Revision questions for Chapter 2

Last updated: October 27, 2022

The questions marked by (*) are more difficult (there is one such question in this revision sheet). 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. The sign "=" is used for both precise and approximate equalities (feel free to do so when answering exam questions).

1. Give two examples of practical problems of supervised machine learning and identify *samples*, *labels*, and *labelled samples* for them. Chapter 2, slides 3–5.
2. What is meant by a classification problem in machine learning? When is a classification problem called binary? When is it called multi-class? Chapter 2, slide 5.
3. What is meant by a regression problem in machine learning? Chapter 2, slide 5.
4. Give two practical examples of classification problems. Chapter 2, slide 5.
5. Give two practical examples of binary classification problems. Chapter 2, slide 5; male vs female.
6. Give two practical examples of multi-class classification problems. Chapter 2, slide 5; classifying web pages into politics, economics, and sport.
7. Give two practical examples of regression problems. Chapter 2, slide 5; predicting tomorrow's temperature.
8. What is meant by a *feature* in machine learning? What is the difference between discrete and continuous features? Chapter 2, slide 6.
9. Define the batch learning protocol. Chapter 2, slide 7.
10. Define the online learning protocol. Chapter 2, slide 8.
11. Give two practical examples of batch learning problems. Chapter 2, slide 5.
12. Give two practical examples of online learning problems. Predicting the weather. Predicting stock prices.
13. What is meant by exploration and exploitation in machine learning? Chapter 2, slide 9.

14. Give the definition of induction in machine learning. Chapter 2, slides 10–11.

15. Give the definition of transduction in machine learning. Chapter 2, slides 10–11.

16. What is the IID assumption in machine learning? Chapter 2, slide 13.

17. What is unsupervised learning? Chapter 2, slide 14.

18. Give two practical examples of unsupervised learning. Company clustering its customers as a first stage of a marketing campaign. Outlier detection (such as recording mistakes).

19. Define the Euclidean distance between two samples. Chapter 2, slide 20.

20. Define the Nearest Neighbour algorithm. Chapter 2, slide 24.

21. Write Python code (or pseudocode) for the Nearest Neighbour algorithm. Chapter 2, slide 24.

22. What is the difference between the Nearest Neighbour algorithm for classification and the Nearest Neighbour algorithm for regression? None.

23. Describe the K Nearest Neighbours algorithm for classification. Chapter 2, slide 29.

24. What is meant by positive and negative samples in machine learning? Chapter 2, slide 25.

25. Consider the training set

    - positive: $(1, 2, 3, 0)$
    - positive: $(1, 4, 2, 3)$
    - negative: $(-2, 3, -4, 3)$
    - negative: $(-1, 1, -3, 2)$

    and a test sample $(0, 1, 0, 1)$. Calculate its classification using the K Nearest Neighbours algorithm with Euclidean distance, first for $K = 1$ and then for $K = 3$.

    Answer: negative for $K = 1$, positive for $K = 3$.

    These are the detailed calculations:

    - The squared distance between the first sample, $(1, 2, 3, 0)$, and the test sample is:

      $$(1-0)^2 + (2-1)^2 + (3-0)^2 + (0-1)^2 = 1 + 1 + 9 + 1 = 12.$$

- The squared distance between the second sample, $(1, 4, 2, 3)$, and the test sample is:

$$(1-0)^2 + (4-1)^2 + (2-0)^2 + (3-1)^2 = 1 + 9 + 4 + 4 = 18.$$

- The squared distance between the third sample, $(-2, 3, -4, 3)$, and the test sample is:

$$(-2-0)^2 + (3-1)^2 + (-4-0)^2 + (3-1)^2 = 4 + 4 + 16 + 4 = 28.$$

- The squared distance between the fourth sample, $(-1, 1, -3, 2)$, and the test sample is:

$$(-1-0)^2 + (1-1)^2 + (-3-0)^2 + (2-1)^2 = 1 + 0 + 9 + 1 = 11.$$

For $K = 1$, the prediction is "negative" since the nearest neighbour (the fourth sample) is negative. For $K = 3$, the prediction is "positive" since among the 3 nearest neighbours (the first, second, and fourth samples) "positive" are in majority.

26. Describe the K Nearest Neighbours algorithm for regression. Chapter 2, slide 29.

27. Consider the following regression problem. The training set is as follows:

    - sample $(1, 2, 3, 0)$ is labelled as -2,
    - sample $(1, 4, 2, 3)$ is labelled as 2,
    - sample $(-2, 3, -4, 3)$ is labelled as 0,
    - sample $(-1, 1, -3, 2)$ is labelled as 1.

    The test sample is $(0, 1, 0, 1)$. Calculate its predicted label using the K Nearest Neighbours algorithm with Euclidean distance, first for $K = 1$ and then for $K = 3$.

    Answer: 1 for $K = 1$, and $1/3$ for $K = 3$.

    The distances are calculated in the answer to Question 25; the rest of the calculations are as follows. When $K = 1$, the predicted label is the label of the nearest (fourth) sample, i.e., 1. When $K = 3$, the predicted label is the mean label for the 3 nearest samples (first, second, and fourth samples), i.e.,

$$(-2 + 2 + 1)/3 = 1/3.$$

28. (*) Consider the following training set: crosses at -4, -3, -2, and -1; noughts at 2, 4, 6, and 8. Sometimes the prediction obtained using the 1 Nearest Neighbour algorithm is different from the prediction obtained using the 3 Nearest Neighbours algorithm. For the given training set

above, find the maximal interval for the new test point where the two predictions would be different.

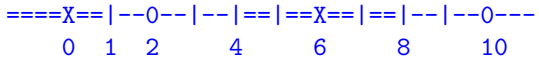
Answer: the interval $(1/2, 1)$ (or $[1/2, 1]$; it does not matter whether you include the end-points).

This interval contains **all** test points where the two predictions are different. The word "maximal" is here to disallow intervals such as $(3/4, 1)$, which are subsets of the maximal interval $(1/2, 1)$.

29. Consider the training set

sample	classification
0	$+1$
2	-1
6	$+1$
10	-1

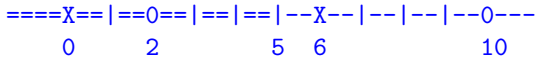
(a) Draw the training samples on the straight line and highlight the regions of test samples that are classified as positive by the Nearest Neighbour algorithm.



The points 1, 4, and 8 demarcate the regions.

Hint: find the predictions for the test samples $-0.5, 0, 0.5, \ldots, 10, 10.5$ and try to argue what happens for the other test samples.

(b) Highlight the regions of test samples that are classified as negative by the 3 Nearest Neighbours algorithm.



The demarcating point is 5. To the left of 5 the training sample 0 is closer than the training sample 10. So to the left of 5 the three nearest neighbours are 0, 2, and 6, while to the right of 5 the three nearest neighbours are 2, 6, and 10.

30. Would you classify the K Nearest Neighbours algorithm as induction or transduction? Explain briefly why. Chapter 2, slide 28.

31. What is the computational complexity of the K Nearest Neighbours algorithm? Give a brief explanation. Chapter 2, slide 31.

32. Give three examples of distances that can be used in the K Nearest Neighbours algorithm. Chapter 2, slide 33. For example, Euclidean, tangent, and kernel.

33. Give three examples of practical application of the K Nearest Neighbours algorithm. Chapter 2, slide 34. Recognizing satellite images. Recognizing patterns in electrocardiography. Text classification (e.g., answering questions such as "is this Web article about politics?").

34. Give two advantages and two disadvantages of the K Nearest Neighbours algorithm. HINT: this is covered in Chapter 4. Chapter 4, slide 22.

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.