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

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

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.*,

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

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

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

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

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

Answer:

$$
\begin{aligned}
& E_{D.x.y}[(y - h(x))^2] \\
&= E_{x.y}[E_D(y - h_D(x))^2] \\
&= E_{x.y}[E_D(y^2 - 2h_D(x)y + h_D(x)^2)] \\
&= E_{x.y}[y^2 - 2E_D(h_D(x))y + E_D(h_D(x)^2)] \\
&= E_{x.y}[E_D(h_D(x))^2 - \bar{h}(x)^2 + \bar{h}(x)^2 - 2E_D(h_D(x))y + y^2] \\
&= E_{x.y}[E_D((\bar{h}(x) - h_D(x))^2) + \bar{h}(x)^2 - 2E_D(h_D(x))y + y^2] \\
&= E_{x.y}[E_D((\bar{h}(x) - h_D(x))^2)] + E_{x.y}[\bar{h}(x)^2 - 2E_D(h_D(x))y + y^2] \\
&= E_{x.D}[(\bar{h}(x) - h_D(x))^2] + E_{x.y}[\bar{h}(x)^2 - 2E_D(h_D(x))y + y^2] \\
&= \textbf{variance} + E_{x.y}[\bar{h}(x)^2 - 2E_D(h_D(x))y + y^2] \\
&= \textbf{variance} + E_x[\bar{h}(x)^2] - 2E_{x,D}[h_D(x)\bar{y}] + E_{x,y}[y^2] + E_x[\bar{y}^2] - E_x[\bar{y}^2] \\
&= \textbf{variance} + E_x[\bar{h}(x)^2 - 2E_D(h_D(x))\bar{y} + \bar{y}^2] + E_{x,y}[y^2] - E_x[\bar{y}^2] \\
&= \textbf{variance} + E_{x.D}[(\bar{y}(x) - \bar{h}(x))^2] + E_{x,y}[y^2] - E_x[\bar{y}^2] \\
&= \textbf{variance} + \textbf{bias}^2 + E_{x,y}[y^2] - E_x[\bar{y}^2] \\
&= \textbf{variance} + \textbf{bias}^2 + E_x[E_y(y^2)] - E_x[E_y(y)^2] \\
&= \textbf{variance} + \textbf{bias}^2 + E_x[E_y(y^2) - E_y(y)^2] \\
&= \textbf{variance} + \textbf{bias}^2 + E_x[E_y(y - \bar{y})^2] \\
&= \textbf{variance} + \textbf{bias}^2 + E_{x.y}(y - \bar{y}(x))^2] \\
&= \textbf{variance} + \textbf{bias}^2 + \textbf{noise}
\end{aligned}
$$

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 ($\boldsymbol{y}$), a matrix of input variables ($\boldsymbol{X}$), and an initial vector of weights ($\boldsymbol{w}_0$). It should output a vector of coefficient values ($\hat{\boldsymbol{w}}$).

In your analysis, include:

1. A plot of $\log(\lambda)$ against the squared error in the 10-folder splited training data. [1 pts]

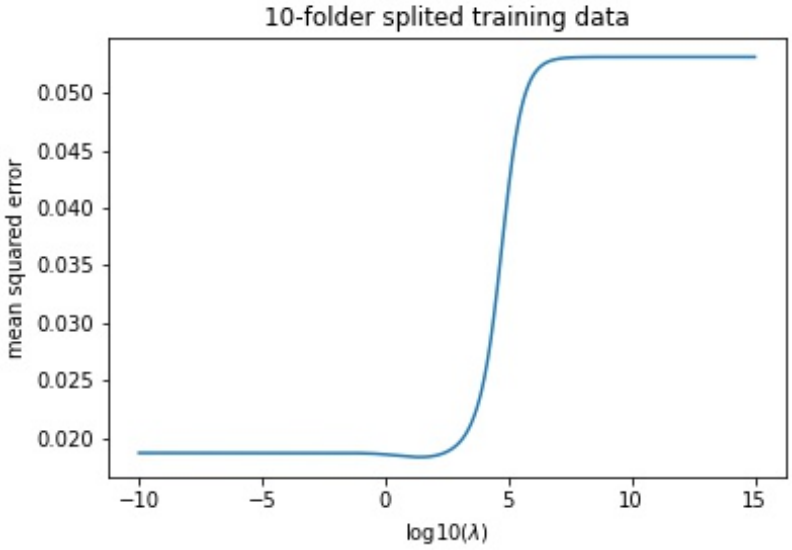

Figure 1: $\log(\lambda)$ against the squared error in the 10-folder splited training data.

2. A plot of $\log(\lambda)$ against the squared error in the test data. [0.5 pts]

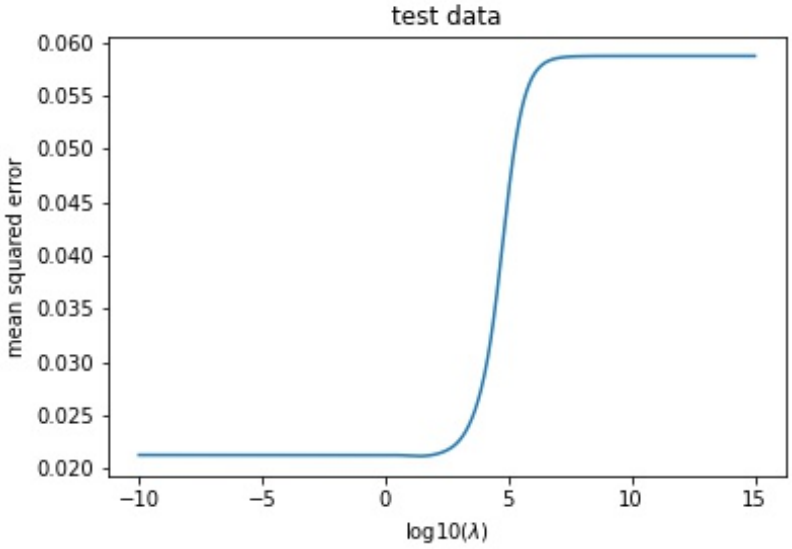

Figure 2: $\log(\lambda)$ against the squared error in the test data.

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]

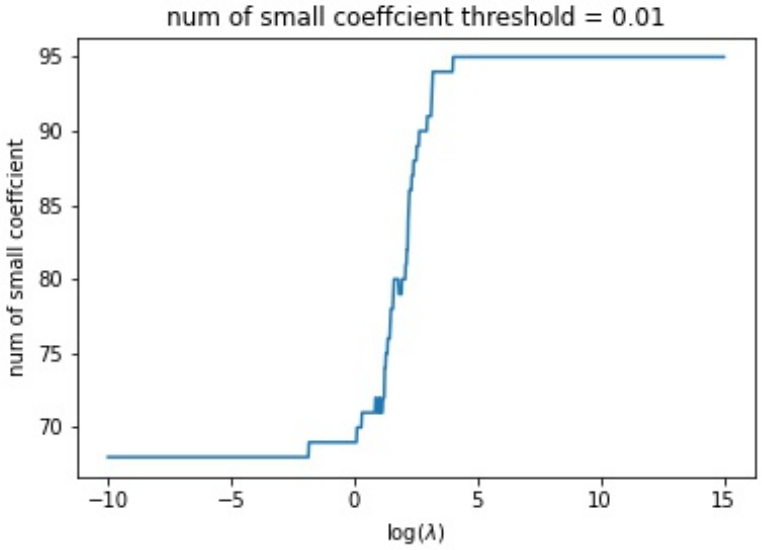

Figure 3: λ against the number of small coefficients

When selecting a λ we should consider two factors: the performance of the test set and the sparsity of coefficient.

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]

   The $\lambda = 22$ gave the best test set performance. The 46th variable PctIlleg has the most positive coefficient. The 40th variable PctKids2Par has the most negative coefficient.

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 $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ and their corresponding target variable $\boldsymbol{t} = (t_1, \ldots, t_N)^T$.

We assume that, given the value of x , the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, w)$ and the variance σ, where $y(x, w)$ is the prediction function. For example, for the linear regression, the $y(x, \boldsymbol{w}) = w_0 + w_1 x$.

Thus, we have

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

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

1. Show that maximizing the log likelihood is equal to minimizing the sum-of-squares error function. [1 pts]
   Answer:
   According to the equation(5) we have:

$$p\left(t|x, \sigma, \omega\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{\left(t - y\left(x, \omega\right)\right)^2}{2\sigma^2}\right] \tag{6}$$

   The likelihood is:

$$L(t|x, \sigma, \omega) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{\left(t - y\left(x, \omega\right)\right)^2}{2\sigma^2}\right]$$

log likelihood is:

$$\ln L(t|x,\sigma,\omega) = -\frac{N}{2}\ln 2\pi\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(t-y(x,\omega))^2$$

Thus,

$$\begin{aligned}\hat{\omega} &= \underset{\omega}{argmax} L\left(\omega|t,x,\sigma,\alpha\right)\\ &= \underset{\omega}{argmin} -\ln(L\left(\omega|t,x,\sigma,\alpha\right))\\ &= \underset{\omega}{argmin}\{-\tfrac{N}{2}\ln 2\pi\sigma - \tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{N}(t-y(x,\omega))^2\}\\ &= \underset{\omega}{argmin}\textstyle\sum_{i=1}^{N}\left[\frac{(t-y(x,\omega))^2}{2\sigma^2}\right]\\ &= \underset{\omega}{argmin}\textstyle\sum_{i=1}^{N}\left(t-y\left(x,\omega\right)\right)^2\end{aligned}$$

$\sum_{i=1}^{N}(t-y(x,\omega))^2$ is the sum-of-squares error function. Thus the maximizing log likelihood is equal to minimizing the sum-of-squares error function.

2. More, if we assume that the polynomial coefficients $\boldsymbol{w}$ is distributed as the Gaussian distribution of the form

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\mathbf{0},\alpha\boldsymbol{I}) \tag{7}$$

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]
Answer:
According to equation (7) we have:

$$p(\omega_0|\alpha) = \frac{1}{\sqrt{2\pi\alpha}}\exp\left[-\frac{\omega_0^2}{2\alpha^2}\right] \tag{8}$$

$$p(\omega_1\alpha) = \frac{1}{\sqrt{2\pi\alpha}}\exp\left[-\frac{\omega_1^2}{2\alpha^2}\right] \tag{9}$$

According to the Bayes' theorem: $p(\omega|t,x) = p(t|x,\omega)p(\omega)$
Then we can construct and maximize the likelihood according to (6) (8) (9):

$$\begin{aligned}L &= \textstyle\prod_{i=1}^{N} p\left(t|x_i,\omega\right)p(\omega)\\ &= \textstyle\prod_{i=1}^{N} p\left(t|x_i,\omega\right)\prod_{j=1}^{2}p(\omega_i)\\ &= (\textstyle\prod_{i=1}^{N} p\left(t|x_i,\omega\right))p(\omega_0)p(\omega_1)\end{aligned}$$

$$\begin{aligned}\hat{\omega} &= \underset{\omega}{argmax} L\left(\omega|t,x,\sigma,\alpha\right)\\ &= \underset{\omega}{argmin} -\ln(L\left(\omega|t,x,\sigma,\alpha\right))\\ &= \underset{\omega}{argmin}\{\tfrac{N}{2}\ln\left(2\pi\sigma\right) + \textstyle\sum_{i=1}^{N}\left[\frac{(t-y(x,\omega))^2}{2\sigma^2}\right] + \ln\left(2\pi\alpha\right) + \tfrac{1}{2\alpha^2}\left(\omega_0^2+\omega_1^2\right)\}\\ &= \underset{\omega}{argmin}\textstyle\sum_{i=1}^{N}\left[\frac{(t-y(x,\omega))^2}{2\sigma^2}\right] + \tfrac{1}{2\alpha^2}\left(\omega_0^2+\omega_1^2\right)\\ &= \underset{\omega}{argmin}\textstyle\sum_{i=1}^{N}\left[\frac{(t-y(x,\omega))^2}{2\sigma^2}\right] + \tfrac{1}{2\alpha^2}\parallel\boldsymbol{\omega}\parallel_2^2\end{aligned}$$

The prediction problem is minimize sum-of-squares error using l2 regression, the regularization parameter is $\frac{1}{2\alpha^2}$.

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 4 depicts the relationship among $\mathcal{D}$, $\mathcal{D}_{train}$ and $\mathcal{D}_{val}$. Suppose we

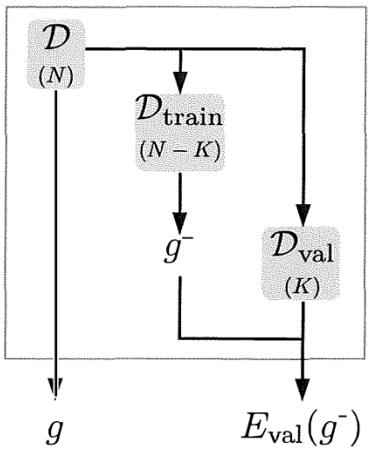

Figure 4: Using a validation set to estimate E_{out}

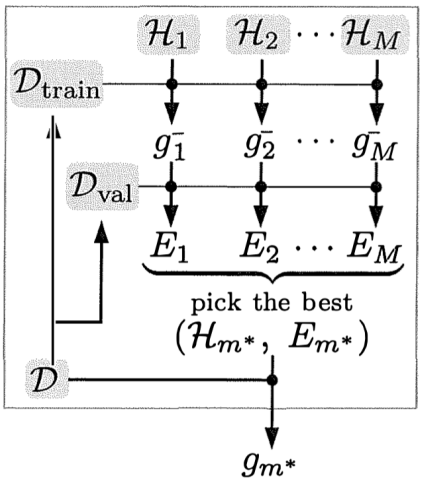

Figure 5: Using a validation set for model selection

have M models $\mathcal{H}_1, \ldots, \mathcal{H}_M$. Given $m \in \{1, \ldots, M\}$, 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 5.

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

$$E_m = E_{val}(g_m^-), \quad m = 1, \ldots, M \tag{10}$$

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 $m = 1, \ldots, M$.

Please answer the following questions:

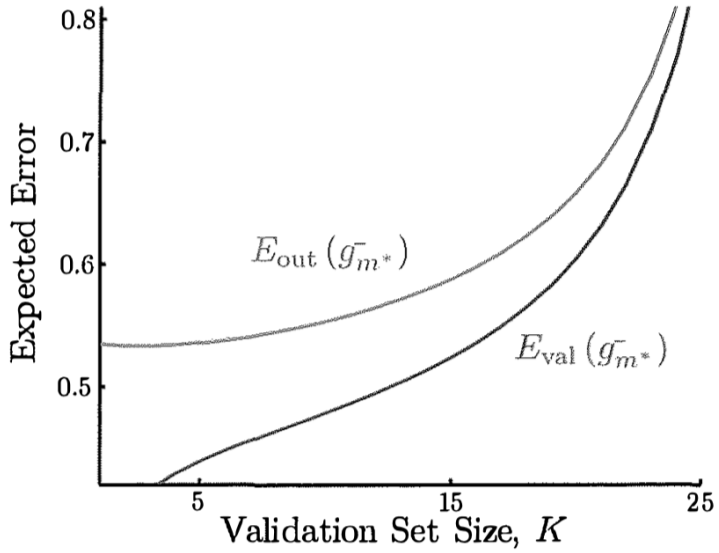

Figure 6: The out-of-sample error $E_{out}(g_{m^*}^-)$ and the validation error $E_{val}(g_{m^*}^-)$ versus the size K of $\mathcal{D}_{val}$.

Problem 4a

Referring to the Figure 6, why are both curves increasing with K? Why do they converge to each other with increasing K? [1 pts]

1.Because when the number of validation set increase, the number of training set will decrease, the training effect will decrease.
2.When the k increase to some value, there are less sets to train the model. The optimistic validation error $E_{val}(g_{m^*}^-)$ is converge to all validation errors $E_{val}(g_m^-)$. Since each single validation error $E_{val}(g_m^-)$ is an unbiased estimate of the $E_{out}(g_m^-)$. We can see that with the increase of K, $E_{val}(g_{m^*}^-)$ converge to $E_{out}(g_m^-)$.

Problem 4b

Referring to the Figure 7, answer the following 3 problems:

1. $\mathbb{E}\left[E_{out}(g_{m^*}^-)\right]$ is initially decreasing. How can this be, if $\mathbb{E}\left[E_{out}(g_m^-)\right]$ is increasing in K for each m? [0.5 pts]
   g_m^- is the result using the training set D_{train}, then using the validation set D_{val} to validate each model, get the best result $g_{m^*}^-$, and the correspond model is H_{m^*}. Then using the all data to train the model H_{m^*} get the result g_{m^*}.
   For the first question, the validation set is very small initially, as the K increase, we could get the better model, thus the $\mathbb{E}\left[E_{out}(g_{m^*}^-)\right]$ is initially decreasing.

2. We see that $\mathbb{E}\left[E_{out}(g_{m^*}^-)\right]$ is initially decreasing, and then it starts to increase. What are possible reason for this? [0.5 pts]
   The reason for the decrease initially has been discussed above. When the number of validation set increase, the number of training set will decrease, the training effect will decrease, the $\mathbb{E}\left[E_{out}(g_{m^*}^-)\right]$ will increase.

3. When $K = 1$, $\mathbb{E}\left[E_{out}(g_{m^*}^-)\right] < \mathbb{E}\left[E_{out}(g_{m^*})\right]$. How can this be, if the learning curves for both models are decreasing? [0.5 pts]
   When K = 1, $g_{m^*}^-$ and g_{m^*} have almost the same size of training set, but $g_{m^*}^-$ is the result after validation, which make sure the $\mathbb{E}[E_{out}(g_{m^*}^-)]$ has small error through small $E_{in}(g_{m^*}^-)$. Thus the result of $g_{m^*}^-$ is better than g_{m^*}.

References

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

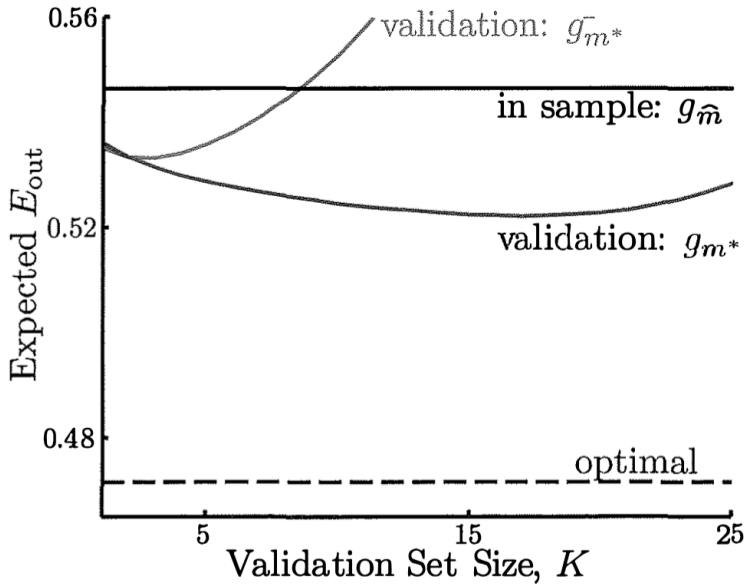

Figure 7: Generalized error versus validation set size K. *validation* $g_{m^*}^-$: the model trained on $\mathcal{D}_{train}$ with the validation set for model selection. *in sample* $g_{\hat{m}}$: the model trained on $\mathcal{D}$ without validation set. *validation* g_{m^*}: the model that is trained 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.