Machine Learning Worksheet 9

Latent Variable Models

1 K-Means and MoG

Problem 1. Consider a mixture of K isotropic Gaussians, each with the same covariance $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. In the limit $\sigma^2 \to 0$ show that the EM algorithm for MoG converges to the K-Means algorithm.

Note that the only difference between the two algorithms is in the E Step!

In the general setting of the MoG model, we have for some data point $\boldsymbol{x}_i$ in the E Step:

$$p(\boldsymbol{z}_i = k|\boldsymbol{x}_i) = \frac{\pi_k \exp\left(\frac{-||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2}{2\sigma^2}\right)}{\sum_l \pi_l \exp\left(\frac{-||\boldsymbol{x}_i - \boldsymbol{\mu}_l||^2}{2\sigma^2}\right)} = \frac{1}{\sum_l \frac{\pi_l}{\pi_k} \exp\left(\frac{-||\boldsymbol{x}_i - \boldsymbol{\mu}_l||^2 + ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2}{2\sigma^2}\right)}$$

If k denotes the component that is closest to $\boldsymbol{x}_i$, then $||\boldsymbol{x}_i - \boldsymbol{\mu}_l||^2 \geq ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2$ for all l, then $-||\boldsymbol{x}_i - \boldsymbol{\mu}_l||^2 + ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \leq 0$ for all l and thus the denominator converges to 1 if $\sigma^2 \to 0$ (because if $l = k$, this part of the sum in the denomiator is always 1, and all other summands converge to 0 because $\exp(-\infty)$ does so).

On the other hand, if k is not resembling the closest component, then $-||\boldsymbol{x}_i - \boldsymbol{\mu}_l||^2 + ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 > 0$ for l denoting the closest component, and whith $\sigma^2 \to 0$ the exponent of this component is

$$\frac{-||\boldsymbol{x}_i - \boldsymbol{\mu}_l||^2 + ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2}{2\sigma^2} \to +\infty$$

and thus the denominator converges to ∞. In total, this results in the hard assignment step of K-Means.

Problem 2. Consider a mixture of K Gaussians

$$p(\boldsymbol{x}) = \sum_k \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Derive $E(\boldsymbol{x})$ and $Cov(\boldsymbol{x})$. It is helpful to remember the identity $Cov(\boldsymbol{x}) = E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})E(\boldsymbol{x})^T$.

For $E(\boldsymbol{x})$ we use a tower formula (see Exercise sheet 2, Problem 2):

$$E(\boldsymbol{x}) = E(E(\boldsymbol{x}|\boldsymbol{z})) = \sum_k \pi_k E(\boldsymbol{x}|\boldsymbol{z})) = \sum_k \pi_k \boldsymbol{\mu}_k$$

Using the identity for the covariance, we first compute $E(\boldsymbol{x}\boldsymbol{x}^T)$, again using the above tower formula:

$$E(\boldsymbol{x}\boldsymbol{x}^T) = \sum_k \pi_k E(\boldsymbol{x}\boldsymbol{x}^T|\boldsymbol{z})$$

Reusing (in the other direction) the identity, we have

$$E(\boldsymbol{x}\boldsymbol{x}^T|\boldsymbol{z}) = \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$$

and thus

$$Cov(\boldsymbol{x}) = \sum_k \pi_k(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T) - E(\boldsymbol{x})E(\boldsymbol{x})^T$$

2 FA/pPCA and PCA

Problem 3. Consider the latent space distribution

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$$

and a conditional distribution for the observed variable $\boldsymbol{x} \in \mathbb{R}^d$,

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Phi})$$

where $\boldsymbol{\Phi}$ is an arbitrary symmetric, positive-definite noise covariance variable. Furthermore, $\boldsymbol{A}$ is a non-singular $d \times d$ matrix and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. Show that for the maximum likelihood solution for the parameters of the model for $\boldsymbol{y}$ specific constraints on $\boldsymbol{\Phi}$ are preserved in the following two cases: (i) $\boldsymbol{A}$ is a diagonal matrix and $\boldsymbol{\Phi}$ is a diagonal matrix (this corresponds to the case of Factor Analysis). (ii) $\boldsymbol{A}$ is orthogonal and $\boldsymbol{\Phi} = \sigma^2\boldsymbol{I}$ (this corresponds to pPCA).

The model for $\boldsymbol{y}$ is a *noiseless* linear transformation. Given that the distribution of $\boldsymbol{x}$ is known, we therefore know the distribution of $\boldsymbol{y}$. Because of the definitions for $\boldsymbol{z}$ and $\boldsymbol{x}|\boldsymbol{z}$ we know that $\boldsymbol{x}$ is a Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Phi}$. And thus, $\boldsymbol{y}$ is also Gaussian with mean $\boldsymbol{A}\boldsymbol{\mu}$ and covariance $\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{A}^T$. Now, assuming that the maximum likelihood solutions for the conditional model for $\boldsymbol{x}$ are $\boldsymbol{\mu}_x$, $\boldsymbol{W}_x$ and $\boldsymbol{\Phi}_x$, by simple *matching patterns* the MLE solutions for $\boldsymbol{y}$ are $\boldsymbol{A}\boldsymbol{\mu}_x$, $\boldsymbol{A}\boldsymbol{W}_x$ and $\boldsymbol{A}\boldsymbol{\Phi}_x\boldsymbol{A}^T$. (i) If $\boldsymbol{A}$ and $\boldsymbol{\Phi}$ are diagonal matrices (Factor Analysis model), the characteristics of $\boldsymbol{x}$ are preserved for $\boldsymbol{y}$. Similarily (ii) if $\boldsymbol{A}$ is orthogonal and $\boldsymbol{\Phi}$ a scaled indentity matrix, the model characteristics are also preserved ($\boldsymbol{A}\boldsymbol{\Phi}_x\boldsymbol{A}^T = \sigma^2\boldsymbol{I}$ in this case).

Problem 4. Show that in the limit $\sigma^2 \to 0$ the posterior mean for the probabilistic PCA model becomes an orthogonal projection onto the same principal subspace as in PCA.

Remember, pPCA is a Factor Analysis model with $\boldsymbol{\Psi} = \sigma^2\boldsymbol{I}$ and $\boldsymbol{W}$ orthonormal. First, we plug the special form of $\boldsymbol{\Psi}$ into the general result for the posterior mean of the latent variable $\boldsymbol{z}$, which is given in the slides:

$$\boldsymbol{m}_i = \boldsymbol{\Sigma}(\boldsymbol{W}^T\sigma^{-2}\boldsymbol{I}(\boldsymbol{x}_i - \boldsymbol{\mu}))$$

with

$$\boldsymbol{\Sigma} = (\boldsymbol{I} + \boldsymbol{W}^T\sigma^{-2}\boldsymbol{I}\boldsymbol{W})^{-1} = \sigma^2(\sigma^2\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{W})^{-1}$$

which gives

$$\boldsymbol{m}_i = (\sigma^2 \boldsymbol{I} + \boldsymbol{W}^T \boldsymbol{W})^{-1} (\boldsymbol{W}^T (\boldsymbol{x}_i - \boldsymbol{\mu}))$$

With $\sigma^2 \to 0$ the maximum likelihood solution for $\boldsymbol{W}$ (given in slides) converges to $\boldsymbol{V}_l \boldsymbol{\Lambda}_l^{1/2}$. So $(\sigma^2 \boldsymbol{I} + \boldsymbol{W}^T \boldsymbol{W})^{-1} \to \boldsymbol{\Lambda}_l^{-1}$, and thus

$$\boldsymbol{m}_i = \boldsymbol{\Lambda}_l^{-1/2} \boldsymbol{V}_l^T (\boldsymbol{x}_i - \boldsymbol{\mu})$$

which is a projection on the same subspace as PCA does.