Ludwig-Maximilians-Universitaet Muenchen
Institute for Informatics
Prof. Dr. Volker Tresp
Rajat Koner
Andreas Lohrer
Sebastian Schmoll

Machine Learning

Summer 2021

Exercise Sheet 3

Exercise 3-1 Linear Regression

Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ be a dataset with N samples of dimension D in which the first column contains only ones (to represent the bias), $\mathbf{y} \in \mathbb{R}^N$ a vector with the target values, and $\mathbf{w} \in \mathbb{R}^D$ the weight vector we want to learn. Given the scalar loss

$$L = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w},$$

show that the analytical solution that minimizes the loss L is $\mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.

Hint: Use the following identities: $(\mathbf{UV})^T = \mathbf{V}^T\mathbf{U}^T$, $\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$, $\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T$ and $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

Possible Solution

$$\begin{aligned} L &= (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda \mathbf{w}^T\mathbf{w} \end{aligned}$$

At the minimum of L, the gradient is 0. Thus:

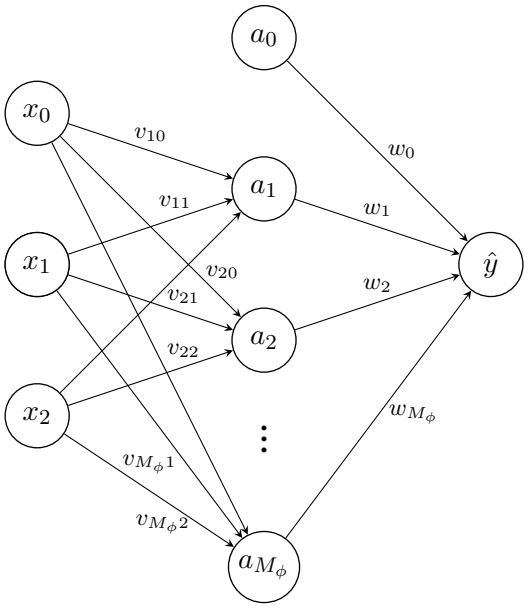
$$\begin{aligned} \nabla_{\mathbf{w}} L \;=\; \frac{\partial L}{\partial \mathbf{w}} &= 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w}^* + 2\lambda\mathbf{w}^* \;=\; 0 \\ &\Leftrightarrow\; 2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w}^* = \mathbf{X}^T\mathbf{y} \\ &\Leftrightarrow\; \mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

Alternatively, use the chain rule to calculate the gradient:

$$\begin{aligned} \nabla_{\mathbf{w}} L \;&=\; -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda\mathbf{w} \;=\; 0 \\ &= -2\mathbf{X}^T\mathbf{y} - 2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w}. \end{aligned}$$

Exercise 3-2 A simple Neural Network

The illustration below depicts a two-layered neural network with inputs $x \in \mathbb{R}^2$ and for each input one bias term $x_0 = 1$, i.e. $\mathbf{x}_i = (1, x_{i,1}, x_{i,2})^T$. Analogously, there is a bias $a_0 = 1$ in the hidden layer.



As activation function for the hidden neurons we employ a sigmoid, i.e.

$$a_h = \sigma(\mathbf{x}_i^T \mathbf{v}_h) = \frac{1}{1 + e^{-\mathbf{x}_i^T \mathbf{v}_h}}; \qquad \forall\, h = 1, \ldots, M_\phi.$$

The output value $\hat{y}$ is calculated as weighted sum of the neurons in the hidden layer: $\hat{y} = \mathbf{a}^T \mathbf{w} = \sum_{j=0}^{M_\phi} a_j w_j$.

(a) Prove that the following holds: $\frac{\partial a_h}{\partial \mathbf{v}_h} = a_h\,(1 - a_h)\,\mathbf{x}_i$

Possible Solution

We need to apply the chain rule: $\frac{\partial a_h}{\partial \mathbf{v}_h} = \frac{\partial \sigma(z)}{\partial z} \cdot \frac{\partial z}{\partial \mathbf{v}_h}$ with $z = \mathbf{x}_i^T \mathbf{v}_h$

First part:

$$\frac{\partial \sigma(z)}{\partial z} = \frac{0 - (-e^{-z})}{(1 + e^{-z})^2} \quad = \quad \frac{e^{-z}}{(1 + e^{-z})^2} \qquad \text{quotient rule!}$$

$$\frac{1}{1 + e^{-z}} \cdot \frac{1 + e^{-z} - 1}{(1 + e^{-z})}$$
$$= \frac{1}{1 + e^{-z}} \left(1 - \frac{1}{1 + e^{-z}}\right)$$
$$= \sigma(z)(1 - \sigma(z))$$

Second part:

$$\frac{\partial z}{\partial \mathbf{v}} = \frac{\partial \mathbf{x}_i^T \mathbf{v}_h}{\partial \mathbf{v}} = \mathbf{x}_i$$

In summary:

$$\frac{\partial a_h}{\partial \mathbf{v}_h} = \sigma(\mathbf{x}_i^T \mathbf{v}_h)(1 - \sigma(\mathbf{x}_i^T \mathbf{v}_h))\,\mathbf{x}_i \;=\; a_h\,(1 - a_h)\,\mathbf{x}_i$$

(b) Express the maximal value of $\hat{y}$ in terms of $\mathbf{w}$ if all weights w_h ($h \in \{0, \ldots, M_\phi\}$) are positive. What's the minimal value?

Possible Solution

- $\hat{y}_{\max}$: The maximal value of each a_h for $h = 1, \ldots M_\phi$ is $\frac{1}{1+0} = 1$. This is the case when $e^{-z_h} = 0$, i.e. for $z_h = \mathbf{x}^T \mathbf{v}_h \to +\infty$. z_0 is the bias term which always has value 1.
  We have: $\hat{y} = \mathbf{a}^T \mathbf{w} = \sum_{h=0}^{M_\phi} a_h \cdot w_h$.
  Thus: $\hat{y}_{\max} = \sum_{h=0}^{M_\phi} 1 \cdot w_h = \sum_{h=0}^{M_\phi} w_h$.
- $\hat{y}_{\min}$: The minimal value of each a_h for $h = 1, \ldots M_\phi$ is $\frac{1}{1+\infty} = 0$. This is the case when $e^{-z_h} = \infty$, i.e. for $z_h = \mathbf{x}^T \mathbf{v}_h \to -\infty$.
  Thus: $\hat{y}_{\min} = \sum_{h=0}^{M_\phi} a_h w_h = w_0 + \sum_{h=1}^{M_\phi} 0 \cdot w_h = w_0$ ($\Rightarrow$ bias-functionality!)

(c) If $v_{h,j} = 0$ for all $j \in \{0, \ldots, M\}, h \in \{1, \ldots, M_\phi\}$, then what is $\hat{y}$?

Possible Solution

For $v_{h,j} = 0$ we have $z_h = \mathbf{x}_i^T \mathbf{0} = 0$ and thus the influence of $\mathbf{x}$ is lost: $a_h = \sigma(0) = \frac{1}{1+e^0} = 0.5$.

With that: $\hat{y} = w_0 + 0.5 \cdot \sum_{h=1}^{M_\phi} w_h$. The prediction is thus the same constant value for any input x.

Exercise 3-3 PyTorch - Feed Forward Neural Network

On the course website you will find an ipython notebook leading you through the implementaion of a simple feed forward neural network in PyTorch for classifying handwritten digits from the MNIST dataset.