
Machine Learning, 2024 Fall

Quiz 8

1 True or False [30 pts]

1. If we observe that the model's performance on the validation set gets worse as training progresses, we can say that our model is overfitting the training set.

solution: False

2. For a classification problem, the more features there are, the easier it is to distinguish between sample categories.

solution: False

3. In handling time series data, overfitting in a model typically means that its predictions for future data will be more accurate.

solution: False

4. If a model uses various types of regularization (such as L1 and L2 regularization), it can completely prevent overfitting.

solution: False

5. Increasing the coefficient of the regularization term always enhances the model's generalization ability.

solution: False

6. The penalty of L2 regularization has a greater impact on small parameters than on large parameters.

solution: False

2 Validation [30 pts]

In machine learning model evaluation, Leave-One-Out Cross-Validation (LOOCV) and k-Fold Cross-Validation are two commonly used methods. Suppose we evaluate a model using LOOCV and 5-fold Cross-Validation. The computational cost of each training is proportional to the square of the number of training samples. When there are n samples,

1. What is the total computational cost of LOOCV?
2. What is the total computational cost of 5-fold Cross-Validation?

solution: 1. Each training uses $n-1$ samples, and the total number of trainings is n :

$$c_{LOOCV} \propto n \cdot (n-1)^2 \approx n^3$$

2. Each training uses 80% of the samples (i.e., $\frac{4n}{5}$ samples), and the total number of trainings is 5:

$$C_{5-Fold} \propto 5 \cdot \left(\frac{4n}{5}\right)^2 = 5 \cdot \frac{16n^2}{25} = \frac{16n^2}{5}$$

3 Decision Tree [40 pts]

There are many factors that influence whether a graduate student stays up late. We have collected some relevant data

id	A	B	C	stays up late
1	Far	Low	No	No
2	Near	High	Yes	No
3	Near	Low	Yes	Yes
4	Near	Low	Yes	Yes
5	Far	High	Yes	No

- A: The time remaining until the deadline.
- B: The performance of their own method.
- C: Whether they have class at 8 AM tomorrow.

Please based on *information gain*, build your decision tree. You can compute the information gain directly, or give reasons. Only results will get ZERO points. $\log_2(3) \approx 1.58496$

[Refer to lecture 12 page 16](#)