HW2 Spring 2023 - Written Assignment

Q1 Transformation
15 Points

If your training data set D consists of $N = 100$ points, $(x^{(i)}, y^{(i)})$ for $i = 1, \ldots, N$ where $\mathbf{x}^{(i)}$ consists of a single feature (i.e. $\mathbf{x}^{(i)} = [x_1^{(i)}]$), and we fit two linear regression models

- model 1: $w_0 + w_1 x_1 + w_2 x_1^2$
- model 2: $w_0 + w_1 x_1$

Q1.1 Transformation Function
4 Points

When using model 1, what transformation function[1] $\Phi(\mathbf{x})$ would we use? (Include x_0 as part of your transformation).

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_1^2 \end{bmatrix}$$

[1]: We learned that by performing a nonlinear feature transformation, we can fit a non-linear model as if it is a linear model.

Save Answer Last saved on **Feb 24 at 3:02 PM**

Q1.2 RSS
3 Points

For model 1, express RSS (residual sum of squares) in terms of $\Phi(x)$.

$$\text{RSS} = \sum_{i=1}^{n} (y_i - (\omega^\top * \phi(x)))^2$$

Save Answer | Last saved on **Feb 24 at 3:04 PM**

Q1.3 Suppose
4 Points

Suppose the true relationship between $\mathbf{x}$ and y is $y = w_0 + w_1 x_1 + \epsilon$. Which model would we expect to have a smaller training error, E_{in}?

Model 1

Model 2

Which model would we expect to have a smaller generalization error, E_{out}?

Model 1

Model 2

Explain.

Model 1 offer a more complex hypothesis compared to the true relationship between x and y. Therefore, it will better overfit on all the training datapoints including the ones with noise and lead to a smaller training error E_{in}.

Our future examples will be better approximated by Model 2 because it better fits the relationship complexity (barring the noise) and therefore it will have a smaller generalization error, E_{out} for new examples presented.

Save Answer | Last saved on **Feb 26 at 4:14 PM**

Q1.4 Suppose #2
4 Points

Suppose the true relationship between $\mathbf{x}$ and y is $y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \epsilon$. Which model would we expect to have a smaller training error, E_{in}?

Model 1

Model 2

Which model would we expect to have a smaller generalization error, E_{out}?

Model 1

Model 2

Explain.

We know that the relationship between x and y is truly a linear function with no quadratic term, then Model 2 may have a smaller training error because it is a simpler model that does not try to fit a quadratic relationship between x and y.

Model 2 relatively generalizes better because it does not overfit on a feature that is not part of the true relationship and describes the true linear relationship with ω_1 and ω_0 accurately. Model 1 overfits on a relationship that is not part of the true relationship so its predictions will be inferior.

Save Answer Last saved on **Feb 26 at 4:02 PM**

Q2 Medical Research
15 Points

A medical researcher wishes to evaluate a new diagnostic test for cancer. A clinical trial is conducted where the diagnostic measurement y of each patient is recorded along with attributes of a sample of cancerous tissue from the patient. Three possible models are considered for the diagnostic measurement:

- Model 1: The diagnostic measurement y depends linearly only on the cancer volume.
- Model 2: The diagnostic measurement y depends linearly on the cancer volume and the patient's age
- Model 3 (Extra Credit): The diagnostic measurement y depends linearly on the cancer volume and the patient's age, but the dependence (slope) on the cancer volume is different for two types of cancer - Type I and Type II. (Hint: Use a variable x_3 which is assigned the value 1 if the cancer if Type I, and x_3 has the value 0 if the cancer is of Type II.)

Q2.1 Define Variables
4 Points

Define variables for the cancer volume, age, and cancer type.

x_1 = Cancer volume

x_2 = Age

x_3 = Cancer type

Write a linear model for the predicted value $\hat{y}$ in terms of these variables for models 1 & 2 above.

Model 1: $\hat{y} = \omega_0 + \omega_1 x_1$

Model 2: $\hat{y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2$

Model 3: $\hat{y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3$

Save Answer Last saved on **Feb 25 at 6:54 PM**

Q2.2 Number of Parameters
4 Points

What is the number of parameters in model 1?

2

What is the number of parameters in model 2?

3

Which model is the most complex?

Model 1

Model 2

Save Answer Last saved on **Feb 25 at 6:54 PM**

Q2.3 First Three Rows
4 Points

Since the models in Q2.1 are linear, given training data, we should have $\hat{\mathbf{y}} = X\mathbf{w}$ where $\hat{\mathbf{y}}$ is the vector of predicted values on the training data, X is a design matrix (feature matrix), and $\mathbf{w}$ is the vector of parameters. To test the different models, data is collected from 100 patients. The records for the first three patients are shown below:

Patient ID	Measurement y	Cancer Type	Cancer Volume	Patient Age
12	5	I	0.7	55
34	10	II	1.3	65
23	15	II	1.6	70
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

For model 1 in Q2.1, based on this data, what are the first three rows of the matrix X?

Model 1: $X = \begin{bmatrix} 1 & 0.7 \\ 1 & 1.3 \\ 1 & 1.6 \end{bmatrix}$

For model 2 in Q2.1, based on this data, what are the first three rows of the matrix X?

Model 2: $X = \begin{bmatrix} 1 & 0.7 & 55 \\ 1 & 1.3 & 65 \\ 1 & 1.6 & 70 \end{bmatrix}$

Model 3: $X = \begin{bmatrix} 1 & 0.7 & 55 & \text{I} \\ 1 & 1.3 & 65 & \text{II} \\ 1 & 1.6 & 70 & \text{II} \end{bmatrix}$

Save Answer Last saved on **Feb 25 at 6:58 PM**

Q2.4 Which Model?
3 Points

To evaluate the models, 10-fold cross validation is used with the following results.

Model	Training MSE	Test MSE
1	2.0	2.01
2	0.7	0.72

Model	Training MSE	Test MSE
3	0.65	0.74

Which model should be selected?

Model 1

Model 2

Save Answer Last saved on **Feb 25 at 6:59 PM**

Q3 Model Comparison
15 Points

Suppose you trained your data[1] on three different models and then plotted how well the different fitted models performed with varying amounts of data:

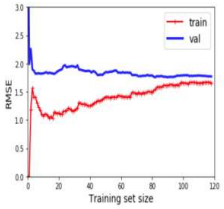

Figure 1: A

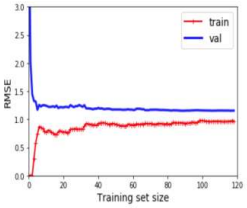

Figure 2: B

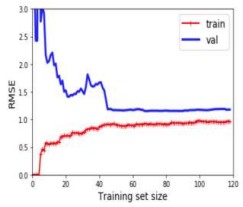

Figure 3: C

(x-axis is "Training set size" and y-axis is "RMSE")

What can you say about overfitting and underfitting?

Model A seems to be underfitting the data since the RMSE is high on both the training set and the validation set. Even by increasing the training set size, we cannot see a decrease in the RMSE for either the training set or the validation set. This means that the training error cannot be reduced without making the hypothesis more complex.

Model B seems to neither overfit nor underfit the data. This is because RMSE is low for both the validation and training sets with only a few data points in the training set.

Model C seems to be overfitting the data since the validation set RMSE is high even with a reasonable training set size (40). The generalization of the model is

poor on the validation set. Also, the RMSE is really low for the training set from the outset. This means that the hypothesis is too complex and does not generalize well to the validation set.

What can you say about the number of examples and the fit of the model?

Model A sees an improvement in the fit of the model as the number of examples increase on the validation set but a decline in the fit of the model on the training set. This is because the model may have an insufficiently complex hypothesis and is underfitting the dataset.

Model B seems to neither overfit nor underfit the data. It reaches a constant error value with relatively few number of examples and then the error increases incrementally for the training set but stays pretty much the same for the validation set. Number of examples has no impact on the fit after a point.

Model C seems to be overfitting the data since the validation set RMSE is high even with a reasonably sized training set (40). The RMSE is really low for the training set from the outset and stays stable as the size of the training set increases. This means that the hypothesis is too complex and does not generalize well to the validation set. Even as you increase the number of examples, the validation set error remains high but training set error remains low.

[1]: The data remains the same amongst the figures

Save Answer Last saved on **Feb 25 at 7:07 PM**

Q4 Flexibility
15 Points

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach[1] for regression or classification?

Advantages of flexible approach (more complex hypothesis class):

- Find non-linear relationships between the input variables, which can be missed by simpler models.

- Ability to tune models to fit complex relationships in the data, which can lead to higher accuracy in predicting outcomes.

- Better identify important features or variables that contribute to the outcome, which can be useful in understanding the underlying relationships in the data.

Disadvantages of flexible approach (more complex hypothesis class):

- Poor generalization on new, unseen data as a result of overfitting and because the model may fit the training data too closely.

- Extra compute resources utilized in order to train the complex model, which also leads to high financial cost.

- Black box problem- it becomes hard to interpret the complex relationships between the variables.

Under what circumstances might a more flexible approach be preferred to a less flexible approach?

It makes sense to take a more flexible approach (more complex hypothesis class) when

- there are known complex (non-linear) relationships between the input and output variables. In this case, lots of variables might take effect, so a flexible model is preferred to fine-tune

- there is plenty of training data available because a flexible approach will tend to overfit on the available data, which if there is plenty, will not occur since the complex variables will capture the complexity in the data.

- high accuracy is required. In this case, it may be a priority to capture all the complex higher-dimensional relationships in the prediction.

When might a less flexible approach be preferred?

It makes sense to take a less flexible approach (less complex hypothesis class) when

- we have limited training data. a less flexible approach may be preferred to avoid overfitting. In these situations, a model that is too complex may memorize the training data instead of learning patterns that generalize well to new data.

- A less flexible approach may be preferred when interpretability is important (avoid black box problem).

- we have a low budget and need computational efficiency. A less flexible approach may be preferred when computational efficiency is a concern because we may not need to train the model multiple times like we would in a case where we have more flexibility.

[1]: i.e. more complex hypothesis class (model class) or a less complex hypothesis class (model class).

Save Answer Last saved on **Feb 25 at 8:39 PM**

Q5 Confidence
15 Points

Consider a binary classification problem ($y \in \{0, 1\}$), where the iid examples

$$D = (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})$$

are divided into two disjoint sets D_{train} and D_{val}

Q5.1 N = 100
8 Points

Suppose you fit a model h using the training set, D_{train}, and then estimated its error using the validation set, D_{val}. If the size of D_{val} was 100 (i.e. $|D_{val}| = 100|$), how confident are you the true error of h is within 0.1 of its average error on D_{val}?

Hoeffding Inequality:

For a given sample of size k from a distribution with values in the range $[a, b]$, the probability that the E_{out} from E_{in} by more than ϵ is bounded by:

$$\delta = 2e^{-2\epsilon^2 * k}$$

where $\epsilon > 0$.

Applying the Hoeffding inequality to estimate the confidence interval for the true error of h:

Using the Hoeffding inequality, we get:

$$\delta = 2e^{-2(0.1)^2 * 100}$$

where $k = 100$ is the sample size.

Simplifying the expression, we get:

$$\delta = 0.27067$$

This means that with probability $1 - \delta = 0.72933$ the true error is within 0.1 of its average error on D_{val}, using the Hoeffding inequality.

Save Answer Last saved on **Feb 25 at 7:12 PM**

Q5.2 N = 200
7 Points

Repeat the previous question where now $|D_{val}| = 200$ (i.e you have 200 examples in your validation set).

Hoeffding Inequality:

For a given sample of size k from a distribution with values in the range $[a, b]$, the probability that the E_{out} from E_{in} by more than ϵ is bounded by:

$$\delta = 2e^{-2\epsilon^2 * k}$$

where $\epsilon > 0$.

Applying the Hoeffding inequality to estimate the confidence interval for the true error of h:

Using the Hoeffding inequality, we get:

$$\delta = 2e^{-2(0.1)^2 * 200}$$

where $k = 200$ is the sample size.

Simplifying the expression, we get:

$$\delta = 0.03663$$

This means that with probability $1 - \delta = 0.96337$ the true error is within 0.1 of its average error on D_{val}, using the Hoeffding inequality.

Save Answer Last saved on **Feb 25 at 10:25 PM**

Q6 Closed Form Solution
15 Points

Suppose you are given the following dataset, where the target variable is MED:

RM	RAD	DIS	MED
6.6	1	4.0	24.0
6.4	2	5.0	21.6
7.2	2	5.0	34.7
6.4	2	5.0	21.6
7.2	2	5.0	34.7

RM	RAD	DIS	MED

Using the data above, write the *equation* derived in the lecture notes (Slide 17 of "Topic 2 part 4 model selection Fall 2022 PDF") to compute the closed form solution for ridge regression where $\lambda = 0.1$. You do not need to actually calculate the coefficient vector - just set up the formula using the numbers given above.

The closed-form solution for ridge regression is written as:

$$w_{\text{ridge}} = (X^\top X + N\lambda I')^{-1} X^\top y$$

$$X = \begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix}$$

$$y = \begin{bmatrix} 24.0 \\ 21.6 \\ 34.7 \\ 21.6 \\ 34.7 \end{bmatrix}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6.6 & 6.4 & 7.2 & 6.4 & 7.2 \\ 1 & 2 & 2 & 2 & 2 \\ 4.0 & 5.0 & 5.0 & 5.0 & 5.0 \end{bmatrix}$$

$$I' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$w_{\text{ridge}} = (\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6.6 & 6.4 & 7.2 & 6.4 & 7.2 \\ 1 & 2 & 2 & 2 & 2 \\ 4.0 & 5.0 & 5.0 & 5.0 & 5.0 \end{bmatrix} * \begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix} + N * 0.1 * \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix})^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6.6 & 6.4 & 7.2 & 6.4 & 7.2 \\ 1 & 2 & 2 & 2 & 2 \\ 4.0 & 5.0 & 5.0 & 5.0 & 5.0 \end{bmatrix} * \begin{bmatrix} 24.0 \\ 21.6 \\ 34.7 \\ 21.6 \\ 34.7 \end{bmatrix}$$

Save Answer Last saved on **Feb 26 at 4:13 PM**

Q7 Gradient Descent
10 Points

Write the gradient descent algorithm (vectorized or not) for ridge regression

Here's the gradient descent algorithm for ridge regression:

Given a training dataset with features $\mathbf{X}$, target values $\mathbf{y}$, regularization parameter λ, and learning rate α, we iteratively update the weights:

for a fixed number of iterations:

Compute the gradient of the loss function with respect to the weights:

$$\nabla J(\mathbf{w}) = \frac{2}{n}(\mathbf{X}^T * \mathbf{X}\mathbf{w} - \mathbf{X}^T * \mathbf{y}) + 2\lambda \mathbf{I}'\mathbf{w}$$

Update the weights using the gradient:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla J(\mathbf{w})$$

.

Note that the L2 regularization term helps to prevent overfitting by penalizing large weights. The regularization parameter λ controls the strength of the regularization.

Save Answer Last saved on **Feb 26 at 4:15 PM**

Q8 Typed & Lateness (Ignore this)
0 Points

Save Answer

Save All Answers

Submit & View Submission ›