Tutoring Session 5

Linear Regression

1 Linear regression

Problem 1: Show that the matrix

$$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$$

takes any vector and projects it onto the space spanned by the columns of $\mathbf{\Phi}$. Use this result to show that the least square solution for linear regression corresponds to an orthogonal projection of the vector $\boldsymbol{T}$ (denoted by $\boldsymbol{Z}$ in class!) onto the manifold S as shown in Figure 1. There, the subspace S is spanned by the basis functions $\phi_j(\boldsymbol{x})$ in which each basis function is viewed as a vector $\boldsymbol{\varphi}_j$ of length N with elements $\phi_j(\boldsymbol{x}_n)$. (Hint: You might want consider what $\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ resembles, e.g. how does it relate to the maximum likelihood solution for linear regression.)

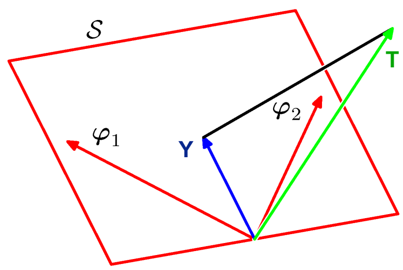

Figure 1: The projection property of $\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$.

2 Ridge regression

Problem 2: Using singular value decomposition of the design matrix $\mathbf{\Phi} = \boldsymbol{UDV}^T$ show that the output on the training set fitted with the ridge regression solution $\hat{\boldsymbol{w}}^{ridge}$ can be written as

$$\sum_j \left(\frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j \boldsymbol{u}_j^T \right) \boldsymbol{z}$$

where $\boldsymbol{u}_j$ are the columns of $\boldsymbol{U}$, d_j the elements of $\boldsymbol{D}$ and λ the cost factor of the $\ell 2$ regularization. What is the interpretation of this formula?

3 Multi-output linear regression

Problem 3: In class, we only considered functions of the form $f : \mathbb{R}^n \to \mathbb{R}$. What about the general case of $f : \mathbb{R}^n \to \mathbb{R}^m$? For linear regression with multiple outputs, write down the loglikelihood formulation and derive the MLE of the parameters.

4 Bayesian Linear Regression

Problem 4: We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases. Prove the following matrix identity

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}}$$

and, using it, show that the uncertainty $\sigma_N^2(\boldsymbol{x})$ associated with the bayesian linear regression function given by eq. (26) on the slides satisfies

$$\sigma_{N+1}^2(\boldsymbol{x}) \leq \sigma_N^2(\boldsymbol{x}) \tag{1}$$

You may want to use

$$\boldsymbol{\Phi}_{N+1}^T\boldsymbol{\Phi}_{N+1} = \boldsymbol{\Phi}_N^T\boldsymbol{\Phi}_N + \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T$$

Problem 5: We know that the posterior for a linear regression algorithm with a likelihood defined by $p(Z \mid W, \beta) = \prod_{n=1}^{N} \mathcal{N}(Z_n \mid W^T\Phi(X_n), \beta^{-1})$ and prior given by $p(w) = \mathcal{N}(W \mid M_0, \mathbf{S}_0)$ is

$$p(W \mid Z) = \mathcal{N}(W \mid M_N, \mathbf{S}_N)$$

where

$$M_N = \mathbf{S}_N(S_0^{-1}M_0 + \beta\Phi^T Z)$$
$$S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi$$

Let's assume β is a known constant. Verify that this is the form of the posterior we would derive.

5 Online Learning

Problem 6: Suppose we are using a linear basis function model where the posterior distribution is given by $p(W \mid Z) = \mathcal{N}(W \mid M_N, \mathbf{S}_N)$ and we have already observed N data points. That means that this posterior can be regarded as the prior for the next observation. By considering an additional data point (X_{N+1}, z_{N+1}), and by completing the square in the exponential, show that the resulting posterior distribution is again given by the posterior mentioned above, but with $\mathbf{S}_N$ replaced by $\mathbf{S}_{N+1}$ and M_N replaced by M_{N+1}.