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

Consider a folder $\mathbb{X} = \{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(N)}\}$ constituted by N webcam snapshots $\mathbf{z}^{(i)}$, each characterizd by the features $\mathbf{x}^{(i)} \in \mathbb{R}^d$ and labeled by the local temperature $y^{(i)} \in \mathbb{R}$ during the snapshot. We would like to find out how to predict the temperature based solely from the feature vector $\mathbf{x}$. To this end, we will use linear predictors of the form: $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ with some weight vector $\mathbf{w} \in \mathbb{R}^d$.

Let us assume that the features $\mathbf{x}$ and label y are related by a simple linear regression model:

$$y = \bar{\mathbf{w}}^T\mathbf{x} + \varepsilon \tag{1}$$

with some non-random weight vector $\bar{\mathbf{w}} \in \mathbb{R}^d$ and random noise ε. We assume that the feature vector and noise are jointly normal with zero mean and covariance matrix $\mathbf{C}$, i.e., $(\mathbf{x}^T, \varepsilon)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$. The feature vectors $\mathbf{x}^{(i)}$ and labels $y^{(i)}$ are independent and identically distributed (i.i.d.) realizations of $\mathbf{x}$ and y.

(a) Consider the predictor $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ for a particular fixed weight vector $\mathbf{w} \in \mathbb{R}^d$. What is the relation between the empirical risk (training error):

$$\mathcal{E}\left(\mathbf{w}|\mathbb{X}\right) := \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2$$

and the generalization error

$$\mathcal{E}\left(\mathbf{w}\right) = \mathbb{E}\{(y - \mathbf{w}^T\mathbf{x})^2\}?$$

(b) Find a closed-form expression for the generalization error which involves the true (but unknown) weight vector $\bar{\mathbf{w}}$ and covariance matrix $\mathbf{C}$.

(c) According to your results in (b), how should we choose the weight vector $\mathbf{w}$ such that the predictor $h^{(\mathbf{w})}$ has small generalization error?

Answer.

(a) Since

$$\begin{aligned}\mathbb{E}\{\mathcal{E}\left(\mathbf{w}|\mathbb{X}\right)\} &= \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\{\left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{E}\left(\mathbf{w}\right) \\ &= \mathcal{E}\left(\mathbf{w}\right), \end{aligned} \tag{2}$$

the empirical risk is an unbiased estimate of the generalization error.

(b) The covariance matrix $\mathbf{C}$ can be written in block form as $\mathbf{C} = \begin{bmatrix} \mathbb{E}[\mathbf{x}\mathbf{x}^T] & \mathbb{E}[\mathbf{x}\varepsilon] \\ \mathbb{E}[\varepsilon\mathbf{x}^T] & \mathbb{E}[\varepsilon^2] \end{bmatrix} := \begin{bmatrix} \mathbf{C}_{\mathbf{x}} & \mathbf{C}_{\mathbf{x}\varepsilon} \\ \mathbf{C}_{\varepsilon\mathbf{x}} & \mathbf{C}_{\varepsilon} \end{bmatrix}$. Using the linearity of the expectation $\mathbb{E}[\cdot]$, we have

$$\begin{aligned}
\mathcal{E}(\mathbf{w}) = \mathbb{E}[(y - \mathbf{w}^T\mathbf{x})^2] &\overset{y=\bar{\mathbf{w}}^T\mathbf{x}+\varepsilon}{=} \mathbb{E}[(\bar{\mathbf{w}}^T\mathbf{x} - \mathbf{w}^T\mathbf{x} + e)^2] \\
&= \mathbb{E}[((\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{x} + e)^2] \\
&= \mathbb{E}[((\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{x})^2] + \mathbb{E}[e^2] + 2\mathbb{E}[(\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{x}e] \\
&= \mathbb{E}[(\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{x}\mathbf{x}^T(\bar{\mathbf{w}} - \mathbf{w})] + \mathbb{E}[e^2] + 2(\bar{\mathbf{w}} - \mathbf{w})^T\mathbb{E}[\mathbf{x}e] \\
&= \mathbb{E}[(\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{x}\mathbf{x}^T(\bar{\mathbf{w}} - \mathbf{w})] + \mathbb{E}[e^2] + 2(\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{C}_{\mathbf{x}\varepsilon} \\
&= (\bar{\mathbf{w}} - \mathbf{w})^T\mathbb{E}[\mathbf{x}\mathbf{x}^T](\bar{\mathbf{w}} - \mathbf{w}) + \mathbf{C}_{\varepsilon} + 2(\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{C}_{\mathbf{x}\varepsilon} \\
&= (\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{C}_{\mathbf{x}}(\bar{\mathbf{w}} - \mathbf{w}) + 2(\bar{\mathbf{w}} - \mathbf{w})^T\mathbf{C}_{\mathbf{x}\varepsilon} + \mathbf{C}_{\varepsilon}. \qquad (3)
\end{aligned}$$

(c) In order to find the optimal weight vector $\mathbf{w}_{opt} = \operatorname{argmin}_{\mathbf{w}} \mathcal{E}(\mathbf{w})$, we set the gradient of (3) equal to zero and solve the resulting equation. This yields $\mathbf{w}_{opt} = \bar{\mathbf{w}} + \mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{x}\varepsilon}$.

Problem 2: Overfitting in Linear Regression

Consider the problem of predicting a real-valued label (target) $y \in \mathbb{R}$ based on the features $\mathbf{x} \in \mathbb{R}^d$. Given a labeled dataset $\mathbb{X}$ consisting of N labeled data points with feature vectors $\mathbf{x}^{(i)} \in \mathbb{R}^d$ and labels $y^{(i)} \in \mathbb{R}$, we learn a linear predictor $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ by minimizing the empirical risk:

$$\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) := \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - h^{(\mathbf{w})}(\mathbf{x}^{(i)}))^2 = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2.$$

If the dataset $\mathbb{X}$ is small compared to the number d of features, i.e., $N \leq d$, the feature vectors $\{\mathbf{x}^{(i)}\}_{i=1}^N$ are typically linearly independent. Show that in this case, there exists a weight vector $\mathbf{w}_0$ so that $\mathcal{E}(h^{(\mathbf{w}_0)}(\cdot)|\mathbb{X}) = 0$.

Answer. Let us denote $\mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^T \in \mathbb{R}^N$ and $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)})^T \in \mathbb{R}^{N \times d}$. The empirical risk $\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X})$ can then be written as

$$\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2. \tag{4}$$

If the feature vectors $\{\mathbf{x}^{(i)}\}_{i=1}^N$ are linearly independent, $\text{rank}(\mathbf{X}) = N$, i.e., there are N columns of $\mathbf{X}$ which are linearly independent. Without loss of generality, we assume that the first N columns of $\mathbf{X}$ are linearly independent. The matrix $\mathbf{X}$ can be written as $\mathbf{X} = [\mathbf{X}_N, \mathbf{X}_{\bar{N}}]$, where $\mathbf{X}_N \in \mathbb{R}^{N \times N}$ and $\text{rank}(\mathbf{X}_N) = N$, which implies invertibility of $\mathbf{X}_N$. If we define $\mathbf{w}_0 = \begin{bmatrix} \mathbf{X}_N^{-1}\mathbf{y} \\ \mathbf{0}_{N-d} \end{bmatrix}$, with $\mathbf{0}_{N-d} = [0, \ldots, 0]^T \in \mathbb{R}^{N-d}$, then

$$\mathbf{X}\mathbf{w}_0 = [\mathbf{X}_N, \mathbf{X}_{\bar{N}}] \begin{bmatrix} \mathbf{X}_N^{-1}\mathbf{y} \\ \mathbf{0}_{N-d} \end{bmatrix} = \mathbf{X}_N\mathbf{X}_N^{-1}\mathbf{y} + \mathbf{X}_{\bar{N}}\mathbf{0}_{N-d} = \mathbf{y}. \tag{5}$$

Thus,

$$\mathcal{E}(h^{(\mathbf{w}_0)}(\cdot)|\mathbb{X}) \overset{(4)}{=} \frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}_0\|^2 \overset{(5)}{=} 0. \tag{6}$$

.

Problem 3: Probability of Sampling Disjoint Datasets

Consider a dataset $\mathbb{X}$ which contains $N = 10$ different labeled webcam snapshots. We then create a training dataset $\mathbb{X}^{(\text{train})}$ by copying 3 randomly selected elements of $\mathbb{X}$. Moreover, we create a validation dataset $\mathbb{X}^{(\text{val})}$ by copying another 2 randomly selected elements of $\mathbb{X}$. What is the probability that the training set and the validation set are disjoint, i.e., they have no snapshot in common?

Answer. Let $\mathcal{B} := \{\mathcal{B}_i : \mathcal{B}_i \subset \mathbb{X}, |\mathcal{B}_i| = 3\}$, i.e., all the subset of $\mathbb{X}$ with size 3.

Given $\mathbb{X}^{(\text{train})} = \mathcal{B}_i \in \mathcal{B}$, the probability that the training set and the validation set are disjoint, i.e., $|\mathbb{X}^{(\text{train})} \cap \mathbb{X}^{(\text{val})}| = 0$, with $|\mathbb{X}^{(\text{val})}| = 2$, is

$$P(|\mathbb{X}^{(\text{train})} \cap \mathbb{X}^{(\text{val})}| = 0 | \mathbb{X}^{(\text{train})} = \mathcal{B}_i) = \frac{\binom{7}{2}}{\binom{10}{2}} = 0.467, \tag{7}$$

where $\binom{7}{2}$ is the number of possibility of choosing 2 snapshots out of 7 remaining snapshots, i.e., $\mathbb{X} \setminus \mathcal{B}_i$, and $\binom{10}{2}$ is the number of possibility of choosing 2 snapshots out of 10 snapshots.

Since we assume the $\mathbb{X}^{(\text{train})}$ is drawn uniformly from $\mathcal{B}$, $P(\mathbb{X}^{(\text{train})} = \mathcal{B}_i) = 1/|\mathcal{B}|$. Thus,

$$\begin{aligned} P(|\mathbb{X}^{(\text{train})} \cap \mathbb{X}^{(\text{val})}| = 0) &= \sum_{\mathcal{B}_i \in \mathcal{B}} P(|\mathbb{X}^{(\text{train})} \cap \mathbb{X}^{(\text{val})}| = 0 | \mathbb{X}^{(\text{train})} = \mathcal{B}_i) P(\mathbb{X}^{(\text{train})} = \mathcal{B}_i) \\ &= 1/|\mathcal{B}| \sum_{\mathcal{B}_i \in \mathcal{B}} P(|\mathbb{X}^{(\text{train})} \cap \mathbb{X}^{(\text{val})}| = 0 | \mathbb{X}^{(\text{train})} = \mathcal{B}_i) \\ &\overset{(7)}{=} 0.467. \end{aligned} \tag{8}$$

Problem 4: The Histogram of the Prediction Error

Consider the dataset $\mathbb{X}$ available at `https://version.aalto.fi/gitlab/junga1/MLBP2017Public/tree/master/Validation/p3data`. For your convenience, this dataset is already split into a training dataset $\mathbb{X}^{(\mathrm{train})} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N^{(\mathrm{train})}}$ (features $\mathbf{x}^{(i)} \in \mathbb{R}^5$ stored in the file "`X_train.txt`", labels $y^{(i)} \in \mathbb{R}$ stored in "`y_train.txt`") and the validation dataset $\mathbb{X}^{(\mathrm{val})}$ (stored in the files "`X_validation.txt`" and "`y_validation.txt`"). We want to predict the label y given the features $\mathbf{x}$ using a linear predictor $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$.

(a) Learn a linear predictor $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ by choosing the weight vector $\mathbf{w}$ such that the empirical risk (using squared error loss)

$$\mathcal{E}\big(h^{(\mathbf{w})}(\cdot)|\mathbb{X}^{(\mathrm{train})}\big) = (1/|\mathbb{X}^{(\mathrm{train})}|) \sum_{(\mathbf{x},y)\in\mathbb{X}^{(\mathrm{train})}} (y - h^{(\mathbf{w})}(\mathbf{x}))^2$$

obtained for the training dataset $\mathbb{X}^{(\mathrm{train})}$ is as small as possible. Denote this optimal weight vector by $\mathbf{w}_{\mathrm{opt}}$.

(b) Select a test set $\mathbb{X}^{(\mathrm{test})}$ by copying $N^{(\mathrm{test})} = 10$ randomly selected data points $(\mathbf{x}^{(i)}, y^{(i)})$ out of the validation dataset $\mathbb{X}^{(\mathrm{val})}$. Evaluate the prediction error of $h^{(\mathbf{w}_{\mathrm{opt}})}$ by computing the empirical risk

$$\mathcal{E}\big(h^{(\mathbf{w}_{\mathrm{opt}})}(\cdot)|\mathbb{X}^{(\mathrm{test})}\big) = (1/|\mathbb{X}^{(\mathrm{test})}|) \sum_{(\mathbf{x},y)\in\mathbb{X}^{(\mathrm{test})}} (y - h^{(\mathbf{w}_{\mathrm{opt}})}(\mathbf{x}))^2$$

obtained for the test dataset $\mathbb{X}^{(\mathrm{test})}$.

(c) Repeat step (b) $K = 100$ times, involving another test dataset $\mathbb{X}^{(\mathrm{test})}$ each time due to randomness, and generate a histogram of the prediction error. In view of the obtained histogram, is it a good idea to evaluate the error only for one single test dataset ?

Answer.

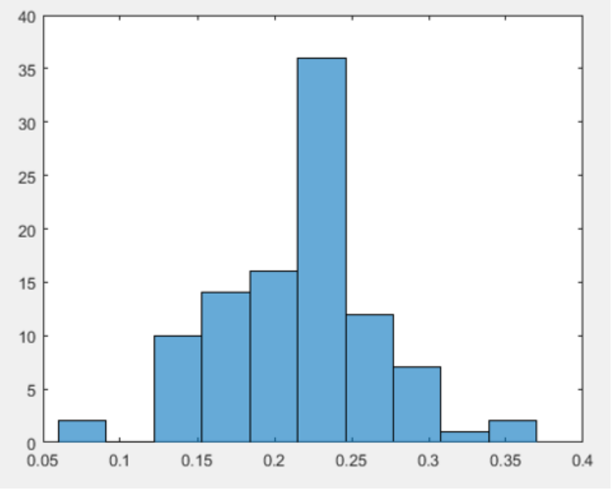

Figure 1: The histogram of the prediction error

Fig. 1 shows the histogram of the prediction error, which spreads out the whole range of $[0, 0.4]$. Therefore, it is not good to evaluate the error only for one single test data set.

Problem 5: K-fold Cross Validation

Consider a dataset $\mathbb{X} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ containing a total of $N = 20$ snapshots (”winter??.jpg” or ”autumn??.jpg” available at `https://version.aalto.fi/gitlab/junga1/MLBP2017Public/tree/master/Validation/WinterFall`) which are either taken either during winter ($y^{(i)} = -1$) or autumn ($y^{(i)} = 1$) . We aim at finding a classifier which classifies an image as “winter” ($\hat{y} = -1$) if $h^{(\mathbf{w})}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) < 1/2$ or as “autumn” ($\hat{y} = 1$) if $h^{(\mathbf{w})}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) \geq 1/2$. Let us collect the image pixels i which belong to the top-left square of size $r \times r$ pixels by $\mathcal{R}_r$. For a given model size r, define the hypothesis space

$$\mathcal{H}^{(r)} := \{h^{(\mathbf{w})}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}), \text{ with } w_i = 0 \text{ for } i \notin \mathcal{R}_r\}.$$

In order to find the best choice for r, we will use “K-fold cross validation” (with $K = 5$) in order to assess the quality of the hypothesis space $\mathcal{H}^{(r)}$ for each $r \in \{1, 10, 20, 50, 100, 200\}$. This works as follows:

- step 1: randomly partition the dataset $\mathbb{X}$ into $K = 5$ equal-size subsets $\mathbb{X}^{(1)}, \ldots, \mathbb{X}^{(K)}$.
- step 2: choose one of the subsets $\mathbb{X}^{(t)}$ as validation set
- step 3: choose the remaining subsets as the training set $\mathbb{X}^{(\mathrm{train}),\mathrm{t}} = \mathbb{X} \setminus \mathbb{X}^{(t)}$
- step 4: find optimal classifier $h^{(\mathbf{w}_{\mathrm{opt,t}})}(\cdot) \in \mathcal{H}^{(r)}$ which minimizes empirical risk

  $$\mathcal{E}\{h^{(\mathbf{w})}|\mathbb{X}^{(\mathrm{train}),\mathrm{t}}\} = (5/N) \sum_{(\mathbf{x},y)\in\mathbb{X}^{(\mathrm{train}),\mathrm{t}}} L((\mathbf{x}, y), h^{(\mathbf{w})}(\cdot))$$

  using logistic loss $L((\mathbf{x}, y), h^{(\mathbf{w})}(\cdot)) = \ln\big(1 + \exp\big(-y(\mathbf{w}^T\mathbf{x})\big)\big)$. You might use gradient descent for determining the optimal weight vector $\mathbf{w}_{\mathrm{opt,t}}$. (see HA3)
- step 5: compute validation error $\mathcal{E}\{h^{(\mathbf{w}_{\mathrm{opt,t}})}|\mathbb{X}^{(t)}\}$
- step 6: repeat from step 2 until every subset $\mathbb{X}^{(t)}$ has been used exactly once for validation
- step 7: compute the average training error $E^{(\mathrm{train})}(r) = (1/5)\sum\limits_{t=1}^{5} \mathcal{E}\{h^{(\mathbf{w}_{\mathrm{opt,t}})}|\mathbb{X}^{(\mathrm{train}),\mathrm{t}}\}$ and the average validation error $E^{(\mathrm{val})}(r) = (1/5)\sum\limits_{t=1}^{5} \mathcal{E}\{h^{(\mathbf{w}_{\mathrm{opt,t}})}|\mathbb{X}^{(t)}\}$

Implement this procedure for each choice $r \in \{1, 10, 20, 50, 100, 200\}$. Plot the average training error $E^{(\mathrm{train})}(r)$ and the average validation error $E^{(\mathrm{val})}(r)$ as functions of the model complexity r. What is the best model complexity for the classification problem at hand? Justify your answer.

Answer. Fig. 2 shows the average training error $E^{(\mathrm{train})}(r)$ and the average validation error $E^{(\mathrm{val})}(r)$ as functions of the model complexity $r \in \{1, 10, 20, 50, 100, 200\}$. As we can see, the validation error is smallest at $r = 20$ which suggests the best model complexity for the classification problem should be $r = 20$.

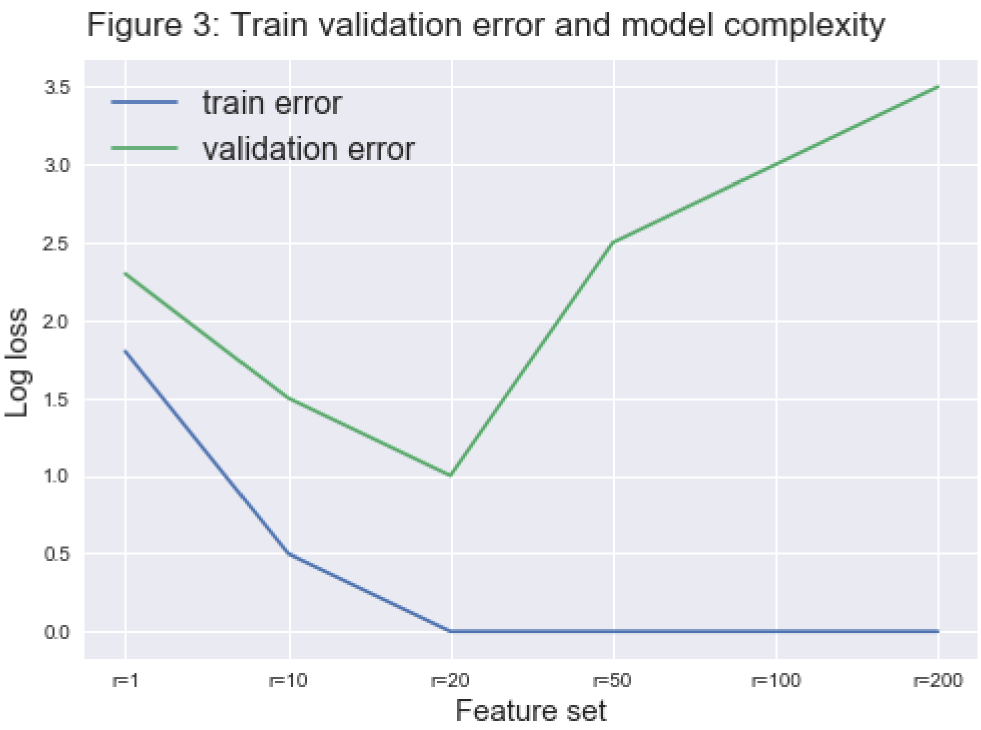

Figure 2: Average training error $E^{(\text{train})}(r)$ and validation error $E^{(\text{val})}(r)$ vs. model complexity.

The Fig. 2 was provided by Sirong Huang. Thank you Sirong.