

Machine Learning, 2021 Spring

Homework 4

Due on 23:59 MAY 5, 2021

Please submit your homework in “pdf” format. Submit the supplementary materials (e.g., files for code) in an **extra** “zip” file.

Problem 1

For a random variable z , let \bar{z} denote its mean, *i.e.*, $\bar{z} = \mathbb{E}[z]$.

Suppose we are given a dataset $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^m$ drawn *i.i.d.* from some *unknown* distribution $P(X, Y)$. Given x , the expected label is defined as

$$\bar{y}(x) = \mathbb{E}_{y|x}[Y] \quad (1)$$

which denotes the label we would expect to obtain.

Next, we run some learning algorithm, such as SVM, linear regression, from which we learned our hypothesis function $h_{\mathcal{D}}$.

Now for a new data point (x, y) sampled from $P(X, Y)$ and out of \mathcal{D} , we want to investigate the expected error between the predicted value $h_{\mathcal{D}}(x)$ and the observation y , *i.e.*,

$$\mathbb{E}_{\mathcal{D}, x, y}[(y - h_{\mathcal{D}}(x))^2]. \quad (2)$$

This error can be decomposed into three parts namely: **variance**, **bias**, and **noise**, where the expectation is taken over all possible training set \mathcal{D} and all (x, y) . Here

$$\begin{aligned} \mathbf{bias}^2 &= \mathbb{E}_x[(\bar{y}(x) - \bar{h}(x))^2] \\ \mathbf{variance} &= \mathbb{E}_{x, \mathcal{D}}[(\bar{h}(x) - h_{\mathcal{D}}(x))^2] \\ \mathbf{noise} &= \mathbb{E}_{x, y}[(y - \bar{y}(x))^2] \end{aligned} \quad (3)$$

where $\bar{h}(x) = \mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)]$ is the “average approximator” by averaging classifiers on all possible training dataset \mathcal{D} .

The error **bias** is the amount by which the expected model prediction differs from the true value or target; while **variance** measures how inconsistent are the predictions from one another, over different training sets, not whether they are accurate or not. **Models that exhibit small variance and high bias underfit the truth target. Models that exhibit high variance and low bias overfit the truth target.**

The data scientist’s goal is to simultaneously reduce bias and variance as much as possible in order to obtain as accurate model as is feasible. However, there is a trade-off to be made when selecting models of different flexibility or complexity and in selecting appropriate training sets to minimize these sources of error.

Question: [2 pts]

Show that

$$\mathbb{E}_{\mathcal{D}, x, y}[(y - h_{\mathcal{D}}(x))^2] = \mathbf{variance} + \mathbf{bias}^2 + \mathbf{noise}. \quad (4)$$

Problem 2

Given the training dataset “data/crime-train.txt” and the test dataset “data/crime-test.txt” (For more information about the datasets, you may refer to “README.md”).

We'd like to use the training dataset to fit a model which can predict the crime rate in new communities, and evaluate model performance on the test set. As there are a considerable number of input variables, overfitting is a serious issue. In order to avoid this, we will use the L2 regularization.

The main goal of this homework is to give you some experience using L2 regularization as a method for variable selection and using 10-folder cross-validation as a technique to get an insight on how the model will generalize to an independent dataset. Your function should accept a scalar value of λ , a vector-valued response variable (\mathbf{y}), a matrix of input variables (\mathbf{X}), and an initial vector of weights (\mathbf{w}_0). It should output a vector of coefficient values ($\hat{\mathbf{w}}$).

In your analysis, include:

1. A plot of $\log(\lambda)$ against the squared error in the 10-folder splited training data. [1 pts]
2. A plot of $\log(\lambda)$ against the squared error in the test data. [0.5 pts]
3. A plot of λ against the number of small coefficients (you can set a threshold), and a brief commentary on the task of selecting λ . [1 pts]
4. For the λ that gave the best test set performance, which variable had the largest (most positive) coefficient? What about the most negative? Discuss briefly. [0.5 pts]

Problem 3

The goal in the prediction problem is to be able to make prediction for the target variable t given some new value of the input variable x on the basis of a set of training data comprising N input values $\mathbf{x} = (x_1, \dots, x_N)^T$ and their corresponding target variable $\mathbf{t} = (t_1, \dots, t_N)^T$.

We assume that, given the value of x , the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$ and the variance σ , where $y(x, \mathbf{w})$ is the prediction function. For example, for the linear regression, the $y(x, \mathbf{w}) = w_0 + w_1x$.

Thus, we have

$$p(t|x, \mathbf{w}, \sigma) = \mathcal{N}(t|y(x, \mathbf{w}), \sigma) \quad (5)$$

Here we only consider the case of a single real-valued variable x . Now you need to use the training data $\{\mathbf{x}, \mathbf{t}\}$ to determine the parameter \mathbf{w} and σ by maximum likelihood.

1. Show that maximizing the log likelihood is equal to minimizing the sum-of-squares error function. [1 pts]
2. More, if we assume that the polynomial coefficients \mathbf{w} is distributed as the Gaussian distribution of the form

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha\mathbf{I}) \quad (6)$$

where α is the parameter of the distribution. Then what is the formulation of the prediction problem? And give us the regularization parameter. Please show us the induction of the procedure.

(Hint. Using Bayes' theorem) [1.5 pts]

Problem 4

In the following problems, we explore the the relationship between the size of the validation set and the expected error.

Let us first look at how the validation set is created. The first step is to partition the data set \mathcal{D} of fixed size N into a training set \mathcal{D}_{train} of size $(N - K)$ and a validation set \mathcal{D}_{val} of size K . We select $N - K$ points at random for training and the remaining for validation. Figure 1 depicts the relationship among \mathcal{D} , \mathcal{D}_{train} and \mathcal{D}_{val} . Suppose we have M models $\mathcal{H}_1, \dots, \mathcal{H}_M$. Given $m \in \{1, \dots, M\}$, denote a $g_m \in \mathcal{H}_m$ as the hypothesis chosen based on data set \mathcal{D} , $g_m^- \in \mathcal{H}_m$ as the hypothesis chosen based on data set \mathcal{D}_{train} as shown in Figure 2.

Now we can evaluate each model on the validation set to obtain the validation errors E_1, \dots, E_M , where

$$E_m = E_{val}(g_m^-), \quad m = 1, \dots, M \quad (7)$$

The validation error can be used as an estimation of the out-of-sample error $E_{out}(g_m^-)$ for each \mathcal{H}_m . It is now a simple matter to select the model with lowest validation error. Let m^* be the index of the model which achieves the minimum validation error. So for \mathcal{H}_{m^*} , $E_{m^*} \leq E_m$ for $m = 1, \dots, M$.

Please answer the following questions:

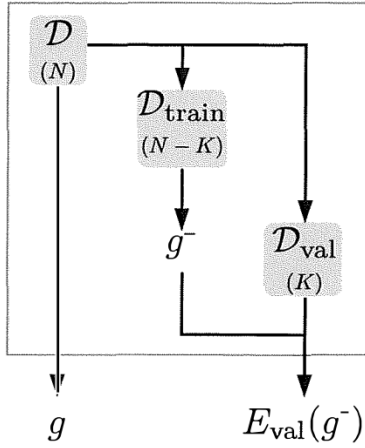


Figure 1: Using a validation set to estimate E_{out}

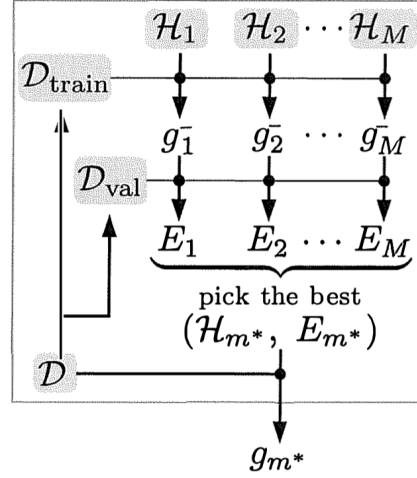


Figure 2: Using a validation set for model selection

Problem 4a

Referring to the Figure 3, why are both curves increasing with K ? Why do they converge to each other with increasing K ? [1 pts]

Problem 4b

Referring to the Figure 4, answer the following 3 problems:

1. $\mathbb{E}[E_{\text{out}}(g_{m^*}^-)]$ is initially decreasing. How can this be, if $\mathbb{E}[E_{\text{out}}(g_m^-)]$ is increasing in K for each m ? [0.5 pts]
2. We see that $\mathbb{E}[E_{\text{out}}(g_{m^*}^-)]$ is initially decreasing, and then it starts to increase. What are possible reason for this? [0.5 pts]
3. When $K = 1$, $\mathbb{E}[E_{\text{out}}(g_{m^*}^-)] < \mathbb{E}[E_{\text{out}}(g_{m^*})]$. How can this be, if the learning curves for both models are decreasing? [0.5 pts]

References

- [1] Abu-Mostafa, Yaser S., 1957-. Learning From Data : a Short Course. [United States] :AMLBook.com, 2012.

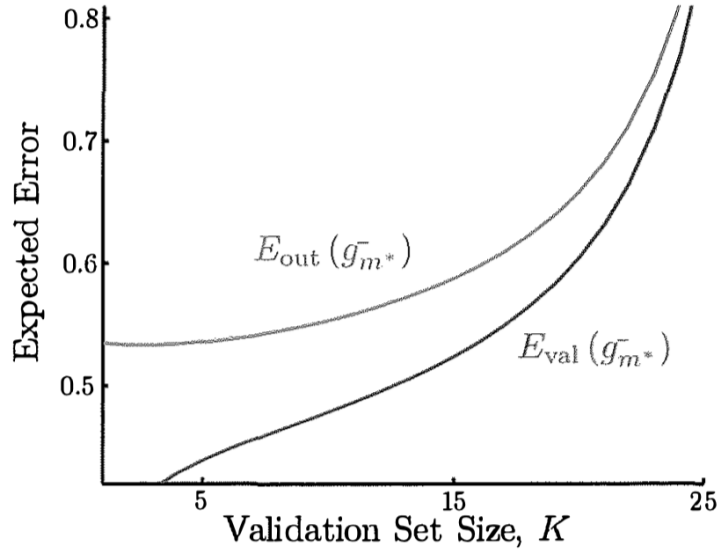


Figure 3: The out-of-sample error $E_{\text{out}}(g_{m^*}^-)$ and the validation error $E_{\text{val}}(g_{m^*}^-)$ versus the size K of \mathcal{D}_{val} .

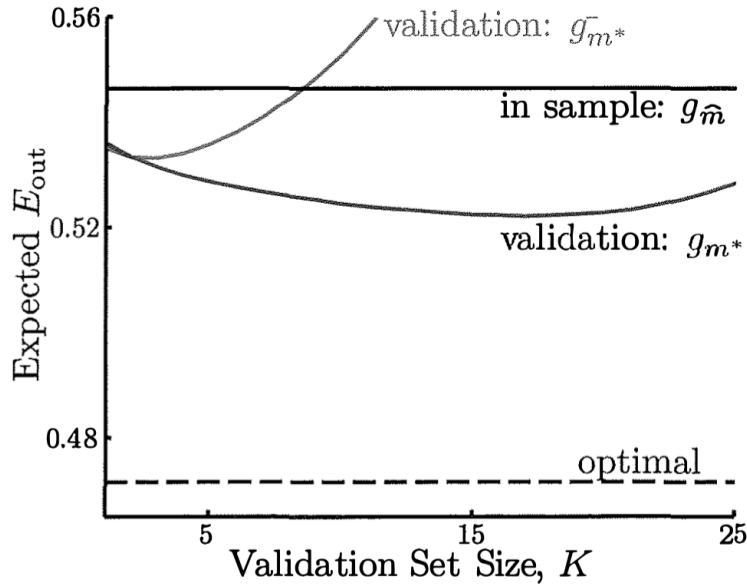


Figure 4: Generalized error versus validation set size K . *validation $g_{m^*}^-$* : the model trained on $\mathcal{D}_{\text{train}}$ with the validation set for model selection. *in sample $g_{\hat{m}}$* : the model trained on \mathcal{D} without validation set. *validation g_{m^*}* : the model that is trained on \mathcal{D} at first for selecting m^* and then is retrained on \mathcal{D} with fixed m^* . The dotted line *optimal* is the optimal model selection, if we could select the model based on the true out of sample error.