Practical Session 3

Linear Regression

1 Weight regularization

Problem 1: Derive the closed form solution for ridge regression error function

$$E_{\text{ridge}}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{\Phi}(x_i) - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

Additionally, discuss the scenario when the number of training samples N is smaller than the number of basis functions M. What computational issues arise in this case? How does regularization address them?

$$\begin{aligned} E_{\text{ridge}}(\boldsymbol{w}) &= \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{\Phi}(x_i) - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \\ &= \frac{1}{2}(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \end{aligned}$$

Taking the gradient

$$\begin{aligned} \nabla_{\boldsymbol{w}}E_{\text{ridge}}(\boldsymbol{w}) &= \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{y} + \lambda\boldsymbol{w} \\ &= (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{y} \end{aligned}$$

Set it to zero

$$\begin{aligned} (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})\boldsymbol{w} &= \boldsymbol{\Phi}^T\boldsymbol{y} \\ \boldsymbol{w} &= (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \end{aligned}$$

In case $N < M$, the covariance matrix $\boldsymbol{\Phi}^T\boldsymbol{\Phi} \in \mathbb{R}^{M\times M}$ will be singular, therefore not invertible. (this may happen even if $N \geq M$, e.g. when some features are correlated).

When regularization is used, $\lambda\boldsymbol{I}$ is added to the covariance matrix, thus fixing the potential degeneracy issue and making the problem tractable.

Problem 2: See Jupyter notebook `practical_03_notebook.ipynb`.

Problem 3: Using singular value decomposition of the design matrix $\boldsymbol{\Phi} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ show that predicted target $\hat{\boldsymbol{y}}$ for the training set when using $\boldsymbol{w}^*_{\text{ridge}}$ can be written as

$$\hat{\boldsymbol{y}} := \boldsymbol{\Phi}\boldsymbol{w}^*_{\text{ridge}} = \sum_{j=1}^{M} \left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda} \boldsymbol{u}_j \boldsymbol{u}_j^T \right) \boldsymbol{y}$$

where $\boldsymbol{u}_j$ are the columns of $\boldsymbol{U}$, d_j the elements of diagonal matrix $\boldsymbol{S}$ and λ the strength of the L_2 regularization. What is the interpretation of this formula?

Based on the SVD of $\boldsymbol{\Phi}$ we can write (the trick is here to rewrite $\lambda\boldsymbol{I}$ as $\lambda\boldsymbol{V}\boldsymbol{V}^T$ and factor matrices out (and remembering that $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$)):

$$\boldsymbol{w}^*_{\text{ridge}} = \boldsymbol{V}(\boldsymbol{S}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}^T\boldsymbol{y}$$

Then

$$\boldsymbol{\Phi}\boldsymbol{w}^*_{\text{ridge}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} = \boldsymbol{U}\boldsymbol{S}(\boldsymbol{S}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{U}^T\boldsymbol{y}$$

First observation, $\boldsymbol{S}^2 + \lambda\boldsymbol{I}$ is a diagonal matrix, with $\sigma_j^2 + \lambda$ on the diagonal, d_j are the singular values of $\boldsymbol{\Phi}$. Therefore, its inverse is again a diagonal, with $1/(\sigma_j^2 + \lambda)$ on the diagonal. And therefore $\boldsymbol{S}(\boldsymbol{S}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}$ is also a diagonal matrix (product of diagonal matrices), with $\frac{\sigma_j^2}{\sigma_j^2+\lambda}$ on the diagonal. This matrix gets multiplied from the right to $\boldsymbol{U}$, i.e. it only scales the columns. Finally the product of two matrices of the form $\boldsymbol{A}\boldsymbol{B}^T$ can be written as the sum of the outer product of the respective columns of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Now let's put the formula in words. First, $\boldsymbol{u}_j^T\boldsymbol{y}$ computes the representation of $\boldsymbol{y}$ with respect to the orthonormal basis $\boldsymbol{U}$, and then reconstructs $\boldsymbol{y}$ in this basis, however with the coordinates *shrunk* ($\lambda > 0$ and thus $\frac{\sigma_j^2}{\sigma_j^2+\lambda} < 1$). A greater amount of shrinkage is applied to the coordinates with smaller singular values. (What does a small singular value mean? We will later discuss that the SVD of $\boldsymbol{\Phi}$ is another way of expressing the *principal components* of the variables in $\boldsymbol{\Phi}$. These are directions in the space spanned by the training examples in which the training data varies, small singular values are directions in which the training data varies very little. Hence, ridge regression shrinks those directions most. The implicit assumption (or justification for this behaviour) is that the output will vary most with those directions that vary most.)

2 Multi-output linear regression

Problem 4: 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.

The observation $\boldsymbol{y}_i$ is a vector with $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{\Sigma})$, $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$, covariance $\boldsymbol{\Sigma}$ is known and fixed for all possible observations. For n i.i.d observed pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, the likelihood is $\prod_i \mathcal{N}(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{\Sigma})$, and thus the negative log-likelihood is const + $\frac{1}{2}\sum_i(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i)$ Let $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$ (the Cholesky decomposition of $\boldsymbol{\Sigma}$), so $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^{-T}\boldsymbol{L}^{-1}$. Using this decomposition write $\boldsymbol{L}^{-1}(\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x_i}) = (\boldsymbol{L}^{-1}\boldsymbol{y}_i - \boldsymbol{L}^{-1}\boldsymbol{W}\boldsymbol{x}_i) = (\hat{\boldsymbol{y}}_i - \hat{\boldsymbol{W}}\boldsymbol{x}_i)$. Using this transformation, the negative log-likelihood now becomes const $+ \frac{1}{2}\sum_i (\hat{\boldsymbol{y}}_i - \hat{\boldsymbol{W}}\boldsymbol{x}_i)^T(\hat{\boldsymbol{y}}_i - \hat{\boldsymbol{W}}\boldsymbol{x}_i)$. Using similar reasoning to the lecture, we can rewrite this as $\mathrm{Tr}(\boldsymbol{\Phi}\hat{\boldsymbol{W}} - \hat{\boldsymbol{Y}})^T(\boldsymbol{\Phi}\hat{\boldsymbol{W}} - \hat{\boldsymbol{Y}})$. Note that $\hat{\boldsymbol{Y}}$ is a matrix that has the vectors $\hat{\boldsymbol{y}}_i$ as its rows. Matrix calculus (derivative of the negative log-likelihood with respect to $\hat{\boldsymbol{W}}$) then gives us $\hat{\boldsymbol{W}}_{MLE} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\hat{\boldsymbol{Y}}$. So these are m single least square problems for every *column* of $\hat{\boldsymbol{Y}}$. Finally, transforming back $\hat{\boldsymbol{W}}_{MLE}$ gives $\boldsymbol{W}_{MLE} = \boldsymbol{L}\hat{\boldsymbol{W}}_{MLE}$.