

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

PARTNER1_NAME (PARTNER1_JHED), PARTNER2_NAME (PARTNER2_JHED)

Instructions

We have provided this L^AT_EX 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 We can write $E_D[(f(x, D) - h(x))^2] =$

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

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:

$$X_i \stackrel{i.i.d}{\sim} F$$

$$\epsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

$$Y_i = X_i^T w + \epsilon_i$$

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?

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

$$w \sim \mathcal{N}(0, \lambda^2 I_d)$$

$$\epsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

$$Y_i = X_i^T w + \epsilon_i$$

Write out the form of the distribution of $\log p(w | Y_i)$ 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 .

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

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

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

- (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?

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:

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

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

- (6) (4 points) Compute the bias and the variance of the maximum likelihood estimate of \hat{w}_0

- (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.

3) Combining Multiple Regressions (10 points)

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

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

Assume X is a fixed matrix. Now instead of training one regression, you randomly split the data into two independent subsets $D_1 = Y_1, \dots, Y_{n/2}$ and $D_2 = Y_{n/2+1}, \dots, Y_n$. 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:

- (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 .

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