

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 $\Sigma = \sigma^2 \mathbf{I}$. In the limit $\sigma^2 \rightarrow 0$ 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 \mathbf{x}_i in the E Step:

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

If k denotes the component that is closest to \mathbf{x}_i , then $\|\mathbf{x}_i - \boldsymbol{\mu}_l\|^2 \geq \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$ for all l , then $-\|\mathbf{x}_i - \boldsymbol{\mu}_l\|^2 + \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \leq 0$ for all l and thus the denominator converges to 1 if $\sigma^2 \rightarrow 0$ (because if $l = k$, this part of the sum in the denominator 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 $-\|\mathbf{x}_i - \boldsymbol{\mu}_l\|^2 + \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 > 0$ for l denoting the closest component, and with $\sigma^2 \rightarrow 0$ the exponent of this component is

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

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

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

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

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

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

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

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

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

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

and thus

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

2 FA/pPCA and PCA

Problem 3. Consider the latent space distribution

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

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

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

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

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

Problem 4. Show that in the limit $\sigma^2 \rightarrow 0$ 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 $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$ and \mathbf{W} orthonormal. First, we plug the special form of $\boldsymbol{\Psi}$ into the general result for the posterior mean of the latent variable z , which is given in the slides:

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

with

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

which gives

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

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

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

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