ID2223
Scalable Machine Learning and Deep Learning

Review Questions 1

FALL 2021

Yuchen Gao
Weikai Zhou

November 12, 2021

KTH EECS Embedded Systems

1 Which of the following is/are true about Normal Equation?

(a) We don't have to choose the learning rate.

(b) It becomes slow when number of features is very large.

(c) No need to iterate.

All (a), (b) and (c) are correct.

Normal equation is $w = (X^T X)^{-1} X^T y$, which is an expression for calculating the w that can minimize J(w). From this equation, we know that we don't need to either choose the learning rate or iterate. Besides, if there are plenty of features, we know that the information stored in X will increase, and it will become time-consuming to calculate $(X^T X)^{-1} X^T y$. Therefore, All (a), (b) and (c) are correct.

2 The following graph (Figure 1) represents a regression line predicting y from x. The values on the graph shows the residuals for each predictions value, i.e., $\hat{y} - y$. Calculate the squared error of the prediction.

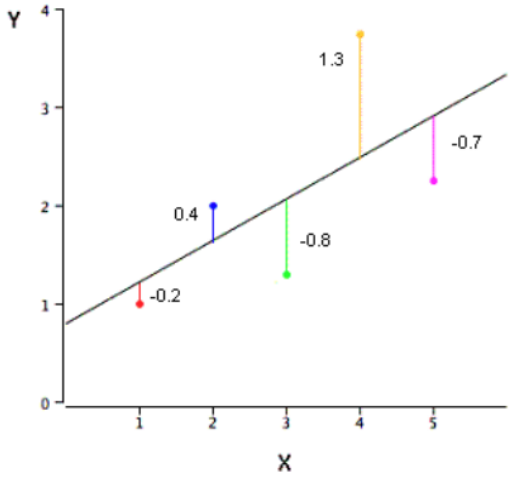


Figure 1: Regression Line

The squared error of the prediction can be calculates as follows

$$\sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 = 0.04 + 0.16 + 0.64 + 1.69 + 0.49$$

$$= 3.02$$

If we divide this number by m, we can get mean squared error(MSE) as

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

$$= \frac{3.02}{5}$$

$$= 0.604$$

3 How does number of observations influence overfitting? Choose the correct answer(s).

(a) In case of fewer observations, it is easy to overfit the data.

(b) In case of fewer observations, it is hard to overfit the data.

(c) In case of more observations, it is easy to overfit the data.

(d) In case of more observations, it is hard to overfit the data.

Both (a) and (d) are correct.

If there are fewer observations, the actual trend of the data may not be explicitly expressed by the observations. And we can let our model hit all the observations easily if we increase the degree of the model. And if there are more observations, the actual trend of the data may be expressed well. And even if we increase the degree of the model, it cannot hit all the observations easily. Therefore, both (a) and (d) are correct.

4 How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

We need to estimate two coefficients. One coefficient is slope w and the other coefficient is intercept b.

5 What is cross validation and how does it work?

Cross validation is a technique to avoid wasting too much training data in validation sets.

Generally, we construct validation set from the training data. For example, we split the training data into two disjoint subsets. One is used to learn the parameters and the other one (the validation set) is used to estimate the test error during or after training, allowing for the hyperparameters to be updated accordingly. Then, the training set is split into complementary subsets. Each model is trained against a different combination of these subsets and validated against the remaining parts. As the figure (Figure 2) showing below, the data set is split into 5 subsets. Each time we choose one subset as the validation set and other four subsets are used to train. Once the model type and hyperparameters have been selected, a final model is trained using these hyperparameters on the full training set, and the test error is measured on the test set.

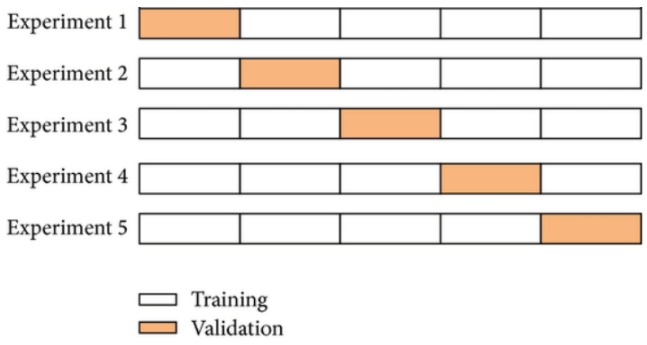


Figure 2: Cross Validation

6 Mathematically show that the softmax function with two classes ($k = 2$) is equivalent to the sigmoid function?

For softmax, we have:

$$\hat{y}_j = p(y = j \mid x; w_j) = \sigma(w_j^T x) = \frac{\exp(w_j^T x)}{\sum_{i=0}^{k-1} \exp(w_i^T x)}$$

Since we only consider $k = 2$ classes, we have:

$$\begin{aligned}
\hat{y}_1 &= \frac{\exp(w_1^T x)}{\exp(w_0^T x) + \exp(w_1^T x)} \\
&= \frac{\frac{\exp(w_1^T x)}{\exp(w_1^T x)}}{\frac{\exp(w_0^T x)}{\exp(w_1^T x)} + \frac{\exp(w_1^T x)}{\exp(w_1^T x)}} \\
&= \frac{1}{\exp\left[-(w_1 - w_0)^T x\right] + 1} \\
&= \frac{1}{1 + \exp\left(-w^T x\right)}
\end{aligned}$$

Similarly, we have

$$\hat{y}_0 = \frac{1}{1 + \exp\left[(w_1 - w_0)^T x\right]} = \frac{1}{1 + \exp\left(w^T x\right)}$$

For binomial, we have:

$$\hat{y} = p(y = 1 \mid x; w) = \sigma(w^T x) = \frac{1}{1 + \exp(-w^T x)}$$

$$1 - \hat{y} = p(y = 0 \mid x; w) = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)} = \frac{\frac{\exp(-w^T x)}{\exp(-w^T x)}}{\frac{1}{\exp(-w^T x)} + \frac{\exp(-w^T x)}{\exp(-w^T x)}} = \frac{1}{1 + \exp(w^T x)}$$

Therefore, the softmax function with two classes ($k = 2$) is equivalent to the sigmoid function.

7 As you know, in binomial logistic regression the cost between the true value y and the predicted value $\hat{y}$ is measured as below:

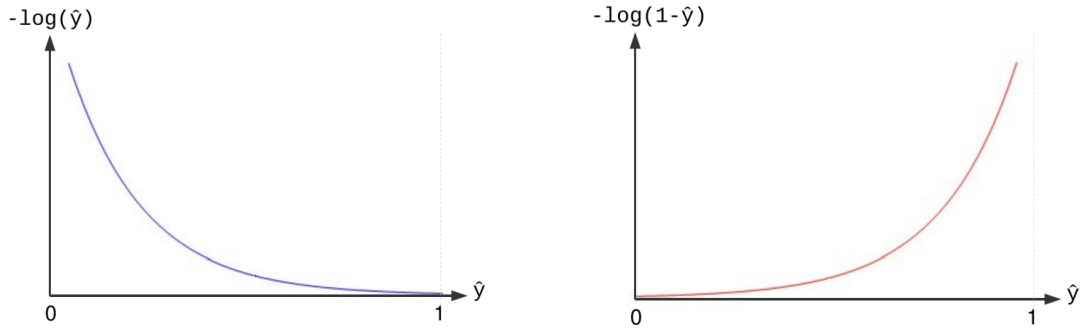
$$\mathbf{cost}(\hat{\mathbf{y}},\ \mathbf{y}) = \begin{cases} -\mathbf{log}(\hat{\mathbf{y}}) & \textbf{if } \mathbf{y = 1} \\ -\mathbf{log}(\mathbf{1} - \hat{\mathbf{y}}) & \textbf{if } \mathbf{y = 0} \end{cases}$$

Explain why -log is a proper function to compute the cost in logistic regression?

The cost function has those property:

- When $\hat{y}$ is close to y, the cost should close to 0;
- When $\hat{y}$ is far from y, the cost should be large enough to represent the deviation value.

We can draw the graph of the functions (Figure 3) like this:



(a) y = 1 (b) y = 0

Figure 3: The picture of cost function

From the figure above, we can find that when $y = 1$, $-\log(\hat{y})$ will be close to 0 if $\hat{y}$ is close to 1 and $-\log(\hat{y})$ will be large if $\hat{y}$ is close to 0. Similarly, when $y = 0$, $-\log(1-\hat{y})$ will be close to 0 if $\hat{y}$ is close to 0 and $-\log(1-\hat{y})$ will be large if $\hat{y}$ is close to 1. Therefore, $-\log$ is a proper function to compute the cost in logistic regression.

8 How are logistic regression cost, cross-entropy, and negative log-likelihood related?

The binomial logistic regression cost for one predicted value is defined as

$$\text{cost}(\hat{y},\ y) = \begin{cases} -\log(\hat{y}) & \text{if } y = 1 \\ -\log(1-\hat{y}) & \text{if } y = 0 \end{cases}$$

It can be written as

$$\text{cost}(\hat{y},\ y) = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

For all the predicted values, the sum of cost can be written as

$$-\sum_{i}^{m}[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})]$$

The likelihood function is

$$L(\theta) = p(X \mid \theta) = \prod_{i}^{m} p(x^{(i)} \mid \theta)$$

The negative log-likelihood is

$$-\log L(\theta) = -\sum_{i=1}^{m} \log p(x^{(i)} \mid \theta)$$

In binomial case, if we consider the value of $\hat{y}^{(i)}$ as the probability, we can have

$$\begin{cases} p(y^{(i)} = 1 \mid x^{(i)}; w) = \hat{y}^{(i)} \\ p(y^{(i)} = 0 \mid x^{(i)}; w) = 1 - \hat{y}^{(i)} \end{cases} \Rightarrow p(y^{(i)} \mid x^{(i)}; w) = (\hat{y}^{(i)})^{y^{(i)}}(1-\hat{y}^{(i)})^{(1-y^{(i)})}$$

So the likelihood is

$$L(\mathrm{w}) = \mathrm{p}(\mathrm{y} \mid \mathrm{x}; \mathrm{w}) = \prod_i^m \mathrm{p}(\mathrm{y}^{(i)} \mid \mathrm{x}^{(i)}; \mathrm{w}) = \prod_i^m (\hat{\mathrm{y}}^{(i)})^{\mathrm{y}^{(i)}} (1 - \hat{\mathrm{y}}^{(i)})^{(1-\mathrm{y}^{(i)})}$$

And the negative log-likelihood is

$$-\log(L(\mathrm{w})) = -\sum_i^m \log \mathrm{p}(\mathrm{y}^{(i)} \mid \mathrm{x}^{(i)}; \mathrm{w}) = -\sum_i^m [\mathrm{y}^{(i)} \log(\hat{\mathrm{y}}^{(i)}) + (1 - \mathrm{y}^{(i)}) \log(1 - \hat{\mathrm{y}}^{(i)})]$$

We can find it is the same as the sum of cost for all the predicted values.

The negative log-likelihood is also called the cross-entropy. The cross-entropy is used to quantify the difference (error) between two probability distributions and it describes how close is the predicted distribution to the true distribution. For the true distribution p and predicted distribution, we have q:

$$H(p, q) = -\sum_j (p_j \log(q_j))$$

In the binomial case, the true probability distribution is $p(\mathrm{y} = 1) = \mathrm{y}$ and $p(\mathrm{y} = 0) = 1 - \mathrm{y}$. The predicted probability distribution is $q(\mathrm{y} = 1) = \hat{\mathrm{y}}$ and $q(\mathrm{y} = 0) = 1 - \hat{\mathrm{y}}$.

Therefore, the cross-entropy of p and q in binomial case is nothing but the logistic cost function.

$$H(p, q) = -\sum_j (p_j \log(q_j)) = -[\mathrm{y} \log(\hat{\mathrm{y}}) + (1 - \mathrm{y}) \log(1 - \hat{\mathrm{y}})] = \mathrm{cost}(\hat{\mathrm{y}},\ \mathrm{y})$$

So we can tell the cross-entropy of p and q is mathematical equivalent to logistic cost function.

Therefore, the logistic regression cost, cross-entropy, and negative log-likelihood coincide with each other under the same condition.

9 Explain how a ROC curve works?

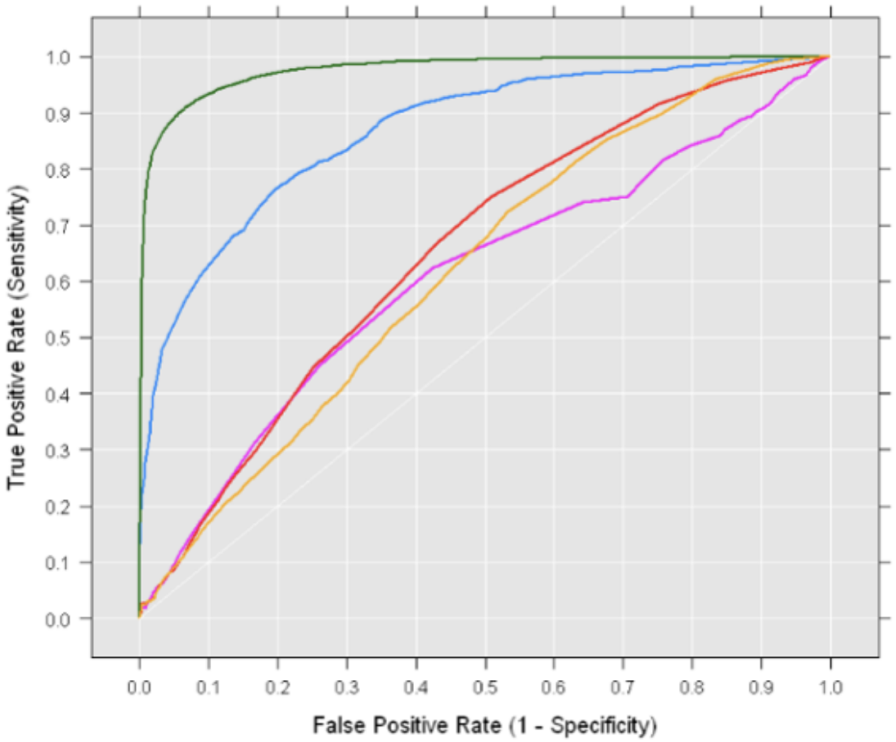


Figure 4: ROC Curve[1]

This picture (Figure 4) shows the ability of the binary classification. The closer the line is to the green line, the better the ability is. As we have defined the True Positive Rate and False Positive Rate:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \ \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

The ROC curve reflects the trade-off between them two. They use different threshold for a model. The higher TPR is, the more FPR the classifier produces. A good classifier moves toward the top-left corner, which means the larger Area under the curve is, the better model we have.

[1]Source: `https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/`