CS5340: Uncertainty Modeling in AI

Tutorial 6

Released: Mar. 6 2024

Problem 1. (Probabilistic PCA)

Principal Components Analysis (PCA) is a famous model applied to dimensionality reduction or visualization where we map data samples to a lower dimensional space[1]. We will now leverage the linear-Gaussian framework to derive a probabilistic form of PCA. *Note:* for this problem, we will be denoting random variables with lower case letters, and bolded lowercase letters to represent vectors, and bolded uppercase letters to represent matrices.

For the probabilistic PCA model, we have D-dimensional data points $\mathbf{x}_i$ for $i = 1, 2, \ldots, N$ and we aim to find some reduced structure for the data. For each data point, we associate a M-dimensional latent variable (where often $M < D$) $\mathbf{z}_i$ that has prior distribution,

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

We define each observed variable $\mathbf{x}$ as,

$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. We can imagine that each data point is obtained by first sampling from the prior $p(\mathbf{z}_i)$ followed by an affine transformation and additive Gaussian noise.

Problem 1.a. Draw the DGM corresponding to the model above. *Hint:* use plate notation for the different data points.

Problem 1.b. Show that the conditional distribution for each observed variable $\mathbf{x}_i$ is given by:

$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})$$

[1]For more information about PCA, you can see various sources including `https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202`.

Problem 1.c. To find the MLE values for the model parameters $\mathbf{W}, \boldsymbol{\mu}$, and σ^2, we would need the marginal distribution $p(\mathbf{X}) = \prod_i^N p(\mathbf{x}_i)$ (assuming i.i.d. data). Due to the latent variables, we will use the EM algorithm[2]. This requires us to marginalize out the latent $\mathbf{z}$'s. To help us along,

1. First, show that the marginal distribution of each data point is again a Gaussian given by

   $$p(\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$$

   where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

2. Then, show that the posterior distribution is also normally distributed,

   $$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$$

   where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

Hint: Given random variables $\mathbf{x}$ and variable $\mathbf{y}$ where:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x) \tag{1}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x}) \tag{2}$$

The marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T) \tag{3}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right) \tag{4}$$

where

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}$$

Problem 1.d. Finally, derive the E-step and the M-step for the EM algorithm applied to probabilistic PCA. *Hint:* If you are really stuck, refer to Chapter 12.2.2. of Bishop's Pattern Recognition and Machine Learning. This portion is not especially difficult but is notationally heavy and requires algebraic manipulation.

[2]We can actually find the solution for fully-observed $\mathbf{X}$ via eigenvector decomposition. However, the EM algorithm can have computational advantages and can be extended to missing data scenarios. For more information, see Chapter 12 of Bishop's Pattern Recognition and Machine Learning.