)$.

2. Recalibrate the cluster mean:

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{|C_k|} \sum_{\mathbf{x} \in C_k} \mathbf{x}$$

Show that this algorithm is an instance of the EM-algorithm for the Euclidean metric $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_2$.

To do so, consider a special mixture of isotropic Gaussians where every mixture component has identical *fixed* covariance $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}, \sigma \in \mathbb{R}$, determine the EM algorithm for this, and then take $\sigma \to 0$.

Problem 2: You have two equal-looking coins, one of which is fair and the other one is biased, but you cannot tell them apart. You pick a coin to start at random, and then flip the two coins alternately (in an ABAB... pattern). After a total of ten flips, you obtain a sequence 0100110111, where 1 corresponds to heads and 0 to tails. That is, out of the $N_e = 5$ even flips, $K_e = 4$ showed up heads, and out of the $N_o = 5$ odd flips, $K_o = 2$ showed up heads.

You want to do maximum likelihood optimization in θ, the probability that the biased coin shows up heads. Derive a general EM-algorithm to determine θ in terms of N_o, N_e, K_o, K_e. In particular, do not assume $N_o = N_e$.

Initialize $\theta^{(0)} = 0.5$. Implement your algorithm in a framework of your choice and report the converged result with the given data.

Hint 1: $\theta^{(1)} = 0.6$.

Hint 2: Reduce your workload by reusing computations from previous coin examples in this course.

2 Variational Auto-Encoders

Problem 3: Compute the KL divergence between two isotropic Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, i.e., Gaussians with diagonal covariance.

Hint: If you exploit the facts you know, you can save yourself a lot of work before walking down the straightforward path.

Problem 4: What do you expect to happen when leaving away the KL term in the evidence lower bound when training a variational auto-encoder?

If that helps you, assume isotropic prior and approximate posterior for both scenarios and use your insights from the previous exercise.