Week 3 Homework Submission
Linear Regression

Iuliia Skobleva
Matriculation Number 03723809

November 10, 2019

Least squares regression

Problem 1

The sum of squares error function for a dataset, where each point is weighted is given by:

$$E_{weighted}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} t_i[\mathbf{w^T}\boldsymbol{\phi}(\boldsymbol{x_i}) - y_i]^2 \tag{1}$$

And we want to find $\mathbf{w} = \arg\min E_{weighted}(\mathbf{w})$. In matrix form, the equation 1 can be written as:

$$\begin{aligned} E_{weighted}(\mathbf{w}) &= (\boldsymbol{y} - \boldsymbol{\Phi w})^{\boldsymbol{T}}\boldsymbol{T}(\boldsymbol{y} - \boldsymbol{\Phi w}) \\ &= (\boldsymbol{y^T} - \boldsymbol{w^T\Phi^T})\boldsymbol{T}(\boldsymbol{y} - \boldsymbol{\Phi w}) \\ &= (\boldsymbol{y^T} - \boldsymbol{w^T\Phi^T})(\boldsymbol{Ty} - \boldsymbol{T\Phi w}) \end{aligned} \tag{2}$$

Using the matrix cookbook, we find the gradient and set it to 0 to find $\mathbf{w}$:

$$\nabla_{\mathbf{w}} E_{weighted} = \boldsymbol{\Phi^T T\Phi w} - \boldsymbol{\Phi^T Ty} \stackrel{!}{=} 0 \tag{3}$$

In the end, we have:

$$\boldsymbol{w} = (\boldsymbol{\Phi^T T\Phi})^{\boldsymbol{-1}}\boldsymbol{\Phi^T Ty} \tag{4}$$

Another question is how this weighting factor t_i can be interpreted in terms of the variance of the noise on the data and data points for which there are exact copies in

the dataset. If we choose the weighting factor appropriately, meaning we give lower weight to data points with high variance, we can derive more accurate predictions, $\mathbf{y} = \mathbf{X}\mathbf{w}$. If we have exact copies in the data set, the analysis will be inaccurate, since we cannot differentiate between points that are just noise (with high variance) and points that are a different class. There, applying the weighting might lead to misleading results.

Ridge Regression

Problem 2

In this problem we augment $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$ with $\sqrt{\lambda}\boldsymbol{I}_{M\times M}$ and the target vector $\mathbf{y}$ with M rows of 0s. Now we can show that the ridge regression estimates can be obtained by ordinary least squared regression.

So, using $E_{RR}(\mathbf{w}) = \frac{1}{2}(\hat{\boldsymbol{y}} - \hat{\mathbf{\Phi}}\boldsymbol{w})^{\boldsymbol{T}}(\hat{\boldsymbol{y}} - \hat{\mathbf{\Phi}}\boldsymbol{w})$, where new $\hat{\mathbf{\Phi}}$ and $\hat{\mathbf{y}}$ are given as, $\hat{\mathbf{\Phi}} = \begin{pmatrix} \mathbf{\Phi} \\ \sqrt{\lambda}\boldsymbol{I}_{M\times M} \end{pmatrix}$ and $\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_{M\times 1} \end{pmatrix}$ we find:

$$\begin{aligned} E_{RR}(\mathbf{w}) &= \frac{1}{2}\left[\begin{pmatrix} \mathbf{\Phi} \\ \sqrt{\lambda}\boldsymbol{I} \end{pmatrix} \mathbf{w} - \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \right]^T \left[\begin{pmatrix} \mathbf{\Phi} \\ \sqrt{\lambda}\boldsymbol{I} \end{pmatrix} \mathbf{w} - \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \right] \\ &= \frac{1}{2}\left[\begin{pmatrix} \mathbf{\Phi}\mathbf{w} - \mathbf{y} \\ \sqrt{\lambda}\boldsymbol{I}\mathbf{w} - \mathbf{0} \end{pmatrix} \right]^T \left[\begin{pmatrix} \mathbf{\Phi}\mathbf{w} - \mathbf{y} \\ \sqrt{\lambda}\boldsymbol{I}\mathbf{w} - \mathbf{0} \end{pmatrix} \right] \\ &= \frac{1}{2}\Big[(\mathbf{\Phi}\mathbf{w} - \mathbf{y})^T(\mathbf{\Phi}\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^T\mathbf{w} \Big] \end{aligned} \tag{5}$$

Thus, we find the least squares loss function with L_2-regularization.

Problem 3

We now want to derive the closed form solution for the expression in equation 5, aka find:

$$\mathbf{w} = \arg\min E_{RR}(\mathbf{w}) \tag{6}$$

This is done by calculating the gradient of $E_{RR}(\mathbf{w})$.

$$\nabla_{\mathbf{w}} E_{RR}(\mathbf{w}) = \nabla_{\mathbf{w}} \frac{1}{2}\Big[(\mathbf{\Phi}\mathbf{w} - \mathbf{y})^T(\mathbf{\Phi}\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^T\mathbf{w} \Big] \stackrel{!}{=} 0 \tag{7}$$

We find with the help of the matrix cookbook:

$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^T\mathbf{y} + \lambda\mathbf{I}\mathbf{w} \stackrel{!}{=} 0 \tag{8}$$

This gives the solution for the closed form expression:

$$\mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{9}$$

Another question is what happens when the number of training samples N is smaller than the number of basis functions M. In this case, our model is too flexible and this leads to overfitting. To avoid that, we need to specify the regularization strength λ and make it large enough so that the weights $\mathbf{w}$ are smaller and the model shows small errors on the validation set.

Comparison of Linear Regression Models

Problem 5

We assume that we have fitted an L_2-regularized linear regression model and obtained the optimal weight vector $\mathbf{w}^*$:

$$\mathbf{w}^* = \arg\min \frac{1}{2}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}_i - y_i)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{10}$$

- If we created a new matrix $\mathbf{X_{new}} = a\mathbf{X}$, what weight vector would produce the same results on $\mathbf{X_{new}}$ as $\mathbf{w}^*$ on $\mathbf{X}$? To find out, we look at equation 9 (which is the closed form solution for ridge regression error function), where instead of having the function $\mathbf{\Phi}$, we have the matrix $\mathbf{X}$ thus getting $\mathbf{w}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^T\mathbf{y}$. The new weight vector would be:

$$\mathbf{w}^*_{\mathbf{new}} = (a^2\mathbf{X}^T\mathbf{X} + \lambda_{new}\mathbf{I})^{-1} \cdot a\mathbf{X}^T\mathbf{y} \tag{11}$$

- The predictions are given by $\mathbf{y} = \mathbf{X}\mathbf{w}$ and for the new matrix it is $\mathbf{y_{new}} = \mathbf{X_{new}}\mathbf{w_{new}}$. The prediction for the old dataset equals:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{w} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \cdot \mathbf{X}^T\mathbf{y} \end{aligned} \tag{12}$$

The solution for the weight vector $\mathbf{w_{new}}$ is found in equation 11. For the prediction we have:

$$\begin{aligned} \mathbf{y_{new}} &= a\mathbf{X}\mathbf{w_{new}} \\ &= a\mathbf{X}(a^2\mathbf{X}^T\mathbf{X} + \lambda_{new}\mathbf{I})^{-1} \cdot a\mathbf{X}^T\mathbf{y} \end{aligned} \tag{13}$$

In the above equation we see that if $\lambda_{new} = a^2\lambda$, then $\mathbf{y} = \mathbf{y_{new}}$.