Machine Learning

Homework 3

Not collected, not graded.

1 From probabilistic PCA to linear regression models

We have seen that pPCA builds a data model for $\mathbf{x} \in \mathbb{R}^D$ based on latent variable $\mathbf{z} \in \mathbb{R}^M$, for $M < D$:

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

where $\mathbf{W}$ is a $D \times M$ matrix, $\boldsymbol{\mu} \in \mathbb{R}^D$, and $\boldsymbol{\varepsilon}$ is Gaussian random noise, $p(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon} \mid 0, \sigma^2 I_D)$. On the other hand, a linear regression model describes the target value $t \in \mathbb{R}$ as a linear model (in $w_0, \mathbf{w}$) from the given input data $\mathbf{x} \in \mathbb{R}^D$, plus noise:

$$t = w_0 + \mathbf{w}^T \mathbf{x} + \boldsymbol{\varepsilon}$$

where w_0 is a bias term, $\mathbf{w} \in \mathbb{R}^D$ are weights and $\boldsymbol{\varepsilon}$ is again Gaussian random noise, $p(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon} \mid 0, \sigma^2 I_D)$.

1. For both models, write down the likelihood, $p(\mathbf{x} \mid \mathbf{z})$ and $p(t \mid \mathbf{x})$, respectively. Differences? Similarities?

2. What is the major difference between the models? *Hints:* try looking at the predictive distribution $p(\mathbf{x})$ and $p(t)$, respectively, by marginalizing. Why do we fail for $p(t)$? What is different between the latent variables $\mathbf{z}$ and the input variables $\mathbf{x}$?

3. Once we introduce basis functions $\phi_j(\mathbf{x})$, write down the linear regression model corresponding to the above fully linear regression. What PCA model does this most closely correspond to?

4. Compare the minimum-error formulation of $\mathbf{W}$ in PCA to the maximum likelihood estimate of $\mathbf{w}$ in linear regression (eigenvectors versus minimal least squares). What are the similarities and what is different? *Hint:* draw a point cloud and a linear "subspace" (linear manifold more precisely, because the origin is not necessarily contained in it), and visualize what errors are minimized in each case.

5. In class we have seen that the maximum likelihood estimate of the weights is found by minimizing the data error function:

   $$\frac{1}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\right)^2,$$

   where we had adopted the convention $\phi_0(\mathbf{x}) = 1$. Let's make the bias term w_0 more explicit by writing:

   $$\frac{1}{2}\sum_{n=1}^{N}\left(t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\mathbf{x}_n)\right)^2,$$

   instead. Determine the maximum likelihood estimate for w_0 from this, explicitly, and discuss it.

2 Bayesian linear regression

The data-likelihood of the linear regression model considered above, based on observed data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and target variables $\mathbf{t} = (t_1, \ldots, t_N)$, is given by

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1}),$$

where β $(= 1/\sigma^2)$ is the noise precision parameter (assumed known). Instead of just maximum likelihood estimation for $\mathbf{w}$, let's now assume a Gaussian prior distribution on the parameters $\mathbf{w} \in \mathbb{R}^M$ (including w_0, again):

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0)$$

for $\mathbf{m}_0$ and $\mathbf{S}_0$ given mean and covariance matrix. The resulting posterior distribution is Gaussian:

$$p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N).$$

1. Show that the mean and covariance of the posterior are given by

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^T\mathbf{t})$$

and

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\Phi^T\Phi,$$

where Φ is the data design matrix. *Hint:* see page 93 of PRML.