Machine Learning: Exercises for Block II (Linear Regression)

Isabel Valera

Exercise 1: Orthogonal

i) Show that the matrix 1 takes any vector $\boldsymbol{v}$ and projects it onto the space spanned by the columns of $\boldsymbol{\Phi}$.

ii) Use this result to show that the least-squares solution given in equation 2 corresponds to an orthogonal projection of the vector $\boldsymbol{t}$ onto the manifold $\mathcal{S}$ as shown in figure 1.

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top} \tag{1}$$

$$\boldsymbol{w}_{ML} = \left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{t} \tag{2}$$

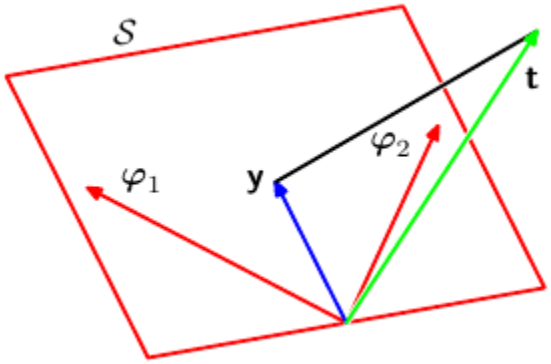

Figure 1

Exercise 2: Sample weights

Consider a data set in which each data point y_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares loss function is as in equation 3.

i) Find an expression for the solution $\boldsymbol{w}^*$ that minimizes this error function.

ii) Given an interpretation of the weighted sum-of-squares error function in terms of data dependent noise variance.

iii) Given an interpretation of the weighted sum-of-squares error function in terms of replicated data points.

$$L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left(y_n - \boldsymbol{w}^\top \varphi(\boldsymbol{x}_n)\right)^2 \tag{3}$$

Exercise 3: Independent noise

Consider the linear model given in equation 4 together with the sum-of-squares loss function given in 5. Now suppose that Gaussian noise ε_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}\left[\varepsilon_i\right] = 0$ and $\mathbb{E}\left[\varepsilon_i \varepsilon_j\right] = \delta_{ij}\sigma^2$ show that minimizing $L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w})$ averaged over the noise distribution is equivalent to minimizing the sum of squares loss for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

$$f(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{D} w_i x_i \tag{4}$$

$$L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(f(\boldsymbol{x}_n, \boldsymbol{w}) - y_n\right)^2 \tag{5}$$

Exercise 4: Linear basis functions

Consider a linear basis function regression model for a multivariate target variable $\boldsymbol{y}$ having a Gaussian distribution as given in 6 where $f(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}^\top \varphi(\boldsymbol{x})$ together with a training data set comprising input basis vectors $\varphi(\boldsymbol{x}_n)$ and corresponding target vectors $\boldsymbol{y}_n$, with $n = 1, ..., N$.

$$p(\boldsymbol{y} \mid \boldsymbol{W}, \boldsymbol{\Sigma}) = \text{Gauss}(\boldsymbol{y} \mid f(\boldsymbol{x}, \boldsymbol{W}), \boldsymbol{\Sigma}) \tag{6}$$

i) Show that the maximum likelihood solution $\boldsymbol{W}_{ML}$ for the parameter matrix $\boldsymbol{W}$ has the property that each column is given by an expression as in equation 2, which was the solution for an isotropic noise distribution. Note that this is independent of the covariance matrix $\boldsymbol{\Sigma}$.

ii) Show that the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \left(\boldsymbol{y}_n - \boldsymbol{W}_{ML}^\top \varphi(\boldsymbol{x}_n)\right) \left(\boldsymbol{y}_n - \boldsymbol{W}_{ML}^\top \varphi(\boldsymbol{x}_n)\right)^\top$$

Exercise 5: Gauss-Markov theorem

i) Let $\boldsymbol{\theta} = \boldsymbol{\alpha}^\top \boldsymbol{w}$ be a linear combination of the parameters $\boldsymbol{w}$. Prove the Gauss-Markov theorem: the least squares estimate of $\boldsymbol{\theta} = \boldsymbol{\alpha}^\top \boldsymbol{w}$ has variance no bigger than that of any other linear unbiased estimate $\boldsymbol{\alpha}^\top \boldsymbol{w}$.

ii) The matrix inequality $\boldsymbol{B} \preceq \boldsymbol{A}$ holds if $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite. Show that if $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix of the least squares estimate of $\boldsymbol{w}$ and $\tilde{\boldsymbol{\Sigma}}$ is the covariance matrix of any other linear unbiased estimate, then $\hat{\boldsymbol{\Sigma}} \preceq \tilde{\boldsymbol{\Sigma}}$.

Exercise 6: Ridge regression

i) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\boldsymbol{w} \sim \text{Gauss}(0, \tau \boldsymbol{I})$ and Gaussian sampling model $\boldsymbol{y} \sim \text{Gauss}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I})$.

ii) Find the relationship between the regularization parameter λ in the ridge formula 7, and the variances τ and σ^2.

$$\sum_{i=1}^{N} \left(y_i - \omega_0 - \sum_{j=1}^{p} x_{ij} \boldsymbol{w}_j \right)^2 + \lambda \sum_{j=1}^{p} \boldsymbol{w}_j^2 \tag{7}$$

Exercise 7: Gaussian parameters

Assume $y_i = \text{Gauss}(w_0 + \boldsymbol{x}_i^\top \boldsymbol{w}, \sigma^2)$, i = 1, 2, ..., N, and the parameters $w_j, j = 1, ..., p$ are each distributed as $\text{Gauss}(0, \tau^2)$, independently of one another. Assuming σ^2 and τ^2 are known, and w_0 is not governed by a prior (or has flat improper prior), show that the (minus) log-posterior density of $\boldsymbol{w}$ is proportional to $\sum_{i=1}^{N} \left(y_i - w_0 - \sum_j x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^{p} w_j^2$ where $\lambda = \frac{\sigma^2}{\tau^2}$

Exercise 8: Artificial data

Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix $\boldsymbol{X}$ with p additional rows $\sqrt{\lambda}\boldsymbol{I}$, and augment $\boldsymbol{y}$ with p zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients towards zero. This is related to the idea of *hints* due to Abu-Mostafa (1995), where model constraints are implemented by adding artificial data examples that satisfy them.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.