Revision questions for Chapter 10

Last updated: December 5, 2022

If you are asked to define some notion, you should explain carefully all notation (if any) that you use in your definition. Answers to some questions are given in blue. All other answers can be found in the module notes (lecture slides or lab worksheets) provided on the module's Moodle page.

1. Describe (using an itemized list or pseudocode) the process of building a decision tree. Chapter 10, slides 11–12.

2. When is a node in a decision tree called *pure*? Make sure to cover both the case of classification and the case of regression. Chapter 10, slide 12.

3. Explain how a prediction is made given a decision tree. Chapter 10, slide 13.

4. Make sure you can answer the question on slides 14–15 of Chapter 10.

5. What is meant by pre-pruning and post-pruning of decision trees? Chapter 10, slide 16.

6. Give three possible criteria for pre-pruning. Chapter 10, slide 16.

7. Give two advantages and a disadvantage of decision trees as compared with other methods of machine learning. Chapter 10, slide 18.

8. What is meant by ensemble methods in machine learning? Chapter 10, slide 19.

9. Name and briefly describe two different ensemble methods. Chapter 10, slides 19–20 and 31.

10. Give a description of random forests as an itemized list or pseudo-code. Chapter 10, slides 21–23 and 25.

11. What is a bootstrapped dataset? How are bootstrapped datasets used in the method of bagging? Chapter 10, slides 21–22.

12. Is the list `[0, 12, 10]` a bootstrapped version of the list `[0, 11, 10]`? No, since `12` is not an element of the list `[0, 11, 10]`. See Chapter 10, slide 22.

13. Is the list `[0, 0, 0]` a bootstrapped version of the list `[0, 11, 10]`? Yes. See Chapter 10, slide 22.

14. Is the list `[0, 11, 10, 11]` a bootstrapped version of the list `[0, 11, 10]`? No, since the two lists are of different sizes. See Chapter 10, slide 22.

15. Is the list `[11, 0, 0]` a bootstrapped version of the list `[0, 11, 11]`? Yes. See Chapter 10, slide 22.

16. List all bootstrapped versions of the dataset $[A, B, C]$, regarding datasets that differ only in the order of their elements as the same dataset. There are 10 such datasets: $[A, A, A]$, $[B, B, B]$, $[C, C, C]$, $[A, A, B]$, $[A, B, B]$, $[B, C, C]$, $[B, B, C]$, $[A, C, C]$, $[A, A, C]$, and $[A, B, C]$.

17. What are the two mechanisms ensuring that the trees in a random forest are sufficiently different? Chapter 10, slide 23.

18. Describe briefly the *soft voting* strategy used in random forests. Chapter 10, slide 25.

19. List three strengths and three weaknesses of random forests as compared with other machine-learning algorithms. Chapter 10, slides 28–29.

20. List the most important parameters of random forests and briefly describe their role. Chapter 10, slide 30.

21. How would you choose the parameter `n_estimators` in the method of random forests? Chapter 10, slide 30.

22. Describe briefly the method of gradient boosting. Chapter 10, slide 31.

23. List two strengths and two weaknesses of gradient boosting machines as compared with other machine-learning algorithms. Chapter 10, slide 32.

24. List the most important parameters of gradient boosting machines and briefly describe their role. Chapter 10, slide 33.

25. Is it possible to overfit random forests by making `n_estimators` too large? No. See Chapter 10, slides 30 and 33.

26. Is it possible to overfit gradient boosting machines by making `n_estimators` too large? Yes. See Chapter 10, slides 30 and 33.

27. How would you build an inductive conformal predictor on top of a Bayesian algorithm? Chapter 10, slide 38.

28. What is the main assumption of the Naive Bayes algorithm? Chapter 10, slide 39.

29. Briefly describe the idea behind the Naive Bayes algorithm. Chapter 10, slide 39.

30. List two strengths and a weakness of the Naive Bayes algorithm as compared with other machine-learning algorithms. Chapter 10, slide 41.

31. Briefly describe the method of logistic regression. Chapter 10, slides 43–45.

32. Briefly describe the method of linear discriminant analysis (LDA). Chapter 10, slide 46.

33. Briefly describe the method of quadratic discriminant analysis (QDA). Chapter 10, slide 47.

34. What is the difference between LDA and QDA? The assumption of equal covariance matrices in the former.

35. Explain briefly how the performance of LDA, QDA, and logistic regression depends on the size of the training set. Chapter 10, slide 50.

Similar lists of questions will be produced for all chapters of the module to help students in revision. There is no guarantee that the actual exam questions will be in this list, or that they will be in any way similar.