CS 475 Machine Learning: Homework 1 Supervised Learning 1 Analytical Questions

Due: Thursday, September 24, 2020, 11:59 pm US/Eastern

30 Points Total Version 1.0

Wenkai Luo (wluo14), Yuetong Liu (yliu390)

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not change any formatting in this document, or we may be unable to grade your work. This includes, but is not limited to, the height of textboxes, font sizes, and the spacing of text and tables. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.

We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

Notation

$\mathbf{x_i}$ One input data vector. $\mathbf{x_i}$ is M dimensional. $\mathbf{x_i} \in \mathbb{R}^{1\times M}$.

We assume $\mathbf{x_i}$ is augmented with a 1 to include a bias term.

$\mathbf{X}$ A matrix of concatenated $\mathbf{x_i}$'s. There are N input vectors, so $\mathbf{X} \in \mathbb{R}^{N\times M}$

y_i The true label for input vector $\mathbf{x_i}$. In regression problems, y_i is continuous.

In general ,y_i can be a vector, but for now we assume it's a scalar: $y_i \in \mathbb{R}^1$.

$\mathbf{y}$ A vector of concatenated y_i's. There are N input vectors, so $\mathbf{y} \in \mathbb{R}^{N\times 1}$

$\mathbf{w}$ A weight vector. We are trying to learn the elements of $\mathbf{w}$.

$\mathbf{w}$ is the same number of elements as $\mathbf{x_i}$ because we will end up computing the dot product $\mathbf{x_i} \cdot \mathbf{w}$.

$\mathbf{w} \in \mathbb{R}^{M\times 1}$. We assume the bias term is included in $\mathbf{w}$.

$h((x))$ The true regression function that describes the data.

i.i.d. Independently and identically distributed.

Bias-variance decomposition We can write $E_D[(f(x,D) - h(x))^2] =$

$(E_D[f(x,D) - h(x))^2 + E_D[(f(x,D) - E_D[f(x,D)])^2]$

where the first term is the bias squared, and the second term is the variance.

Notes: In general, a lowercase letter (not boldface), a, indicates a scalar.

A boldface lowercase letter, $\mathbf{a}$, indicates a vector.

A boldface uppercase letter, $\mathbf{A}$, indicates a matrix.

1) Methods of Estimation (10 points)

In class, we discussed estimating the parameters of a Gaussian Linear regression using Maximum Likelihood Estimations and showed its equivalence with Ordinary Least Squares.

(1) (2 points) Consider the following situation, where F is some distribution:

$$X_i \overset{i.i.d}{\sim} F$$
$$\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$$
$$Y_i = X_i^T w + \epsilon_i$$

Assuming σ^2 is known, what is the Maximum Likelihood Estimate (MLE) of w in this situation? How does our MLE estimate of w change if we instead assume $\mathbf{X}$ is a fixed (non-random) matrix?

$$\begin{aligned}\mathbf{w}_{MLE} &= \underset{\mathbf{w}}{\text{argmax}} \prod_{i=1}^{n} P(y_i, \mathbf{x}_i|\mathbf{w}) = \underset{\mathbf{w}}{\text{argmax}} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})P(\mathbf{x}_i)\\ &= \underset{\mathbf{w}}{\text{argmax}} \sum_{i=1}^{n} \log\left[P(y_i|\mathbf{x}_i,\mathbf{w})\right] + \log\left[P(\mathbf{x}_i)\right]\\ &= \underset{\mathbf{w}}{\text{argmax}} \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(\mathbf{x}_i^T\mathbf{w} - y_i)^2 + \log\left[F\right]\end{aligned}$$

where the MLE of $\mathbf{w}$ is the arguement that maxmize the $\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(\mathbf{x}_i^T\mathbf{w} - y_i)^2 + \log\left[F\right]$.

When $\mathbf{X}$ is a fixed matrix, $\log\left[F\right]$ becomes 0, thus the MLE of $\mathbf{w}$ can be computed by following equations:

$$\begin{aligned}\mathbf{w}_{MLE} &= \underset{\mathbf{w}}{\text{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{w} - y_i)^2\\ &= (\mathbf{X}\mathbf{X}^\mathbf{T})^{-1}\mathbf{X}\mathbf{Y}^\mathbf{T} \qquad \text{(According to the Lecture note)}\end{aligned}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$

(2) (3 points) Instead, consider the following case: $\mathbf{X}$ is a fixed matrix, w is a d-dimensional vector.

$$w \sim \mathcal{N}(0, \lambda^2 I_d)$$
$$\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$$
$$Y_i = X_i^T w + \epsilon_i$$

Write out the form of the distribution of $\log p(w \mid Y_i)$ for n i.i.d samples. What expression for w do you get if you maximize this quantity instead of maximising the likelihood? What kind of regularization is being applied here? Write out the regularization penalty in terms of λ.

$\sum_{i=1}^{n} \log p(\mathbf{w} \mid Y_i) = \log p(\mathbf{w} \mid D) = \log \frac{p(D|\mathbf{w})p(w)}{p(D)} = \log p(D \mid \mathbf{w}) + \log p(\mathbf{w}) - \log p(D)$ where D is the data set sampled from the true distribution.

$$\begin{aligned}\mathbf{w} &= \underset{\mathbf{w}}{\operatorname{argmax}} P(\mathbf{w}|D) = \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i, \mathbf{w}) + \log P(\mathbf{w}) \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{w})^2 + \frac{\sigma^2}{\lambda^2}\frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &= (\mathbf{X}\mathbf{X}^{\mathbf{T}} + \frac{\sigma^2}{\lambda^2}\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^T(\text{According to the lecture note})\end{aligned}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$.

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \frac{1}{2}(y_i - \mathbf{x}_i^T\mathbf{w})^2 + \frac{\sigma^2}{\lambda^2}\frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Based on the object function, it can be concluded that it's a L2 regularizer with the following regularization penalty:

$\mathrm{E}_W(\mathbf{w}) = \frac{\sigma^2}{\lambda^2}\frac{1}{2}\mathbf{w}^T\mathbf{w}$

(3) (3 points) Now, repeat the same steps as above, but under the following assumption:

$$w_i \overset{i.i.d}{\sim} Laplace(0, \lambda)$$

What kind of regularization is being applied here? Write out the regularization penalty in terms of λ.

Based on the assumpotion: $w_i \overset{i.i.d}{\sim} Laplace(0, \lambda)$, $p(\mathbf{w})$ can be calculated as $p(\mathbf{w}) = \frac{1}{(2\lambda)^d} e^{-\frac{\sum_{i=1}^{d} |w_i|}{\lambda}}$

$$\begin{aligned}\mathbf{w} &= \underset{\mathbf{w}}{\operatorname{argmax}} P(\mathbf{w}|D) = \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) + \log P(\mathbf{w}) \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{w})^2 + \frac{2\sigma^2}{\lambda}\frac{1}{2}\sum_{i=1}^{d}|w_i|\end{aligned}$$

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{w})^2 + \frac{2\sigma^2}{\lambda}\frac{1}{2}\sum_{i=1}^{d}|w_i|$$

Based on the object function, it can be concluded that it's a L1 regularizer with the following regularization penalty:

$\mathrm{E}_W(\mathbf{w}) = \frac{2\sigma^2}{\lambda}\frac{1}{2}\sum_{i=1}^{d}|w_i|$

(4) (2 points) Suppose you are given a training set and a held out set, and you train a Linear Regression model with L2 regularization on the training set. You try many different values for the regularization penalty λ, and you choose the λ that gives you the best Mean Squared Error (MSE) on a held-out validation set. Now, suppose you are given a new held-out test set. How do you expect your trained model's MSE on this new test set to compare to (a) the model's MSE on your training set? (b) the model's MSE on your validation set? Why?

(a)MSE(test) >MSE(training)

(b)MSE(test) >MSE(validation)

The MSE on both training set and validation set is minimized based on trained model. However, the held-out test set is independent of the training set and validation set. Therefore, its MSE is larger than training set and validaiton set.

2) Bias Variance Trade-off and Regularization (10 points)

In class, we discussed hypothesis classes, the bias variance trade-off, and regularization. Now, we will explore the bias variance trade-off and its interaction with regularization.

Consider the Gaussian Linear Regression model discussed in class:

$$\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$
$$Y_i = w_0 + w_1 X_i + \epsilon_i$$

Suppose the variance σ^2 is known and $\mathbf{X}$ is fixed (non-random).

Based on some domain knowledge, you believe that a simple hypothesis might work well to model this data, and you decide to specify your hypothesis class as zero-order polynomials.

Explicitly, you fit a zero order polynomial to the data $f(X) = \hat{w_0}$.

(5) (3 points) Derive the maximum likelihood estimate for $\hat{w_0}$

Based on the question, we can assume that the model should be:

$$f(x) = w_0 + \epsilon_i, \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

$$p(D|\hat{w_0}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \hat{w_o})^2}{2*\sigma^2}}$$

$$\hat{w}_{oMLE} = \underset{\hat{\mathbf{w_0}}}{\text{argmax}} \log p(D|\hat{w_0}) = \underset{\hat{\mathbf{w_0}}}{\text{argmin}} \sum_{i=1}^{n} (y_i - \hat{w_0})^2 \rightarrow \frac{\partial}{\partial \hat{w_0}} \sum_{i=1}^{n} (y_i - \hat{w_0})^2 = 0$$

$$\rightarrow \sum_{i=1}^{n} (y_i - \hat{w_0}) = 0 \rightarrow \hat{w}_{oMLE} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

(6) (4 points) Compute the bias and the variance of the maximum likelihood estimate of $\hat{w_0}$

$$\text{Bias} = \text{E}\left[\hat{w_o}\right] - w_0 = \frac{1}{n} \sum_{i=1}^{n} (w_0 + w_1 X_i) - w_0 = \frac{1}{n} * n w_0 + \sum_{i=1}^{n} w_1 X_i - w_0 = \sum_{i=1}^{n} w_1 X_i$$

$$\text{Variance} = \text{Var}\left[\hat{w}_{oMLE}\right] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}\left[y_i\right] (y_i \text{ are independent to each other})$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

(7) (3 points) For a new observation (X^*, Y^*), you use the estimate for $\hat{w_0}$ to predict Y^*. Given that you're not using w_1, write out the bias-variance decomposition for the expected value of the mean square error.

Based on the lecture note and provided information that $\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$, the MSE can be decomposed as following equation:

$$\begin{aligned}
\mathrm{E}\left[(Y^* - f(X^*))^2\right] &= \mathrm{Var}[Y^*] + \mathrm{E}[(p(X^*) - \mathrm{E}[f(X^*)])^2] + \mathrm{Var}[f(X^*)] \\
&= \text{observation variance} + \text{model bias}^2 + \text{model variance} \\
\mathrm{Var}[Y^*] &= \mathrm{Var}[\epsilon_i] = \sigma^2 \\
\mathrm{E}[(p(X^*) - \mathrm{E}[f(X^*)])^2] &= (w_0 + w_1 X^* - \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 X_i))^2 = (w_1 X^* - \sum_{i=1}^{n} w_1 X_i)^2 \\
\mathrm{Var}[f(X^*)] &= \mathrm{Var}\left[\hat{w}_{oMLE}\right] = \frac{\sigma^2}{n}
\end{aligned}$$

3) Combining Multiple Regressions (10 points)

Suppose you are given the following data, where F is some distribution:

$$\epsilon_i \overset{i.i.d}{\sim} F$$
$$Y_i = X_i^T w + \epsilon_i$$

Assume X is a fixed matrix. Now instead of training one regression, you randomly split the data into two independent subsets $D_1 = Y_1, \ldots Y_{n/2}$ and $D_2 = Y_{n/2+1} \ldots Y_n$. You then train a regression model on D_1 to learn $w^{(1)}$, and an independent model on D_2 to learn $w^{(2)}$.

The prediction function now looks like:

$$f(X_i) = \frac{1}{2} f^{(1)}(X_i; w^{(1)}) + \frac{1}{2} f^{(2)}(X_i; w^{(2)})$$

Now, given a new point (X^*, Y^*), you want to evaluate the Mean Squared Error:

$$\mathbb{E}[(Y^* - f(X^*))^2]$$

(1) (5 points) Write out the bias-variance decomposition for the MSE:

Based on the lecture note and provided information that $\mathrm{E}[\epsilon_i] = 0$, the MSE can be decomposed as following equation:

$$\begin{aligned}
\mathrm{E}\left[(Y^* - f(X^*))^2\right] &= \mathrm{Var}[Y^*] + \mathrm{E}[(p(X^*) - \mathrm{E}[f(X^*)])^2] + \mathrm{Var}[f(X^*)] \\
&= \text{observation variance} + \text{model bias}^2 + \text{model variance} \\
\mathrm{Var}[Y^*] &= \mathrm{Var}[\epsilon_i] = \mathrm{E}[\epsilon_i^2] \\
\mathrm{E}[(p(X^*) - \mathrm{E}[f(X^*)])^2] &= [(X^*)^T w - (\frac{1}{2}(X^*)^T w_1 + \frac{1}{2}(X^*)^T w_2)]^2 \\
\mathrm{Var}[f(X^*)] &= \frac{1}{4}\mathrm{Var}[f_1(X^*)] + \frac{1}{4}\mathrm{Var}[f_2(X^*)]
\end{aligned}$$

(2) (3 points) Assuming the combined regression function f is unbiased, compare the variance of f with the variance of a linear regression model trained on the entire dataset D.

$$\begin{aligned}
\mathrm{Var}[f(X)] &= \frac{1}{4}\mathrm{Var}[f^{(1)}(X)] + \frac{1}{4}\mathrm{Var}[f^{(2)}(X)] \quad (f^{(1)}(X) \perp f^{(2)}(X)) \\
\mathrm{Var}[f^{(1)}(X_i; w^{(1)})] &\propto \mathrm{Var}[D] \quad (D \text{ is the whole dataset, when } f^{(1)} \text{ is trained on } D) \\
\mathrm{Var}[f^{(1)}(X_i; w^{(1)})] &\propto \mathrm{Var}[\frac{D}{2}] = \frac{1}{4}\mathrm{Var}[D] \quad (\text{when } f^{(1)} \text{ is trained on } \frac{D}{2}) \\
\mathrm{Var}[f(X)] &= \frac{1}{16}\alpha(\mathrm{E}[f^{(1)}(X)])\mathrm{Var}[D] + \frac{1}{16}\alpha(\mathrm{E}[f^{(2)}(X)])\mathrm{Var}[D]
\end{aligned}$$

$\alpha(\mathrm{E}[f^{(\bullet)}(X)])$is a function related to $E[f^{(\bullet)}(X)]$

Assume the dataset is large enough, according to CLT, the mean of D_1 and D_2 are similar to D. Then the variance of the combined model is smaller than the simple model.Otherwise, the mean of D_1 and D_2 are difference from D. Then the combined model could has larger variance

(3) (2 points) Based on your response to the question above, what are the advantages and disadvantages of combining two independent regressions?

Advantage:
When D_1 and D_2 have mean close is D, the variance of combined model is smaller.
Disadvantage:
However, the bias of the combine regression model might be larger that a single regression model obtained from the larger training set due to law of larger number, which says that $\hat{\theta}$ is close to actual θ if sample size approach infinite.
Moreover, the combined model is unstable. When D_1 and D_2 have large variance, the combined model could have higher variance than the simple regression model.