Exercise 04 of Machine Learning [IN 2064]

Name: Yiman Li
Matr-Nr: 03724352
cooperate with Kejia Chen(03729686)

Problem 1

To find the minimun of the function $E_{weighted}(\boldsymbol{w})$, first define a scalar factor matrix:

$$\boldsymbol{T} = \mathrm{diag}(t_1, t_2, \dots, t_n) \in \mathbb{R}^{N \times N}$$

Then compute the gradient $\nabla_{\boldsymbol{w}} E_{weighted}(\boldsymbol{w})$:

$$\begin{aligned}\nabla_{\boldsymbol{w}} E_{weighted}(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{N} t_i [\boldsymbol{w}^T \phi(\boldsymbol{x}_i) - y_i]^2 \\ &= \nabla_{\boldsymbol{w}} \frac{1}{2} (\boldsymbol{\Phi w} - \boldsymbol{y})^T \boldsymbol{T} (\boldsymbol{\Phi w} - \boldsymbol{y}) \\ &= \nabla_{\boldsymbol{w}} \frac{1}{2} (\boldsymbol{w}^T \boldsymbol{\Phi}^T \boldsymbol{T} \boldsymbol{\Phi w} - 2\boldsymbol{w}^T \boldsymbol{\Phi}^T \boldsymbol{T y} + \boldsymbol{y}^T \boldsymbol{y}) \\ &= \boldsymbol{\Phi}^T \boldsymbol{T} \boldsymbol{\Phi w} - \boldsymbol{\Phi}^T \boldsymbol{T y}\end{aligned} \tag{1}$$

Now set the gradient to zero and solver for $\boldsymbol{w}$ to obtain the minimizer

$$\boldsymbol{\Phi}^T \boldsymbol{T} \boldsymbol{\Phi w} - \boldsymbol{\Phi}^T \boldsymbol{T y} = 0 \tag{2}$$

This leads to the closed form solution for ridge regression error function:

$$\boldsymbol{w}^* = (\boldsymbol{\Phi}^T \boldsymbol{T} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{T y} \tag{3}$$

(1) In terms of the noise on the data, assuming the target variable y is given by a deterministic function $f(\boldsymbol{x}, \boldsymbol{w})$ with additive Gaussian noise so that

$$y_i = f(\boldsymbol{x_i}, \boldsymbol{w}) + \epsilon_i \tag{4}$$

where ϵ_i is a zero mean Gaussian random variable with precision β_i. Thus we can write

$$p(y_i | \boldsymbol{x_i}, \boldsymbol{w}, \beta_i) = \mathcal{N}(t | y(\boldsymbol{x}, \boldsymbol{w}), \beta_i^{-1}) \tag{5}$$

Taking the logarithm of the likelihood function, we have

$$\begin{aligned}\ln p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\beta}) &= \sum_{i=1}^{N} \ln \mathcal{N}(y_i | \boldsymbol{w}^T \phi(\boldsymbol{x_i}), \beta_i^{-1}) \\ &= \frac{1}{2} \sum_{i=1}^{N} \ln \beta_i - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \beta_i [\boldsymbol{w}^T \phi(\boldsymbol{x}_i) - \boldsymbol{y}_i]^2\end{aligned} \tag{6}$$

So $E_weighted(\boldsymbol{w})$ share the similar form with the last item of the equation 6, so we can just interpret t_i as the precision of each Gaussian random variable β_i.
(2) In terms of the data points for which there are exact copies in the dataset, the corresponding weights will be enlarged, since they may play a more mportant role in the estimation.

Problem 2

Augment the design matrix $\boldsymbol{\Phi} \in \mathrm{R}^{N\times M}$ with M additional rows $\sqrt{\lambda}\boldsymbol{I}_{M\times M}$, then we get:

$$\boldsymbol{\Phi'} = \begin{pmatrix} \boldsymbol{\Phi}_{N\times M} \\ \sqrt{\lambda}\boldsymbol{I}_{M\times M} \end{pmatrix} \in \mathrm{R}^{(N+M)\times M} \tag{7}$$

Augment $\boldsymbol{y}$ with M zeros, then we get:

$$\boldsymbol{y}' = \begin{pmatrix} \boldsymbol{y}_{N\times 1} \\ 0_{M\times 1} \end{pmatrix} \in \mathrm{R}^{(N+M)\times 1} \tag{8}$$

Note that $\boldsymbol{y'}^T\boldsymbol{y'} = \boldsymbol{y}^T\boldsymbol{y}$.
Then the ordinary least squares regression are:

$$\begin{aligned}
E_{LS}(\boldsymbol{w}) &= \frac{1}{2}(\boldsymbol{\Phi'}\boldsymbol{w} - \boldsymbol{y}')^T(\boldsymbol{\Phi'}\boldsymbol{w} - \boldsymbol{y}') \\
&= \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{\Phi}'^T\boldsymbol{\Phi'}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{\Phi}'^T\boldsymbol{y}' + \boldsymbol{y}'^T\boldsymbol{y}') \\
&= \frac{1}{2}(\boldsymbol{w}^T(\boldsymbol{\Phi}^T_{M\times N}\ \sqrt{\lambda}\boldsymbol{I}_{M\times M})\begin{pmatrix} \boldsymbol{\Phi}_{N\times M} \\ \sqrt{\lambda}\boldsymbol{I}_{M\times M} \end{pmatrix}\boldsymbol{w} - 2\boldsymbol{w}^T(\boldsymbol{\Phi}^T_{M\times N}\ \sqrt{\lambda}\boldsymbol{I}_{M\times M})\begin{pmatrix} \boldsymbol{y}_{N\times 1} \\ 0_{M\times 1} \end{pmatrix} + \boldsymbol{y}'^T\boldsymbol{y}') \\
&= \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} + \lambda\boldsymbol{w}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}) \\
&= \frac{1}{2}(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}
\end{aligned} \tag{9}$$

which is exactly the form of ridge regression.

Problem 3

To find the minimun of the function $E_{ridge}(\boldsymbol{w})$, compute the gradient $\nabla_{\boldsymbol{w}}E_{ridge}(\boldsymbol{w})$:

$$\begin{aligned}
\nabla_{\boldsymbol{w}}E_{ridge}(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}}(\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{y_i})^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}) \\
&= \nabla_{\boldsymbol{w}}(\frac{1}{2}(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}) \\
&= \boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{y} + \lambda\boldsymbol{w}
\end{aligned} \tag{10}$$

Now set the gradient to zero and solver for $\boldsymbol{w}$ to obtain the minimizer

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{y} + \lambda\boldsymbol{w} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{y} = 0 \tag{11}$$

This leads to the closed form solution for ridge regression error function:

$$\boldsymbol{w}^* = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \tag{12}$$

So when the number of training samples N is smaller than the number of basis functions M, then the coefficients may be finely tuned to the data by developing large positive and negative values so that the corresponding polynomial function matches each of the data points exactly, that is, overfitting problem due to the lack of data.
The rgularization involves adding a penalty term to the error function in order to discourage the coefficients from reaching large value, and the coefficient λ governs the relative importance of the regularization term compared with the sum-of-squares error term.

Problem 4

When we want to predict M target variables, then use the same set of basis functions to model all of the components of the target vector so that

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{W}^T\phi(\boldsymbol{x}) \tag{13}$$

where $\boldsymbol{y}$ is a M-dimensional column verctor, $\boldsymbol{W}$ is an $N \times M$ matrix of parameters, and $\phi(\boldsymbol{x})$ is an N-dimensional column vector with elements $\phi_j(\boldsymbol{x})$, with $\phi_0(\boldsymbol{x}) = 1$. Suppose we take the conditional distribution of the target vector to be an isotropic Gaussian of the form

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{W}, \beta) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}^T\phi(\boldsymbol{x}), \beta^{-1}\boldsymbol{I}) \tag{14}$$

If we have a set of observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$, we can combine these into a matrix $\boldsymbol{Y}$ of size $N \times M$ such that the n_{th} row is given by $\boldsymbol{y}_n^T$. Similarly, we can combine the input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ into a matrix $\boldsymbol{X}$. The log likelihood function is then given by

$$\begin{aligned} \ln p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{W}^T\phi(\boldsymbol{x}_n), \beta^{-1}\boldsymbol{I}) \\ &= \frac{NM}{2}\ln\frac{\beta}{2\pi} - \frac{\beta}{2}\sum_{n=1}^{N}||\boldsymbol{y}_n - \boldsymbol{W}^T\phi(\boldsymbol{x}_n)||^2 \end{aligned} \tag{15}$$

As before, we can maximize this function with respect to $\boldsymbol{W}$, giving

$$\boldsymbol{W}_{MLE} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{Y} \tag{16}$$

Examine this result for each target variable y_k, we have

$$\boldsymbol{w}_m = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}_m = \boldsymbol{\Phi}^\dagger\boldsymbol{y}_m \tag{17}$$

where y_m is an N-dimensional column vector with components y_{nm} for $n = 1, \ldots, N$. Thus the solution to the regression problem decouples between the different target variables, and we need only compute a single pseudo-inverse matrix $\Phi^\dagger$, which is shared by all of the vector $\boldsymbol{w}_m$.

We can also maximize the log likelihood function 15 with respect to the noise precision parameter β, giving

$$\frac{1}{\beta_{MLE}} = \frac{1}{NM}\sum_{n=1}^{N}||\boldsymbol{y}_n - \boldsymbol{W}^T\phi(\boldsymbol{x}_n)||^2 \tag{18}$$

Problem 5

a) For a simple linear regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$, when $\boldsymbol{X}_{new}$ is scalled by a while $\boldsymbol{y}$ remains the same, here we define a matrix $\boldsymbol{A} = \text{diag}(a, a, \ldots, a) \in \mathbb{R}^D$, then

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}_{new}\boldsymbol{w}_{new} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{w}_{new} \tag{19}$$

so we have

$$\boldsymbol{w}_{new} = \boldsymbol{A}^{-1}\boldsymbol{X}^{-1}\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{w} \tag{20}$$

Respectively, $\boldsymbol{w}_i^{new} = \frac{1}{a}\boldsymbol{w}_i$

b) For L_2-regularized linear regression model with the optimal weight vector $\boldsymbol{w}_{new}^* \in \mathbb{R}^D$ as

$$\begin{aligned} \boldsymbol{w}_{new}^* &= \arg\min\frac{1}{2}\sum_{i=1}^{N}(a \cdot \boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \\ &= \arg\min\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \end{aligned} \tag{21}$$

Now set the gradient to zero and solve for $\boldsymbol{w}$ to obtain the minimizer

$$\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{y} + \lambda_{new}\boldsymbol{w} \overset{!}{=} 0 \tag{22}$$

So we get the weight vector

$$\begin{aligned} \boldsymbol{w}_{new} &= (\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X} + \lambda_{new}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{y} \\ &= \boldsymbol{A}^{-1}\boldsymbol{w} \\ &= \boldsymbol{A}^{-1}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \end{aligned} \tag{23}$$

So we can get $\lambda_{new}\boldsymbol{I} = \lambda\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{I}$, namely, $\lambda_{new} = a^2\lambda$.

Problem 6

See the pages below.