Non-Assessed Exercise

UNIVERSITY OF MANCHESTER
DEPARTMENT OF COMPUTER SCIENCE

DATA70121: Machine Learning and Statistics I

Lecture 9: Model Assessment and Selection (II)

Sample Answers

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

Lecture 9: Model Assessment and Selection (II)

Multiple Choice Questions

1. An analytic method for model assessment and selection in machine learning always work on *training* data without need of *validation* data. True or False?

   A. True

   B. False

   A

2. For model assessment and selection, *Akaike's Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC) are applicable to any learning models. True or False?

   A. True

   B. False

   B

3. The *best subset selection* method always provides the optimal feature subset for a given data set. True or False?

   A. True

   B. False

   A

4. The *forward stepwise selection* (FSS) always yields the same sub-optimal feature subset as obtained by the *backward stepwise selection* (BSS). True or False?

   A. True

   B. False

   B

5. Both FSS and BSS search for the same number of models during feature subset selection. True or False?

   A. True

   B. False

   A

6. *Adjusted R-squared* always increases as we add features to a linear regression model. True or False?

   A. True

   B. False

   B

7. In statistical learning, *AIC* is based on which principle?

   A. Maximum likelihood estimation

   B. Bayesian inference

   C. Residual sum of squares

   D. Occam's razor

   A

8. Which of the following is a correct interpretation of *BIC*?

   A. BIC estimates the relative amount of information lost by a given model.

   B. BIC estimates the complexity of a model.

   C. BIC directly estimates the out-of-sample prediction error.

   D. BIC is an approximation to the log of the Bayes factor.

   D

9. In the context of *AIC* and *BIC*, what does "model complexity" refer to?

   A. The number of parameters in the model

   B. The computational complexity of fitting the model

   C. The amount of data required to fit the model

   D. The complexity of the hypotheses that the model can express

   A

10. In context of *AIC* and *BIC* used in machine learning, which statement below is *true*?

    A. AIC and BIC always select the same model.
    B. Neither AIC nor BIC can be used for model selection.
    C. AIC tends to select larger models than BIC.
    D. BIC tends to select larger models than AIC.

    C

11. Both *AIC* and *BIC* aim to strike a balance between model complexity and:

    A. Model interpretability
    B. Model size
    C. Training time
    D. Prediction accuracy

    D

12. Which of the following criteria would you prefer if you prioritise a better *out-of-sample* prediction accuracy?

    A. AIC
    B. BIC
    C. Either could be preferred depending on the situation.
    D. Neither of them

    C

13. Which feature subset selection method starts with a full model and sequentially removes predictors based on *AIC* or *BIC*?

    A. Best subset selection
    B. Forward stepwise selection (FSS)
    C. Backward stepwise selection (BSS)
    D. All of the above methods

    C

14. In context of *feature subset selection* for a regression task, which of the following statements is *CORRECT*?

    A. The best subset selection always yields the smallest RSS on training data.

    B. The FSS always leads to the smallest RSS on training data.

    C. The BSS always leads to the smallest RSS on training data.

    D. None of the above are correct.

    A

15. You apply the BSS method to a dataset with d features in total. In order to find out an sub-optimal feature subset, which of the following statements is *CORRECT*?

    A. $d(d+1)$ models have to be compared.

    B. $d(d+1)+1$ models have to be compared.

    C. $d(d+1)/2$ models have to be compared.

    D. $d(d+1)/2+1$ models have to be compared.

    D

16. AIC is used to select the best regression model from three candidate models, A, B and C with 30, 20 and 10 parameters, trained on the same dataset with 100 examples. The RSS for models A, B and C are 200, 220 and 230, respectively. Which of three models is the *best* model?

    A. Model A.

    B. Model B.

    C. Model C.

    D. Cannot be decided.

    C

17. What kind of models do *AIC* and *BIC* aim to find?

    A. The simplest models.
    B. The most complex models.
    C. Models that minimise the prediction error.
    D. Models that balance goodness of fit and model complexity.

C,D

18. Which of the following reasons justify the use of subset selection methods in statistical learning?

    A. Improvement of model interpretability.
    B. Reduction of computational costs.
    C. Reduction of overfitting.
    D. Enhancement of prediction accuracy.
    E. Removal of multicollinearity.

A,C,D,E

Explanation to Answers

1. Such a methodology uses model complexity for estimate generalisation performance based on only training data.

2. Neither of two criteria works on non-probabilistic model.

3. This is an exhaustive search method that works on ALL possible subsets.

4. Due to the different search order, it is possible for two methods to end up with different sub-optimal subsets.

5. Both search for $p(p+1)/2+1$ models for feature subset selection, where p is the number of features.

6. Unlike regular R-squared, *adjusted R-squared* accounts for model complexity (i.e., number of predictors) and can decrease if the addition of a predictor does not significantly improve the model.

7. *AIC* is derived from the principle of maximum likelihood estimation and provides a measure of how well a model fits the data.

8. *BIC* is intended to approximate the log of the Bayes factor, a quantity used in Bayesian statistics to compare models.

9. In the context of *AIC* and *BIC*, "model complexity" refers to the number of parameters in the model; the more parameters the higher flexibility in a model.

10. *AIC* tends to select larger models than *BIC* because it has a smaller penalty for model complexity. *BIC* includes a more substantial penalty term for models with more parameters, particularly as sample size grows.

11. Both *AIC* and *BIC* aim to penalise overly complex models to avoid overfitting and prioritise models that achieve good prediction accuracy simultaneously.

12. Both *AIC* and *BIC* are used for model selection with the goal of optimizing out-of-sample prediction. Depending on the specific scenario and the relative importance of bias (*underfitting*) and variance (*overfitting*), either *AIC* or *BIC* could be preferred.

13. *Backward stepwise selection* begins with a full model containing all predictors and iteratively removes one predictor at a time based on the *AIC* or *BIC*.

14. *Best subset selection* is an exhaustive search method that compares ALL possible combination of features so it can find the least RSS on a given training data set.

15. Total number of models to be compared is $[d+(d-1)+...+2+1]+1 = d(d+1)/2+1$.

16. Calculate with $AIC = (RSS+2p)/n$. Thus, we have $AIC(A) = 2.6$, $AIC(B) = 2.6$ and $AIC(C) = 2.5$ (least).

17. *AIC* and *BIC* aim to find models that minimise the prediction error and models that balance goodness of fit and model complexity. They do not necessarily aim to find the simplest or the most complex models.

18. Four correct answers are explained as follows:

    - **A**: Subset selection methods can improve model interpretability by removing unnecessary predictors, thereby simplifying the model.
    - **C**: They can also reduce overfitting, which occurs when a model is too complex and performs well on the training data but poorly on new data.
    - **D**: Enhancing prediction accuracy is another benefit, as removing irrelevant predictors can sometimes improve predictions.
    - **E**: Subset selection also helps to remove multicollinearity, where two or more predictors are highly correlated with each other, which can interfere with the interpretability of model coefficients.