Non-Assessed Exercise

UNIVERSITY OF MANCHESTER
DEPARTMENT OF COMPUTER SCIENCE

DATA70121: Machine Learning and Statistics I

Lecture 8: Model Assessment and Selection (I)

Sample Answers

You are strongly suggested making a serious attempt before seeing sample answers.

Lecture 8: Model Assessment and Selection (I)

Multiple Choice Questions

1. In machine learning, a model is established to gain the best performance on the data used in its learning process. True or False?

   A. True

   B. False

   B

2. A high-bias machine learning model is likely to be underfitting. True or False?

   A. True

   B. False

   A

3. In context of machine learning, *inductive bias* refers to a phenomenon that people always want a learning model that performs the best on the observed data during learning. True or False?

   A. True

   B. False

   B

4. In machine learning, over-fitting means poor generalisation. True or False?

   A. True

   B. False

   A

5. In machine learning, *variance* is the error due to overly simple assumptions in the learning algorithm. True or false?

   A. True

   B. False

   B

6. *Cross-validation* can be used with any machine learning algorithm. True or false?

   A. True
   B. False

   A

7. In statistical learning, *bias-variance trade-off* for a machine learning model refers to:

   A. The trade-off between its training speed and accuracy
   B. The trade-off between the amount of data and the quality
   C. The trade-off between the model's complexity and its performance.
   D. The trade-off between the memory and the time during its training

   C

8. Training a linear regression model leads to high bias. Which of the following actions might help to reduce the bias?

   A. Increase the number of training examples
   B. Add polynomial features
   C. Use a simpler model
   D. Reduce the number of training examples

   B

9. When it comes to bias and variance, which of the following is generally true?

   A. Increasing the model's complexity will decrease both bias and variance.
   B. Increasing the model's complexity will decrease bias and increase variance.
   C. Increasing the model's complexity will increase both bias and variance.
   D. Increasing the model's complexity will increase bias and decrease variance.

   B

10. What is the primary purpose of using *held-out validation* in a machine learning context?

    A. To increase the accuracy of the model
    B. To prevent overfitting by evaluating the model's performance on unseen data
    C. To add more features to the model
    D. To reduce the computational load on the system

    B

11. What is a potential disadvantage of *held-out validation*?

    A. It prevents the model from overfitting.
    B. It uses all data points for training
    C. It may waste of data, as a portion of the dataset is not used for training.
    D. It cannot handle a large dataset.

    C

12. Which of the following statements is *true* about *cross-validation* and *held-out validation*?

    A. Cross-validation is a type of held-out validation.
    B. Held-out validation is a type of cross-validation.
    C. Cross-validation and held-out validation are entirely different methods.
    D. None of the above statements is true.

    B

13. In practice, what is a good value for k in k-fold cross-validation?

    A. 100
    B. 5
    C. 2
    D. 1

    B

14. In which situations, would it be beneficial to use LOOCV?

    A. When computation resources are limited
    B. When the model's performance has high variance
    C. When the dataset is large
    D. When the dataset is small

    D

15. What does *cross-validation* do that *held-out validation* does not?

    A. It provides a single estimate of the model performance.
    B. It provides multiple estimates of the model performance.
    C. It uses all data for training.
    D. It uses all data for testing.

    B

16. In 10-fold cross validation, the training takes 1,000 seconds and the test spends two seconds in each fold. Which of the following options regarding the overall time is *CORRECT*?

    A. 10,002 seconds
    B. 10,012 seconds
    C. 10,020 seconds
    D. 10,200 seconds

    C

17. Which of the followings are the risks of a high variance model?

    A. The model will likely underfit the data.
    B. The model will likely overfit the data.
    C. The model might perform poorly on unseen data.
    D. The model might perform well on unseen data.
    E. The model will likely be too simple.

    B,C

18. Which of the following are *correct* for repeated k-fold cross-validation?

    A. Should be used when the dataset is very large.

    B. Can help in reducing the variability of the performance estimate.

    C. Repeats the process of k-fold cross-validation with different random splits of the dataset.

    D. Always provides a biased performance estimate.

    E. Should be used when computational resources are limited.

B,C

Explanation to Answers

1. The ultimate goal of machine learning is maximising the generalisation.

2. High bias often refers to oversimplification of the model, which does not learn enough from the training data, leading to underfitting.

3. Inductive bias refers to the expectation that the learned model performs the best on unobserved data in test.

4. Overfitting indicates high performance on training data but poor performance on test data.

5. Variance is the error due to overly complex assumptions in the learning algorithm.

6. Cross-validation is a generic technique and can be used with any machine learning algorithm.

7. It highlights the relationship between a model's complexity (high or low) and its performance, in terms of generalization and overfitting.

8. Adding polynomial features makes the model more complex, allowing it to fit the training data more closely and potentially reducing bias.

9. Increasing the model's complexity generally decreases bias but increases variance, leading to a trade-off between the two.

10. The primary purpose of held-out validation is to evaluate the generalization capability of a machine learning model to prevent overfitting.

11. The primary purpose of held-out validation is to evaluate the generalization capability of a machine learning model to prevent overfitting.

12. Held-out validation is essentially a type of cross-validation where the data is split into two parts: a training set and a validation set. However, in cross-validation, the data is divided into k parts, with each part serving as the validation set once.

13. A common choice for k in k-fold cross-validation is 5 or 10. Choosing a small value for k can lead to a larger bias in the error estimation while choosing a large value for k can lead to a larger variance.

14. LOOCV is typically beneficial when the dataset is small, as it makes use of nearly all data for training in each iteration.

15. Cross-validation provides multiple estimates of the model performance, one for each fold or iteration, while held-out validation provides a single estimate.

16. Each fold in 10-fold cross validation will take 1000 seconds for training and 2 seconds for testing hence 1002 seconds totally. So 10 folds will take 10*1002 = 10,020.

17. A high variance model is likely to overfit the data as it captures the noise and specifics of the training data. As a result, it might perform poorly on unseen data as it fails to generalise well.

18. Repeated k-fold cross-validation helps reduce the variability of the performance estimate (**B**) by repeating the process of k-fold cross-validation with different random splits of the dataset (**C**). It does not need to be used specifically with large datasets (**A**), and it is not recommended when computational resources are limited (**E**) due to the increased computation time. It doesn't always provide a biased performance estimate (**D**); the bias depends on other factors such as the choice of 'k' and the nature of the data.