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

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

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

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

1. $E_{D,x,y}[(y-h(x))^2]$

$= E_{D,x,y}[(y-\bar{y}(x)+\bar{y}(x)-h(x))^2]$

$= E_{D,x,y}[(y-\bar{y}(x))^2+2(y-\bar{y}(x))(\bar{y}(x)-h(x))+(\bar{y}(x)-h(x))^2]$

$= E_{D,x,y}[(y-\bar{y}(x))^2]+E_{D,x,y}[(\bar{y}(x)-h(x))^2]+2E_{D,x,y}[(y-\bar{y}(x))(\bar{y}(x)-h(x))]$

$= E_{x,y}[(y-\bar{y}(x))^2]$ [Since y and $\bar{y}(x)$ are functions of x and y] $+ E_{D,x}[(\bar{y}(x)-h(x))^2]+2\int_D\int_x\int_y (y-\bar{y}(x))(\bar{y}(x)-h(x))f(x)f(y|x)f(D)\,dy\,dx\,dD$

$= noise + E_{D,x}[(\bar{y}(x)-\bar{h}(x)+\bar{h}(x)-h_D(x))^2]+2\int_D\int_x[(\bar{y}(x)-h(x))f(x)f(D)\left[\int_y(y-\bar{y}(x))f(y|x)dy\right]dx\,dD$

$= noise + E_{D,x}[(\bar{y}(x)-\bar{h}(x))^2+2(\bar{y}(x)-\bar{h}(x))(\bar{h}(x)-h_D(x))+(\bar{h}(x)-h_D(x))^2]+0$ [$\because \int_y(y-\bar{y}(x))f(y|x)dy = \int_y y f(y|x)dy - \bar{y}(x)\int_y f(y|x)dy$

$= noise + E_{D,x}[(\bar{y}(x)-\bar{h}(x))^2]+E_{D,x}[(\bar{h}(x)-h_D(x))^2]+2E_{D,x}[(\bar{y}(x)-\bar{h}(x))(\bar{h}(x)-h_D(x))]$ $= \bar{y}(x)-\bar{y}(x)=0$]

$= noise + E_x[(\bar{y}(x)-\bar{h}(x))^2]$ [Since $\bar{y}(x)$ and $\bar{h}(x)$ are functions of x] $+ variance + 2\int_x\int_D(\bar{y}(x)-\bar{h}(x))(\bar{h}(x)-h_D(x))f(x)f(D)\,dD\,dx$

$= noise + bias^2 + variance + 2\int_x(\bar{y}(x)-\bar{h}(x))f(x)\left[\int_D(\bar{h}(x)-h_D(x))f(D)dD\right]dx = 0$ [$\because \int_D(\bar{h}(x)-h_D(x))f(D)dD = \bar{h}(x)\int_D f(D)dD - \int_D h_D(x)f(D)dD$

$= variance + bias^2 + noise$ Q.E.D $= \bar{h}(x)-\bar{h}(x)=0$]

Problem 2

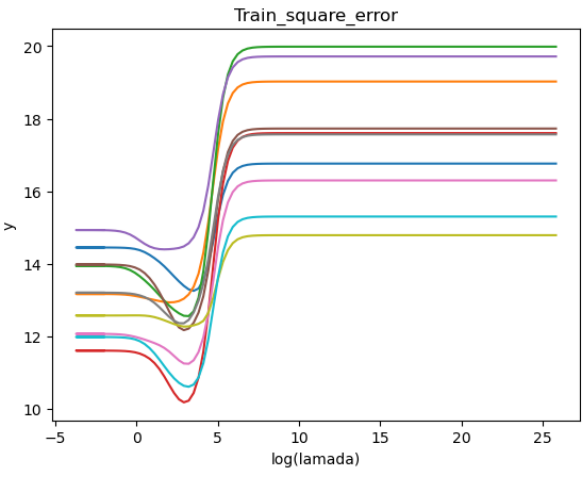
Given the training dataset "data/crime-train.txt" and the test dataset "data/crime-test.txt" (For more information about the datasets, you may refer to "README.md").

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 ($\hat{\boldsymbol{w}}$).

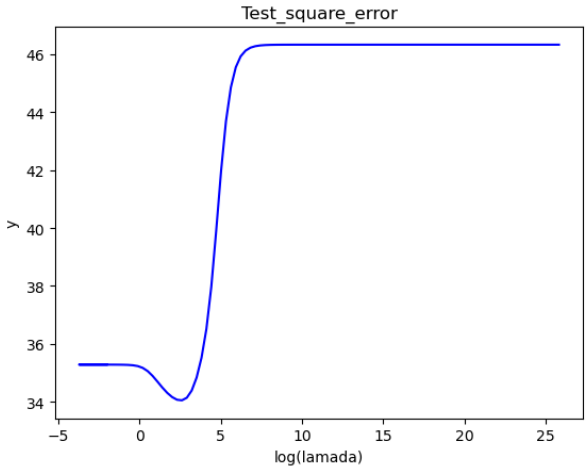
In your analysis, include:

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

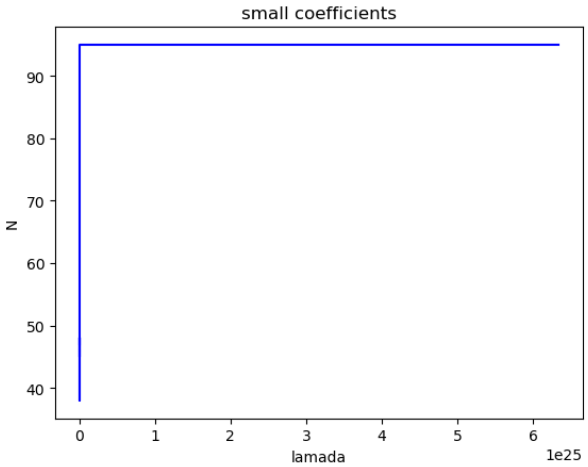


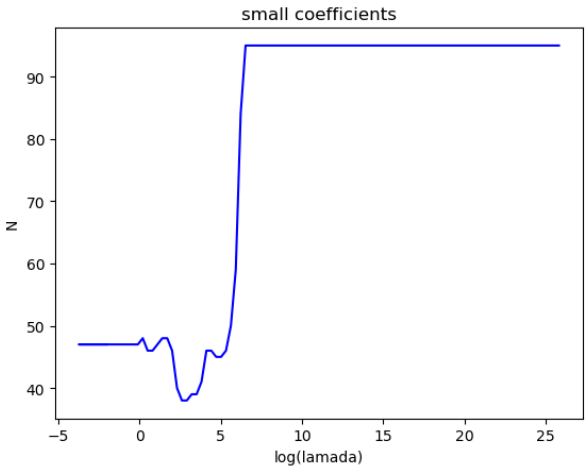
t

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]





(1).I think the graph with λ as x-axis is not so clear about the trend so I also add the graph of $\log(\lambda)$ as x-axis.
(2).The threshold:0.0001
Brief analysis:
(1).Since λ is the penalty coefficient of norm of coefficient,when it is small, it means not give to much restrict to large coefficient,so the variance maybe big because of the extreme situation which had learned out some large coefficients.The generalization is not too bad in this range.
(2).When λ is not too small or too big,it gives some restricts to avoid the extreme situation,which cause the variance down.And it also not give to much restricts so most of the situation of coefficient can be learnt.The generalization is best in this range.[In this case is about 1e3-1e5].
(3).When λ is too big,strict penalty on the coefficient makes them tend to 0.Square error is big and the generalization is bad.

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]
(1).best λ=838.8608
(2).largest coefficient: PctIlleg=0.03244458909756975
smallest coefficient: PctKids2Par=-0.020147945079381974
(3).The same as the analysis in 3,best λ will appear when it is not too small or too big.So it can give some restricts to extreme situation to decrease the variance and also not give too much restricts so that most of the coefficient can be learnt.The generalization is best when λ in this range.

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]

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

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

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]

3.(1) $L(t;x,w,\sigma) = \prod_{i=1}^{N} p(t_i|x_i,w,\sigma)$

$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y(x_i,w)-t_i)^2}{2\sigma^2}}$

$\ln L(t;x,w,\sigma) = \sum_{i=1}^{N} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y(x_i,w)-t_i)^2}{2\sigma^2}}\right)$

$= -\frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}||Xw-t||^2 \qquad X=(x_1,\cdots x_N)^T, x_i=\begin{pmatrix}1\\ x_i\end{pmatrix}, i=1,\cdots N, w=\begin{pmatrix}w_0\\ w_1\end{pmatrix}$

$\max_w \ln L(t;x,w,\sigma) = -\frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}||Xw-t||^2 \Leftrightarrow \max_w -\frac{1}{2\sigma^2}||Xw-t||^2 \Leftrightarrow \min_w \frac{1}{2}||Xw-t||^2$ [sum-of-square error function]

Q.E.D.

(2). $\because p(w|t,x,\sigma) = \frac{p(t|w,x,\sigma)p(w)}{p(t)} \propto p(t|w,x,\sigma)p(w)$

$L(w|t,x,\sigma) = \prod_{i=1}^{N} p(t_i|w,x_i,\sigma)p(w) = \prod_{i=1}^{N} p(t_i|w,x_i,\sigma)\prod_{j=0}^{1} p(w_j)$

$w^* = \arg\max_w L(w|t,x,\sigma)$

$= \arg\max_w e^{\sum_{i=1}^{N}\left(-\frac{(y(x_i,w)-t_i)^2}{2\sigma^2}\right)} \prod_{j=0}^{1} e^{-\frac{w_j^2}{2\alpha^2}}$

$= \arg\max_w \left[\sum_{i=1}^{N}\left(-\frac{(y(x_i,w)-t_i)^2}{2\sigma^2}\right) + \sum_{j=0}^{1} -\frac{w_j^2}{2\alpha^2}\right]$

$= \arg\min_w ||Xw-t||^2 + \frac{1}{2\alpha^2}||w||_2^2 \qquad X=(x_1,\cdots x_N)^T, x_i=\begin{pmatrix}1\\ x_i\end{pmatrix}, i=1,\cdots N, w=\begin{pmatrix}w_0\\ w_1\end{pmatrix}$

which is a L_2-regularization problem

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 $\mathcal{H}_1,\ldots,\mathcal{H}_M$. 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 2.

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

Figure 2: Using a validation set for model selection

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

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

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 $\mathcal{H}_{m^*}$, $E_{m^*} \le E_m$ for $m = 1,\ldots,M$.

Please answer the following questions:

Figure 3: 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 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. $\mathbb{E}\left[E_{out}(g_{m^*}^-)\right]$ 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]
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]
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]

4. (a). 1° Because the size of the train set : $N-k$ is decreasing when K is increasing

2° Because the size of the validation set : K is increasing, by law of large numbers $E_{val}(g_{m^*}^-) \to E_{out}(g_{m^*}^-)$, $K \to \infty$

(b). 1° Since $g_{m^*}^-$ is the smallest E_{in} among M hypotheses, $E_{in}(g_{m^*}^-) \approx E_{out}(g_{m^*}^-)$ for small M and larger K, fixed N, $E_{in}(g_{m^*}^-)$ is decreasing when K is increasing, so $E[E_{out}(g_{m^*}^-)]$ is initially decreasing.

2° The reason for initially decreasing is discussed in (a). The reason for $E[E_{out}(g_{m^*}^-)]$ starts to increase when K is large is because the size of the train set : $N-k$ is decreasing when K is increasing, when K is too large, set for training is too small.

3° The reason for this is that when $K=1$, $g_{m^*}^-$ and g_{m^*} has almost the same size of the train set to produce almost the same hypothesis, that is $g_{m^*}^- \approx g_{m^*}$, but the choose for $g_{m^*}^-$ is based on $\underset{g_n^-, n=1,\cdots,N}{\operatorname{argmin}} E_{val}(g_n^-)$ which guarantee the smallest E_{in} among M hypotheses, that means given a guarantee to not too big $E[E_{out}(g_{m^*}^-)]$ [decrease the effect of noise] but $E[E_{out}(g_{m^*}^-)]$ doesn't have this guarantee, so when the size of train set is similar ($K \ll N$), $E[E_{out}(g_{m^*}^-)] < E[E_{out}(g_{m^*})]$ can be possible.

Figure 4: 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.

References

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