Machine Learning, 2022 Spring
Assignment 3

Notice

Due 23:59 (GMT + 8), May 11, 2022
Plagiarizer will get 0 points.
LaTeXis highly recommended. Otherwise you should write as legibly as possible.

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 $\mathcal{D} = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{m}$ drawn i.i.d. from some unknown distribution $P(X, Y)$. Given x, the expected label is defined as

$$\bar{y}(x) = \mathbb{E}_{y|x}|Y|$$

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.,

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

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

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

where $\bar{h}(x) = \mathbb{E}_{\mathcal{D}}\left[h_{\mathcal{D}}(x)\right]$ 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:

Show that

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

Solution:

At first, we prove the following equation

$$\mathbb{E}\left[z^2\right] = \mathbb{E}\left[(z-\bar{z})^2 + \bar{z}^2\right]$$

Proof:

$$\begin{aligned}\mathbb{E}\left[(z-\bar{z})^2\right] &= \mathbb{E}\left[z^2 - 2z\bar{z} + \bar{z}^2\right] \\ &= \mathbb{E}\left[z^2\right] - 2\mathbb{E}[z]\bar{z} + \bar{z}^2 \\ &= \mathbb{E}\left[z^2\right] - 2\bar{z}^2 + \bar{z}^2 \\ &= \mathbb{E}\left[z^2\right] - \bar{z}^2\end{aligned}$$

Then, the proof is finished. For the original problem, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D},x,y}\left[\left(y - h_{\mathcal{D}}(x)\right)^2\right] = \mathbb{E}_{\mathcal{D},x,y}\left[y^2 - 2h_{\mathcal{D}}(x)y + h_{\mathcal{D}}(x)^2\right] \\ &\overset{(I)}{=} \mathbb{E}_{x,y}\left[y^2\right] - 2\mathbb{E}_{\mathcal{D},x}\left[h_{\mathcal{D}}(x)\right)\mathbb{E}_{x,y}[y] + \mathbb{E}_{\mathcal{D},x}\left[h_{\mathcal{D}}(x)^2\right] \\ &\overset{(II)}{=} \mathbb{E}_{x,y} \mid (y-\bar{y}(x))^2 + \bar{y}(x)^2\Big] - 2\mathbb{E}_{\mathcal{D},x}\left[h_{\mathcal{D}}(x)\right]\mathbb{E}_{\bar{x},y}[y] + \mathbb{E}_{\mathcal{D},x}\left[\left(h_{\mathcal{D}}(x) - \bar{h}(x)\right)^2 + \bar{h}(x)^2\right] \\ &=\text{noise } + \mathbb{R}_x\left[\bar{y}(x)^2\right] - 2\mathbb{E}_{\mathcal{D}}, \left[h\mathcal{D}(x) \mid \mathbb{R}_{x,y}[y] + \text{ variance } + \mathbb{R}_x\left[\bar{h}(x)^2\right]\right. \\ &=\text{noise } + \mathbb{E}_x\left[\bar{y}(x)^2\right] - 2\mathbb{E}_x[\bar{h}(x)]\mathbb{E}_x[\bar{y}(x)] + \text{ variance } + \mathbb{R}_x\left[\bar{h}(x)^2\right] \\ &=\text{noise + variance } + \mathbb{E}_z\left[(\bar{y}(x) - h(x))^2\right] \\ &=\text{noise + variance } + \text{ bias }^2,\end{aligned}$$

where (I) is true because y and $h(x)$ are independent variables, and (II) is true due to the proved equation.

Problem 2

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

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

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

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

1. Show that maximizing the log likelihood is equal to minimizing the sum-of-squares error function.

2. More, if we assume that the polynomial coefficients w is distributed as the Gaussian distribution of the form

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

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)

Solution:

1. The likelihood function is

$$\begin{aligned}L(\boldsymbol{w}, \sigma \mid \boldsymbol{x}, \boldsymbol{t}) &\propto p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{w}, \sigma) \\ &\propto \prod_{i=1}^{N} p\left(t_i \mid x_{\boldsymbol{i}}, \boldsymbol{w}, \sigma\right)\end{aligned}$$

where the parameter w and σ is de termined. Then maximizing the log likelihood is

$$\begin{aligned}
\hat{w} &= \arg\max_{w} \ln(L(w, \sigma \mid \boldsymbol{x}, t)) \\
&= \arg\max_{w} \ln\left(\prod_{i=1}^{N} p\left(t_i \mid x_i, w, \sigma\right)\right) \\
&= \arg\min_{w} -\sum_{i=1}^{N} \ln\left(p\left(t_i \mid x_i, w, \sigma\right)\right) \\
&= \arg\min_{w} -\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(t_i - y\left(x_i, w\right)\right)^2}{2\sigma}\right)\right) \\
&= \arg\min_{w} \sum_{i=1}^{N} \left(t_i - y\left(x_i, w\right)\right)^2
\end{aligned}$$

2. The likelihood function is

$$\begin{aligned}
L(\boldsymbol{w}, \sigma \mid \boldsymbol{x}, \boldsymbol{t}) &\propto p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{w}, \sigma) p(\boldsymbol{w}, \sigma) \\
&\propto \prod_{i=1}^{N} p\left(t_i \mid x_i, w, \sigma\right) \prod_{k=0}^{1} p\left(w_k \mid \alpha\right)
\end{aligned}$$

Then maximizing the log likelihood is

$$\begin{aligned}
\hat{w} &= \arg\min\mathrm{x}_{w} \ln(L(w, \sigma \mid x, t)) \\
&= \arg\mathrm{misx}_{w} \ln\left(\prod_{i=1}^{N} p\left(t_i \mid x_i, w, \sigma\right) \prod_{k=0}^{1} p\left(w_k \mid \alpha\right)\right) \\
&= \arg\min_{w} -\sum_{i=1}^{N} \ln\left(p\left(t_i \mid x_i, w, \sigma\right)\right) - \sum_{k=0}^{1} \ln\left(p\left(w_i \mid \alpha\right)\right) \\
&= \arg\min_{w} -\sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(t_i - y\left(x_i, w\right)\right)^2}{2\sigma}\right)\right) - \sum_{k=0}^{1} \ln\left(\frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{w_k^2}{2\alpha}\right)\right) \\
&= \arg\min_{w} \sum_{i=1}^{N} \frac{\left(t_i - y\left(x_i, w\right)\right)^2}{2\sigma} + \sum_{k=0}^{1} \frac{w_k^2}{2\alpha} \\
&= \arg\min_{w} \sum_{i=1}^{N} \left(t_i - y\left(x_i, w\right)\right)^2 + \frac{\sigma}{\alpha} \sum_{k=0}^{1} w_k^2
\end{aligned}$$

The formulation of the prediction problem is

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{N} \left(t_i - y\left(x_t, w\right)\right)^2 + \frac{\sigma}{\alpha} \sum_{k=0}^{1} w_k^2,$$

where the regularization parameter is σ/α.

Problem 3 (Lecture 11 - p58)

$$Y_i = x_i^T \beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

$$\beta_k \sim \mathcal{N}(0, 1/(2\lambda)) \tag{2}$$

(1) Find the expression for $\widetilde{\lambda}(\sigma, \lambda)$
(2)What regularization does it correspond to if $\beta_k \sim Laplace(0, b)$?

Solution:

(1)

$$\begin{aligned}\hat{\beta} &= \arg\max e^{\sum \frac{-(Y_i - x_i^T\beta)^2}{2\sigma^2}} \prod e^{-\lambda\beta_k^2} \\ &= \arg\min \sum \frac{(Y_i - x_i^T\beta)^2}{2\sigma^2} \sum \lambda\beta_k^2 \\ &= \arg\min \sum (Y_i - x_i\beta)^2 + \sum 2\sigma^2\lambda\beta_k^2 \\ &= \arg\min ||Y - X\beta||_2^2 + \hat{\lambda}||\beta||_2^2\end{aligned}$$

We get $\hat{\lambda} = 2\sigma^2\lambda$.
(2)

$$\begin{aligned}\hat{\beta} &= \arg\max e^{\sum \frac{-(Y_i - x_i^T\beta)^2}{2\sigma^2}} \prod e^{-\frac{|\beta_k|}{b}} \\ &= \arg\min \sum \frac{(Y_i - x_i^T\beta)^2}{2\sigma^2} \sum \frac{|\beta_k|}{b} \\ &= \arg\min \sum (Y_i - x_i\beta)^2 + \sum 2\sigma^2 \frac{|\beta_k|}{b} \\ &= \arg\min ||Y - X\beta||_2^2 + \hat{b}||\beta||_2^2\end{aligned}$$

L1 regularization.

Problem 4 (Lecture 11 - p42)

Generate a set of data (x_i, y_i) according to $y = \theta^T x + z, \theta \sim \mathcal{N}(0,1), z \sim \mathcal{N}(0, 0.1), \theta \in \mathcal{R}^{10}, x \in \mathcal{R}^{10}$. Use ridge regression to estimate the value of θ and plot the path (Regularization coefficient-θ_i).