Revision questions for Chapter 10

Last updated: November 25, 2022

If you are asked to define some notion, you should explain carefully all notation (if any) that you use in your definition.

1. Describe (using an itemized list or pseudocode) the process of building a decision tree.
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.
3. Explain how a prediction is made given a decision tree.
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?
6. Give three possible criteria for pre-pruning.
7. Give two advantages and a disadvantage of decision trees as compared with other methods of machine learning.
8. What is meant by ensemble methods in machine learning?
9. Name and briefly describe two different ensemble methods.
10. Give a description of random forests as an itemized list or pseudo-code.
11. What is a bootstrapped dataset? How are bootstrapped datasets used in the method of bagging?
12. Is the list `[0, 12, 10]` a bootstrapped version of the list `[0, 11, 10]`?
13. Is the list `[0, 0, 0]` a bootstrapped version of the list `[0, 11, 10]`?
14. Is the list `[0, 11, 10, 11]` a bootstrapped version of the list `[0, 11, 10]`?
15. Is the list `[11, 0, 0]` a bootstrapped version of the list `[0, 11, 11]`?
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.
17. What are the two mechanisms ensuring that the trees in a random forest are sufficiently different?
18. Describe briefly the *soft voting* strategy used in random forests.
19. List three strengths and three weaknesses of random forests as compared with other machine-learning algorithms.

20. List the most important parameters of random forests and briefly describe their role.

21. How would you choose the parameter `n_estimators` in the method of random forests?

22. Describe briefly the method of gradient boosting.

23. List two strengths and two weaknesses of gradient boosting machines as compared with other machine-learning algorithms.

24. List the most important parameters of gradient boosting machines and briefly describe their role.

25. Is it possible to overfit random forests by making `n_estimators` too large?

26. Is it possible to overfit gradient boosting machines by making `n_estimators` too large?

27. How would you build an inductive conformal predictor on top of a Bayesian algorithm?

28. What is the main assumption of the Naive Bayes algorithm?

29. Briefly describe the idea behind the Naive Bayes algorithm.

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

31. Briefly describe the method of logistic regression.

32. Briefly describe the method of linear discriminant analysis (LDA).

33. Briefly describe the method of quadratic discriminant analysis (QDA).

34. What is the difference between LDA and QDA?

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

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.