CS-E3210- Machine Learning Basic Principles
Home Assignment - "Validation"

Your solutions to the following problems should be submitted as one single pdf which does not contain any personal information (student ID or name). The only rule for the layout of your submission is that for each problem there has to be exactly one separate page containing the answer to the problem. You are welcome to use the LaTeX-file underlying this pdf, available under `https://version.aalto.fi/gitlab/junga1/MLBP2017Public`, and fill in your solutions there.

Problem 1: The Training Error is not the Generalization Error

Answer.

(a) Based on Law of Large Numbers,

$$\lim_{N\to\infty} \mathcal{E}\left(\mathbf{w}|\mathbb{X}\right) = \mathcal{E}\left(\mathbf{w}\right) \tag{1}$$

(b)

$$\begin{aligned}\mathcal{E}\left(\mathbf{w}\right) =& \mathsf{E}\{((\bar{\mathbf{w}}-\mathbf{w})^T\mathbf{x}+\varepsilon)^2\}\\ =& \mathsf{E}\{(\bar{\mathbf{w}}-\mathbf{w})^T\mathbf{x}\mathbf{x}^T(\bar{\mathbf{w}}-\mathbf{w}) + 2\varepsilon(\bar{\mathbf{w}}-\mathbf{w})^T\mathbf{x}+\varepsilon^2\}\\ =& (\bar{\mathbf{w}}-\mathbf{w})^T\mathsf{E}[\mathbf{x}\mathbf{x}^T](\bar{\mathbf{w}}-\mathbf{w}) + 2(\bar{\mathbf{w}}-\mathbf{w})^T\mathsf{E}[\mathbf{x}\varepsilon] + \mathsf{E}[\varepsilon^2]\end{aligned} \tag{2}$$

$$\begin{aligned}(\mathbf{x}^T,\varepsilon)^T &\sim \mathcal{N}(\mathbf{0},\mathbf{C})\\ \mathsf{E}[\mathbf{x}\mathbf{x}^T] = Var(\mathbf{x}) + \mathsf{E}[\mathbf{x}]^2 &= \mathbf{C}_{11}\\ \mathsf{E}[\mathbf{x}\varepsilon] = \mathsf{E}[\mathbf{x}]\mathsf{E}[\varepsilon] + Cov(\mathbf{x},\varepsilon) &= \mathbf{C}_{12} = \mathbf{C}_{21}\\ \mathsf{E}[\varepsilon^2] &= \mathbf{C}_{22}\end{aligned} \tag{3}$$

$$\mathcal{E}\left(\mathbf{w}\right) = (\bar{\mathbf{w}}-\mathbf{w})^T\mathbf{C}_{11}(\bar{\mathbf{w}}-\mathbf{w}) + 2(\bar{\mathbf{w}}-\mathbf{w})^T\mathbf{C}_{12} + \mathbf{C}_{22} \tag{4}$$

(c) We choose $\mathbf{w} = \bar{\mathbf{w}}$, such that the predictor $h^{(\mathbf{w})}$ has small generatlization error $\mathcal{E}\left(\mathbf{w}\right) = \mathbf{C}_{22}$.

Problem 2: Overfitting in Linear Regression

Answer.
Gathering $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ as $\mathbf{X}$. The columns of $\mathbf{X}$ form a linearly independent set. According to the invertible matrix theorem, $\mathbf{X}$ and $\mathbf{X}^T$ are invertible.

$$\begin{aligned}
\nabla_w \mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) &= 0 \\
\frac{1}{N}\nabla_w \|\mathbf{Xw} - \mathbf{y}\|_{\mathbf{2}}^{\mathbf{2}} &= \mathbf{0} \\
\nabla_w (\mathbf{Xw} - \mathbf{y})^{\mathbf{T}}(\mathbf{Xw} - \mathbf{y}) &= \mathbf{0} \\
\nabla_w (\mathbf{w}^T\mathbf{X}^{\mathbf{T}}\mathbf{Xw} - \mathbf{2w}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y} + \mathbf{y}^{\mathbf{T}}\mathbf{y}) &= \mathbf{0} \\
2\mathbf{X}^{\mathbf{T}}\mathbf{Xw} - \mathbf{2X}^{\mathbf{T}}\mathbf{y} &= \mathbf{0} \\
\mathbf{w} &= \mathbf{X}^{-1}\mathbf{y}
\end{aligned} \tag{5}$$

When $\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$, $\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) = \frac{1}{N}\|\mathbf{y} - \mathbf{Xw}\|_{\mathbf{2}}^{\mathbf{2}} = \mathbf{0}$.

Problem 3: Probability of Sampling Disjoint Datasets

Answer.
Sampling without replacement:

$$\begin{aligned} P =& \frac{\mathbf{C}_3^{10}\mathbf{C}_2^7}{\mathbf{C}_3^{10}\mathbf{C}_2^{10}} = \frac{\binom{10}{3}\binom{7}{2}}{\binom{10}{3}\binom{10}{2}} \\ =& \frac{42}{90} \\ \approx& 46.7\% \end{aligned} \tag{6}$$

Sampling with replacement:

$$\begin{aligned} P =& \frac{H_3^{10}H_2^7}{H_3^{10}H_2^{10}} = \frac{C_2^{7+2-1}}{C_2^{10+2-1}} \\ =& \frac{28}{55} \\ \approx& 50.9\% \end{aligned} \tag{7}$$

Problem 4: The Histogram of the Prediction Error

Answer.

(a) To minimize empirical risk $\mathcal{E}(h(\cdot)|\mathbb{X})$,we can directly solve where its gradients are 0. From the Homework 2, Problem 1,

$$\begin{aligned} \nabla_w \mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}^{(train)}) &= 0 \\ \mathbf{w}_{\text{opt}} &= (\mathbf{X^T X})^{-1}\mathbf{X^T y} \\ (\mathbf{X^T X} &\text{ is invertible}) \end{aligned} \tag{8}$$

(b) The sampling method used is sampling from $\mathbb{X}^{(val)}$ without replacement.
The empirical risk is 0.25482349589. The prediction error is small.

(c)

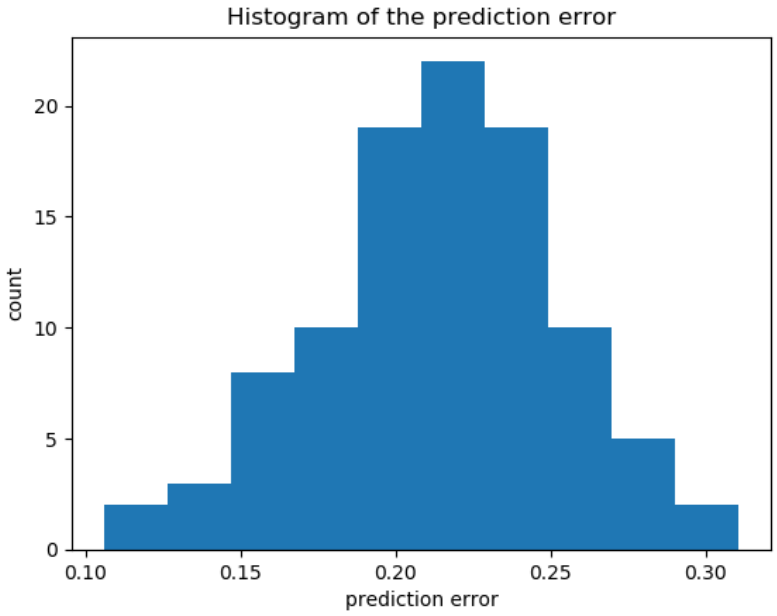

Figure 1: Histogram of the prediction error

From the obtained histogram, we can see that the prediction error fluctuates. And after repeating the experiments many times, there is no obvious rules found in prediction error distribution. So it is not a good idea to evaluate the error only for one single test dataset. It is better to evaluate the error for multiple test dataset and then to compute their mean.

Problem 5: K-fold Cross Validation

Answer.

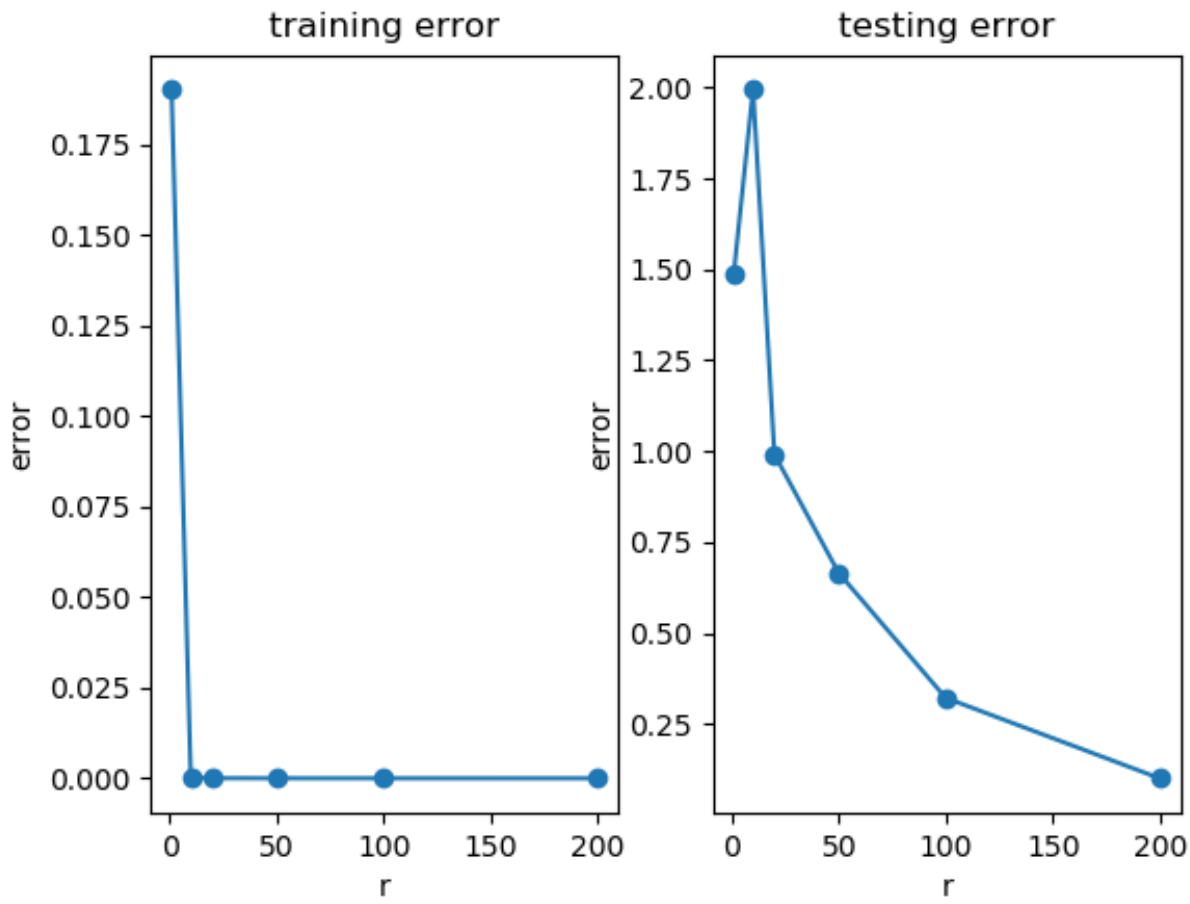

Figure 2: Histogram of the prediction error

We can see that the testing error is minimized when r=200. So the best model complexity is r=200.