Revision questions for Chapter 6

Last updated: January 28, 2022

The question marked by (*) is more difficult. 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 a precise reference is given. The sign "=" is used for both precise and approximate equalities (feel free to do so when answering exam questions).

1. What is meant by data normalization in machine learning? (Remember that in this course "normalization" is understood in the wide sense and includes the transformations perfomed by `Normalizer`, `StandardScaler`, etc., in `scikit-learn`.)

2. Why is normalization of features not essential for the method of Least Squares? Chapter 6, slide 3.

3. Why is normalization of features essential for Ridge Regression and the Lasso? Chapter 6, slide 3.

4. Briefly describe the class `StandardScaler` in `scikit-learn`, paying particular attention to its `fit` and `transform` methods. Chapter 6, slides 6 and 23.

5. Briefly describe the class `RobustScaler` in `scikit-learn`, paying particular attention to its `fit` and `transform` methods. Chapter 6, slides 7 and 24–25.

6. Briefly describe the class `MinMaxScaler` in `scikit-learn`, paying particular attention to its `fit` and `transform` methods. Chapter 6, slide 7.

7. Briefly describe the class `Normalizer` in `scikit-learn`, paying particular attention to its `fit` and `transform` methods. Chapter 6, slides 8 and 26.

8. Give an example of a dataset for which the use of the class `Normalizer` has a better justification than the use of classes perfoming normalization of features (such as `StandardScaler`). Handwritten digits (Chapter 6, slide 8).

9. Consider the following training set:

feature 1	feature 2	label
−3	2	male
0	5	female
3	8	male
0	8	male

What is its normalized version, in the sense of `MinMaxScaler`? Apply the same transformation to the test set

feature 1	feature 2
1	-1
0	4
2	5

This is similar to Chapter 6, slides 9–11 and 13–14. The normalized version is:

feature 1	feature 2	label
0	0	male
0.5	0.5	female
1	1	male
0.5	1	male

Applying the same transformation to the test set, we obtain:

feature 1	feature 2
0.667	-0.5
0.5	0.333
0.833	0.5

10. For the training set

feature 1	feature 2	label
-10	0	1.6
10	2	2.8

find its normalized version in the sense of `StandardScaler`. Apply the same transformation (emulating the `transform` method) to the test set

feature 1	feature 2
-20	-2
10	4
0	0

The mean of feature 1 is 0 and the mean of feature 2 is 1. The standard deviation of feature 1 is 10 and the standard deviation of feature 2 is 1. The transformation is:

- Divide feature 1 by 10.
- Subtract 1 from feature 2.

The normalized version of the training set is:

feature 1	feature 2	label
-1	-1	1.6
1	1	2.8

Applying the same transformation to the test set, we obtain:

feature 1	feature 2
-2	-3
1	3
0	-1

11. Consider the following training set:

feature 1	feature 2	label
-3	4	male
4	3	female
4	4	male

What is its normalized version, in the sense of `Normalizer`? Apply the same transformation to the test set

feature 1	feature 2
-4	3
3	-3

The normalized version of the training set is:

feature 1	feature 2	label
-0.6	0.8	male
0.8	0.6	female
0.707	0.707	male

The normalized version of the test set is:

feature 1	feature 2
-0.8	0.6
0.707	-0.707

12. What is meant by data snooping in machine learning? Chapter 6, slide 16.

13. What is wrong with the following code for data normalization?

```
X = MinMaxScaler().fit_transform(boston.data)
X_train, X_test, y_train, y_test = train_test_split(X,
  boston.target)
```

Correct the code. Chapter 6, slide 18. One way to correct the code is:

```
X_train, X_test, y_train, y_test = train_test_split(boston.data,
  boston.target)
scaler = MinMaxScaler().fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

14. Discuss disadvantages of normalizing the training and test sets separately when using classes `StandardScaler`, `RobustScaler`, and `MinMaxScaler`. Chapter 6, slides 19–22.

15. Is it admissible to normalize training and test set separately when using the class `Normalizer`? Explain briefly why or why not.

16. What is wrong with the following code for data normalization?

```
X_train, X_test, y_train, y_test = train_test_split(X,
  boston.target)
X_train = MinMaxScaler().fit_transform(X_train)
X_test = MinMaxScaler().fit_transform(X_test)
```

Correct the code. This is another way of asking Question 14. See Chapter 6, slides 19–22, and the code in the answer to Question 13.

17. Explain the use of a validation set for parameter selection in machine learning. Chapter 6, slides 29–30.

18. What are disadvantages of the use of the test set for parameter selection (i.e., of choosing the parameters that give the best results on the test set)? Chapter 6, slide 29. In practice, the test set is not available. And even when it is, this would amount to data snooping.

19. Explain the use of cross-validation for parameter selection in machine learning. Chapter 6, slide 31.

20. How would you perform parameter selection using grid search in a hierarchical manner to improve its computational efficiency? Chapter 6, slide 32. You start with a small and course grid and then move to a finer one (or finer ones).

21. Suppose the result of grid search using cross-validation to select parameters `C` and `gamma` is:

		gamma				
		0.5	1.0	1.5	2.0	2.5
C	0.5	0.92	0.92	0.92	0.92	0.92
	1.0	0.92	0.92	0.92	0.92	0.92
	1.5	0.92	0.92	0.92	0.92	0.92
	2.0	0.92	0.92	0.92	0.92	0.92
	2.5	0.92	0.92	0.92	0.92	0.92

(The entries in the table are the accuracy of the algorithm for different values of the parameters.) How suitable was this grid for selecting the optimal values of the two parameters? Explain briefly why. See Chapter 6, slide 33. The grid covers too small a range of (`C`, `gamma`). It would be better to use a log scale for the parameters (such as using the values 0.01, 0.1, 1, 10, and 100 for both parameters, in place of 0.5, 1.0, 1.5, 2.0, and 2.5).

22. Answer Question 21 for the following grid for parameters `A` and `B`:

		B				
		0.01	0.1	1	10	100
A	0.01	0.59	0.66	0.74	0.73	0.77
	0.1	0.68	0.70	0.73	0.80	0.85
	1	0.68	0.72	0.77	0.81	0.84
	10	0.69	0.77	0.80	0.85	0.89
	100	0.72	0.85	0.88	0.91	0.92

Chapter 6, slide 33. The results keep improving as we increase the values of the two parameters, so there is a good chance they can be improve further as we keep increasing them beyond 100.

23. Answer Question 21 for the following grid for parameters `A` and `B`:

		B				
		0.01	0.1	1	10	100
A	0.01	0.18	0.39	0.64	0.48	0.21
	0.1	0.36	0.58	0.80	0.61	0.43
	1	0.63	0.82	0.95	0.83	0.57
	10	0.41	0.60	0.81	0.63	0.44
	100	0.23	0.44	0.63	0.47	0.31

Chapter 6, slide 33. It looks like the grid covers the optimal values of (`A`, `B`), but it is too crude; now we should try a finer grid around $(\texttt{A}, \texttt{B}) = (1, 1)$.

24. List three desiderata for the method of inductive conformal prediction. Which of them are satisfied automatically? Chapter 6, slide 36.

25. Briefly explain why conformal prediction is not feasible in combination with feature normalization and parameter selection. Chapter 6, slide 37.

26. Compare and contrast conformal prediction and inductive conformal prediction. Chapter 6, slides 38 and 39.

27. Compare and contrast conformal prediction and cross-conformal prediction. Chapter 6, slide 38.

28. Compare and contrast inductive conformal prediction and cross-conformal prediction. Chapter 6, slide 38.

29. What is an *inductive conformity measure*? Define the inductive conformal predictor based on a given inductive conformity measure. Chapter 6, slides 40–41.

30. What is an *inductive nonconformity measure*? Define the inductive conformal predictor based on a given inductive nonconformity measure. Chapter 6, slides 40–42.

31. Give three examples of inductive nonconformity measures. Chapter 6, slide 43 (using Nearest Neighbour). Assignment 2 (the inductive conformity measure based on the Lasso). Alternatively, use Ridge Regression instead of the Lasso.

32. Give two examples of inductive conformity measures. Modify the inductive conformity measure on slide 43 of Chapter 6 as follows: take the distance to the nearest neighbour (in the training set proper) of a different class as the conformity measure. Put the minus sign in front of the inductive conformity measure based on the Lasso and used in Assignment 2.

33. In the context of inductive conformal prediction, what is the minimal possible p-value for a training set proper of size $n - m$ and calibration set of size m? The answer is $1/(m + 1)$. This is analogous to Question 5 for Chapter 3.

34. Take the distance to the nearest neighbour (in the training set proper) of a different class as the conformity measure: $A(\zeta, (x, y))$ is the smallest distance from x to x' such that $(x', y') \in \zeta$ for some $y' \neq y$.

    - The training set proper is: -2 and -1 are labelled A, 2 is labelled B, and 6 is labelled C.
    - The calibration set is: -2 and 0 are labelled A, 3 is labelled B, and 5 and 7 are labelled C.
    - The test sample is 9.

    Compute the three p-values for the test sample. Summarize them in terms of the point prediction, confidence, and credibility. Follow Chapter 6, slides 43–45. First assume the label of 9 is A.

Calibration/test sample	Label	Conformity score
-2	A	4
0	A	2
3	B	3
5	C	3
7	C	5
9	A (?)	3

    The rank of the test sample is 4, and the p-value is $4/6 = 0.667$. Now assume that the label of 9 is B. Since we are dealing with an inductive

conformal predictor, we can copy the rows of the previous table corresponding to the calibration samples, and only need to update the last row (the one corresponding to the test sample).

Calibration/test sample	Label	Conformity score
−2	A	4
0	A	2
3	B	3
5	C	3
7	C	5
9	B (?)	3

The rank of the test sample is again 4, and the p-value is again $4/6 = 0.667$. Finally assume that the label of 9 is C. Copy the rows of the previous table corresponding to the calibration samples, and update the last row.

Calibration/test sample	Label	Conformity score
−2	A	4
0	A	2
3	B	3
5	C	3
7	C	5
9	C (?)	7

The test sample is the least strange, and the p-value is $6/6 = 1$. To summarize, the point prediction is C, the confidence is 0.333, and the credibility is 1. (The confidence is low since this is not a good conformity measure in this context.)

35. In the case of regression, a simple inductive nonconformity measure is

$$A(\zeta, (x, y)) = |y - \hat{y}|,$$

where $\hat{y}$ is the prediction for the true label y computed from ζ as training set. Why is this inductive nonconformity measure sometimes regarded to be insufficiently flexible? How would you improve its flexibility? Chapter 6, slide 46. The inductive conformity measure A leads to prediction intervals of a constant width (see slide 48). A more flexible inductive nonconformity measure is

$$A(\zeta, (x, y)) = |y - \hat{y}| / \sigma,$$

where $\sigma > 0$ is an estimate of the accuracy of $\hat{y}$ computed from ζ as training set.

36. Give a pseudo-code (or Python code) for the inductive conformal predictor based on the inductive conformity measure

$$A(\zeta, (x, y)) = |y - \hat{y}|$$

of the previous question. Chapter 6, slides 47–48.

37. (*) Prove that the code in Question 36 is correct. Chapter 6, slides 49–50.

38. Make sure you can do the exercise on slide 51 of Chapter 6. Chapter 6, slides 52–53.

39. Describe the `scikit-learn` class `GridSearchCV` paying particular attention to its methods `fit`, `predict`, and `score`. See Lab Worksheet 6, Section 2.

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.