Revision questions for Chapter 4

Last updated: January 28, 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 course notes (lecture slides or lab worksheets) provided on the course's Moodle page, in which case precise references are given.

1. What is meant by *overfitting* in machine learning? Chapter 4, slides 4–6. In words: *overfitting* is where the model complexity is above the "sweet spot" (achieving the optimal accuracy on unseen data).
2. What is meant by *underfitting* in machine learning? Chapter 4, slide 6. In words: *underfitting* is where the model complexity is below the "sweet spot".
3. Define the *training accuracy* in machine learning. Chapter 4, slide 7.
4. Define the *generalization accuracy* in machine learning. Chapter 4, slide 7.
5. Compare and contrasts the behaviour of the training and generalization accuracy as functions of model complexity. Chapter 4, slide 7.
6. How does the optimal model complexity depend on the size of the dataset? Chapter 4, slide 8.
7. What is meant by the *decision boundary* in machine learning? Chapter 4, slide 14.
8. Define *classifiers*, *regressors*, and *estimators* in the context of `scikit-learn`. Chapter 4, slide 20.
9. Give two strengths and two weaknesses of the K Nearest Neighbours algorithm. Chapter 4, slide 22.
10. Define the *learning curve* of a classifier. Chapter 4, slide 25.
11. Define the *learning curve* of a regressor. Chapter 4, slide 25 (implicitly) and Lab Worksheet 4.
12. Describe the method of *cross-validation* for evaluating generalization performance. Make sure to cover both the case of classification and the case of regression. Chapter 4, slides 26–28.
13. Explain what the following sequence of commands in `scikit-learn` is doing (assuming that all relevant modules have been imported):

```
iris = load_iris()
knn = KNeighborsClassifier(n_neighbors=3)
np.mean(cross_val_score(knn, iris.data, iris.target, cv=5))
```

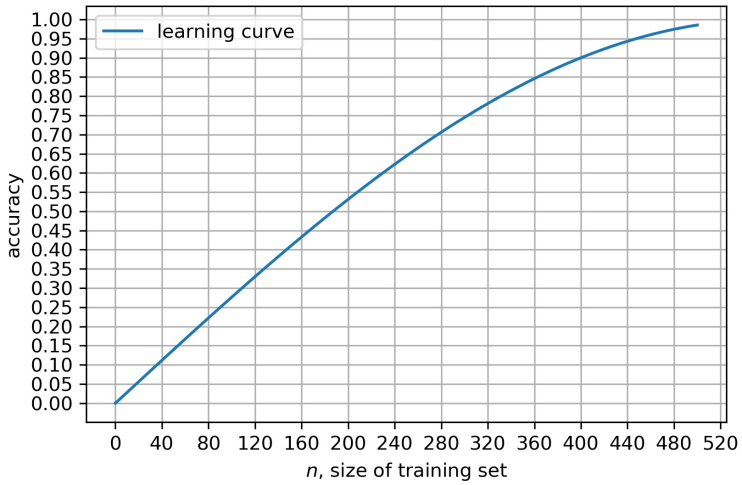
Make sure to explain the role of options such as `cv=5`. Chapter 4, slides 26–28.

14. Explain why the cross-validation procedure leads to a downward bias in the estimate of the generalization performance. Chapter 4, slide 31.

15. Describe the method of *leave-one-out cross-validation* for evaluating generalization performance. Chapter 4, slide 32.

16. How does the bias of the K-fold cross-validation procedure depend on the number K of folds? Explain briefly why. Chapter 4, slides 31–32.

17. Estimate the downward bias of 10-fold cross-validation for a training set of size $n = 400$ and this learning curve:



- The accuracy for the full training set is about 0.90.
- The accuracy for training sets of size $\frac{K-1}{K}n = 360$ is about 0.85.
- Therefore, the downward bias is about $0.90 - 0.85 = 0.05$.

18. Compare and contrast cross-validation and conformal prediction as answers to the question "How confident can we be in our prediction?" Chapter 4, slide 33.

Similar lists of questions will be produced for all chapters of the course 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.