Machine Learning Worksheet 10

Unsupervised Learning

Problem 1: Where would you use PCA and where ICA on real world data?

Problem 2: 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 3: 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.

You may use the solution for the posterior of the FA model:

$$p(\boldsymbol{z}_i|\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{z}_i|\boldsymbol{m}_i, \boldsymbol{\Sigma})$$

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

$$\boldsymbol{m}_i = \boldsymbol{\Sigma}(\boldsymbol{W}^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}))$$

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 $\boldsymbol{z}$:

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

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.