UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Revision Practice
COMP24112 Machine Learning

May 2023

Time: Recommended to practice after revising both the lecture content and in-class tutorial questions.

Marking Scheme Included

Do not publish

This practice contains similar types of questions that will appear in your final exam.

The use of electronic calculators is permitted provided they are not programmable and do not store text.

---

1. Which of the following tasks is unsupervised?

   A. To train a classifier to recognise faces.

   B. To train a regression model to predict stock market.

   C. To map a group of 10-dimensional data points to a 2-dimensional space to maximise data variance along each dimension.

   D. None of the above.

   C

2. Consider the following table summarising the testing results for a two-class classification task:

		Actual Class	
		Class A	Class B
Predicted Class	Class A	70	15
	Class B	10	5

   What is the sensitivity score for classifying the class B?

   A. 10.0 %

   B. 15.0 %

   C. 25.0%

   D. 75.0 %

   C

3. Which of the following statements is NOT true?

   A. Training accuracy minus test accuracy provides an estimate of the degree of overfitting.

   B. Training error does not provide a good estimate of the true error.

   C. The mean value of a set of test error rates taken over different testing sets is a good estimate of the true error.

   D. The mean value of a set of training error rates taken over different training sets is a good estimate of the true error.

   D

4. Training a 2-class k-NN with 80 samples from class A and 160 samples from class B, what is the training accuracy if k=240?

   A. 75%.

   B. 100%.

   C. 0%.

   D. 67%.

   D

5. Bootstrap is:

   A. An optimisation algorithm that avoids overfitting.

   B. A data partitioning method for getting a better estimate of a model's performance.

   C. A probability distribution function.

   D. A neural network training technique.

   B

6. Which of the following is NOT a hyperparameter?

   A. Coefficient vector $\mathbf{w}$ of a linear model.

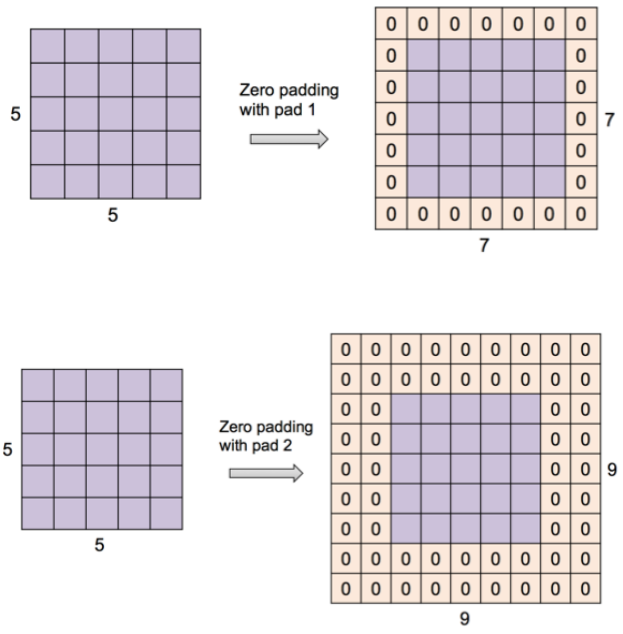
   B. Reguarlisation parameter of a support vector machine (SVM).

   C. Regularisation parameter of lasso.

   D. The number of neurons in the 2nd hidden layer of a neural network.

   A

7. Taking an $N \times N$ input, there is a common operation called zero padding in the convolutional layer. It adds zeros to the border of the input to form a new input. The convolutional filter is then applied to the new input. For instance, taking a 5×5 input, zero padding with pad 1 adds zeros to the border of the input to expand the input to a $(5+2) \times (5+2)$array, while zero padding with pad 2 forms a new input of $(5+4) \times (5+4)$:



What is the resulting size after applying zero padding with pad 3 to a 7×7 input?

A. 10×10

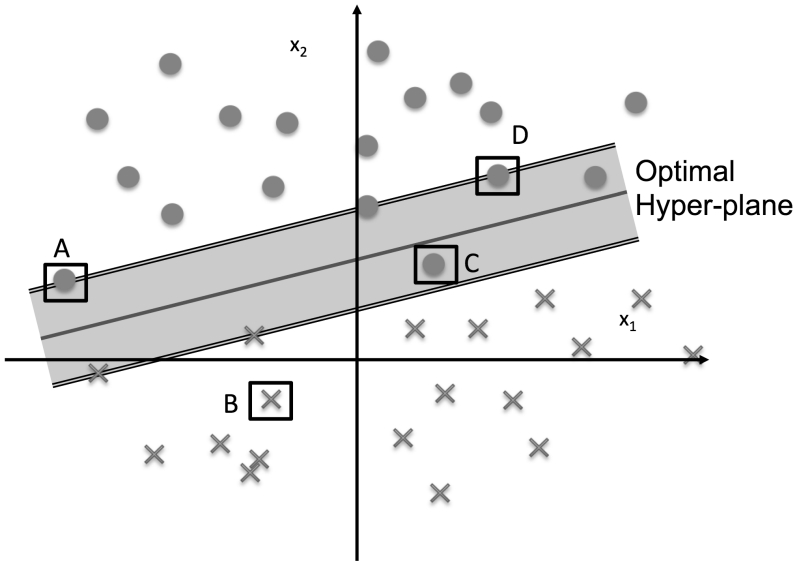
B. 13×13

C. 21×21

D. 4×4

B

8. The figure below displays the training samples and the learned SVM hyperplane. Which of the four highlighted samples is NOT a support vector?



A. Sample A.
B. Sample B.
C. Sample C.
D. Sample D.

B

9. Which of the following statement is true?

A. Convolutional neural network only supports 1D neurons.
B. The perceptron algorithm can be directly used for multi-class classification.
C. Multilayer perceptron is a nonlinear model.
D. In a 2-dimensional space, a soft-margin SVM with polynomial kernel generates a straight line as classification boundary.

C

10. Given a neural network with 10 input units, 5 hidden units, 1 output units, and a sigmoid activation function, how many weights does it contain (including bias)?

    A. 51.

    B. 55.

    C. 61.

    D. 16.

    C

11. Which of the following techniques is used in neural network training?

    A. Normal equations.

    B. Backpropagation.

    C. Quadratic programming.

    D. Iterative re-weighted least squares.

    B

12. Which of the following statements on deep learning is NOT true?

    A. Deep learning is a representation learning technique.

    B. Deep learning is about learning using neural networks.

    C. The perceptron algorithm is a typical deep learning model.

    D. A convolutional neural network with 22 hidden layers is a deep learning model.

    C

13. Which of the following is a linear model?

    A. $\frac{1}{3x_1+4x_2}$.

    B. $3x_1+4x_2+1$.

    C. $3x_1+\exp(x_2)+1$.

    D. $3x_1+x_1x_2+1$.

    B

14. A logistic regression model is trained to do three-class classification. Given a training sample characterised by a two-dimensional feature vector $\boldsymbol{x} = [3,4]$ from class 3, it predicts $p(\text{class } 1|\boldsymbol{x}) = 0.5$, $p(\text{class } 2|\boldsymbol{x}) = 0.3$ and $p(\text{class } 3|\boldsymbol{x}) = 0.2$. What is the cross entropy loss computed using this sample?

    A. -log(0.5).

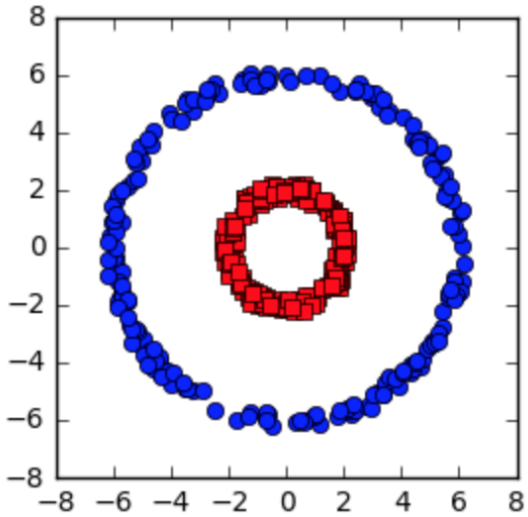
    B. -log(0.3).

    C. -log(0.2).

    D. None of the above.

    C

15. Which of the following statements is true?



    A. The two classes of data points can be separated using a perceptron algorithm.

    B. The two classes of data points can be separated using a soft-margin SVM with Gaussian kernel.

    C. The two classes of data points can not be separated using a hard-margin SVM with Gaussian kernel.

    D. The two classes of data points can not be separated using a k-NN classifier.

    B

16. A logistic regression model is trained to do three-class classification. Given a training sample characterised by a two-dimensional feature vector $\boldsymbol{x} = [3,4]$ from class 3, it predicts $p(\text{class } 1|\boldsymbol{x}) = 0.5$, $p(\text{class } 2|\boldsymbol{x}) = 0.3$ and $p(\text{class } 3|\boldsymbol{x}) = 0.2$. What is the cross entropy loss computed using this sample?

    A. -log(0.5).

    B. -log(0.3).

    C. -log(0.2).

    D. None of the above.

    C

17. Given two clusters $\{-1,-3\}$ and $\{3,5\}$ of one dimensional objects, use the *single* link with *Minkowski* distance to calculate their distance. As a result, the distance between two clusters is

    A. 2

    B. 4

    C. 6

    D. 8

    B

18. A 4-input neuron has weights of 1, 2, 3 and 4 and a bias parameter of 1. The activation function is set as the identity function. Given an input vector of [1, 2, 3, -4], the output of this neuron will be:

    A. -1,

    B. 16,

    C. -16,

    D. 123.

    A

19. Consider a dataset with four samples, where each sample is a point in a 2D space with a binary class label:

$$x_1 = [-1,-1],\ y_1 = -1$$
$$x_2 = [-1,+1],\ y_2 = +1$$
$$x_3 = [+1,-1],\ y_3 = +1$$
$$x_4 = [+1,+1],\ y_4 = -1$$

Which classifier cannot be used to classify these samples successfully?

A. A single layer perceptron.

B. A multilayer perceptron with one hidden layer.

C. A multilayer perceptron with two hidden layers.

D. A support vector machine with Gaussian kernel.

A

20. There is a training dataset containing 4, 5, 3 and 3 training samples belonging to class 1, 2, 3 and 4, respectively. If you try to classify the entire training dataset using a 15-NN classifier. What is the resulting training error?

A. 73.3

B. 66.7%

C. 50%

D. 33.3%

B

21. A separating plane is learned by the support vector machine: $y = 3x_1 + 4x_2 + 1$. What is the correct value of its margin (d as shown in the above figure)?

A. $d = 0.1$.

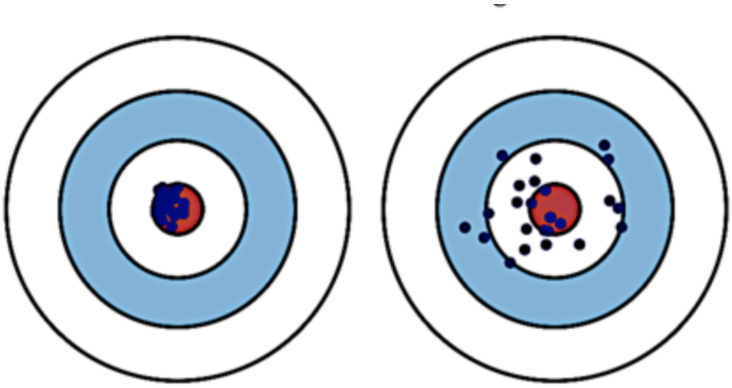
B. $d = 0.2$.

C. $d = 0.3$.

D. $d = 0.4$.

D

22. The red area in the figure below corresponds to the ground truth poihts, the blue points are predicted by a machine learning model. Compared to the prediction in the figure on the left, the right prediction has



A. Lower bias, lower variance.

B. Similar bias, higher variance.

C. Higher bias, higher variance.

D. Higher bias, similar variance.

B

23. There are three clusters of one-dimensional objects:

$$\text{cluster } 1 = \{-1, 1\},\ \text{cluster } 2 = \{2, 5\},\ \text{cluster } 3 = \{0, 3\}.$$

Use the *complete* link with *Euclidean* distance to calculate the distance between the two clusters. According to the Agglomerative Algorithm, which two clusters should be merged?

A. Clusters 1, 2.

B. Clusters 2, 3.

C. Clusters 1, 3.

D. Can not be decided.

C

24. Assume you have trained a linear model $f(\mathbf{x}) = \mathbf{w}\mathbf{x}^T + b$ with weight vector $\mathbf{w} = [1, -1, 1]$ and bias $b = 0$ to assign a 3-dimensional data point to either class A or class B, by using the logistic regression approach. It assigns the point $\mathbf{x} = [0, 0, 2]$ to class A with the estimated posteriors $p(A|\mathbf{x}) = 88\%$ and $p(B|\mathbf{x}) = 12\%$. Given a new data point $\mathbf{x} = [3, 1, -1]$, what would the model predict? Note: A logistic sigmoid function has the form of $\sigma(x) = \frac{1}{1+\exp(-x)}$.

A. The computed posterior is $p(A|\mathbf{x}) = 86\%$ and $p(B|\mathbf{x}) = 14\%$, and the predicted label is class A.

B. The computed posterior is $p(A|\mathbf{x}) = 14\%$ and $p(B|\mathbf{x}) = 86\%$, and the predicted label is class B.

C. The computed posterior is $p(A|\mathbf{x}) = 73\%$ and $p(B|\mathbf{x}) = 27\%$, and the predicted label is class A.

D. The computed posterior is $p(A|\mathbf{x}) = 27\%$ and $p(B|\mathbf{x}) = 73\%$, and the predicted label is class B.

C

25. Find a least squares regression line for the three data points (-1, -2), (1, 2), (0, 0.3) using the linear model $\hat{y} = w_1 x + w_0$. What are the optimal values of w_0 and w_1 derived by setting the partial derivatives of the sum-of-squares error function as zero? Note: keep at most two decimal places.
Answer: _____

Model answer and marking scheme for Q25
$w_0 = 0.1$ and $w_1 = 2$

On how to derive it:
The sum-of-squares error function is

$$\begin{aligned} O &= \frac{1}{2}\left[(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2\right] \\ &= \frac{1}{2}\left[(w_1 x_1 + w_0 - y_1)^2 + (w_1 x_2 + w_0 - y_2)^2 + (w_1 x_3 + w_0 - y_3)^2\right] \\ &= \frac{1}{2}(-w_1 + w_0 + 2)^2 + \frac{1}{2}(w_1 + w_0 - 2)^2 + \frac{1}{2}(w_0 - 0.3)^2 . \end{aligned}$$

Derive the partial derivatives:

$$\begin{aligned} \frac{\partial O}{\partial w_0} &= (-w_1 + w_0 + 2) + (w_1 + w_0 - 2) + (w_0 - 0.3) = 3w_0 - 0.3, \\ \frac{\partial O}{\partial w_1} &= -(-w_1 + w_0 + 2) + (w_1 + w_0 - 2) = 2w_1 - 4. \end{aligned}$$

Set the partial derivatives to zero and solve the linear equations:

$$\begin{aligned} 3w_0 - 0.3 = 0, \\ 2w_1 - 4 = 0, \end{aligned}$$

which gives $w_0 = 0.1$ and $w_1 = 2$.

26. For the same question body as above, but to find the optimal values of w_0 and w_1 using the gradient descent approach with a learning rate of 1. Starting from the initial guess of $w_0^{(0)} = w_1^{(0)} = 0$, what are the updated values of w_0 and w_1 in the 4th iteration? Note: keep at most two decimal places.
Answer: _____

Model answer and marking scheme for Q26
$w_0 = -1.5$ and $w_1 = 0$

Gradient descent update equations with $\eta = 1$ are

$$w_0^{(k+1)} = w_0^{(k)} - \eta \left.\frac{\partial O}{\partial w_0}\right|_{w_0^{(k)}} = w_0^{(k)} - \eta(3w_0^{(k)} - 0.3) = -2w_0^{(k)} + 0.3$$
$$w_1^{(k+1)} = w_1^{(k)} - \eta \left.\frac{\partial O}{\partial w_1}\right|_{w_1^{(k)}} = w_1^{(k)} - \eta(2w_1^{(k)} - 4) = -w_1^{(k)} + 4$$

The first updating iteration:

$$w_0^{(1)} = -2w_0^{(0)} + 0.3 = 0.3$$
$$w_1^{(1)} = -w_1^{(0)} + 4 = 4$$

The second updating iteration:

$$w_0^{(2)} = -2w_0^{(1)} + 0.3 = -0.3$$
$$w_1^{(2)} = -w_1^{(1)} + 4 = 0$$

The third updating iteration:

$$w_0^{(3)} = -2w_0^{(2)} + 0.3 = 0.9$$
$$w_1^{(3)} = -w_1^{(2)} + 4 = 4$$

The fourth updating iteration:

$$w_0^{(4)} = -2w_0^{(3)} + 0.3 = -1.5$$
$$w_1^{(4)} = -w_1^{(3)} + 4 = 0$$

27. For the same question body as above, is the learning rate of 1 a good choice? Choose between "yes" and "No".
Answer: _____

Model answer and marking scheme for Q27
No

Reason: The weights computed by gradient descent with learning rate 1 oscillate around the true optimal weights ($w_0 = 0.1$ and $w_1 = 2$) without approaching it. For instance, w_1 changes between 4 and 0, and it will never approach the optimal value 2 (e.g., in the 5th updating iteration it has $w_1^{(5)} = -w_1^{(4)} + 4 = 4$). Thus, this learning rate setting is not good.

28. You are to cluster eight points: $x_1 = (2,10), x_2 = (2,5), x_3 = (8,4), x_4 = (5,8), x_5 = (7,5), x_6 = (6,4), x_7 = (1,2)$ and $x_8 = (4,9)$. Suppose, you assigned x_1, x_4 and x_7 as initial cluster centres for K-means clustering ($K = 3$). Using K-means with the Manhattan distance, compute the three cluster centroids after the first round of the algorithm. Note: keep one decimal place only.
Answer: _____

Model answer and marking scheme for Q28
(2,10), (6,6), (1.5, 3.5)

On how to derive it:
Calculate distances between points and centroids:

$$d_{21} = 5, d_{31} = 12, d_{51} = 10, d_{61} = 10, d_{81} = 3;$$

$$d_{24} = 6, d_{34} = 7, d_{54} = 5, d_{64} = 5, d_{84} = 2;$$

$$d_{27} = 4, d_{37} = 9, d_{57} = 9, d_{67} = 7, d_{87} = 10.$$

At the first round, three clusters are formed:

$$\{x_1\}, \{x_3, x_4, x_5, x_6, x_8\}, \{x_2, x_7\}.$$

Based on the grouping, update the three centroids $A : (2,10)$,
$B : \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6,6)$, and $C : \left(\frac{2+1}{2}, \frac{2+5}{2}\right) = (1.5, 3.5)$.

29. For the same question body as above, what are the three clusters after two rounds of the algorithm.
Answer: _____

Model answer and marking scheme for Q29

$$\{x_1, x_8\}, \{x_3, x_4, x_5, x_6\}, \{x_2, x_7\}.$$

On how to derive it:
Using the new centroids, recalculate the distances as follows:

$$d_{1A} = 0, d_{2A} = 5, d_{3A} = 12, d_{4A} = 5, d_{5A} = 10, d_{6A} = 10, d_{7A} = 9, d_{8A} = 3,$$

$$d_{1B} = 8, d_{2B} = 5, d_{3B} = 4, d_{4B} = 3, d_{5B} = 3, d_{6B} = 2, d_{7B} = 9, d_{8B} = 5,$$

$$d_{1C} = 7, d_{2C} = 2, d_{3C} = 7, d_{4C} = 8, d_{5C} = 7, d_{6C} = 5, d_{7C} = 2, d_{8C} = 8.$$

After the second round of the algorithm, three updated clusters are

$$\{x_1, x_8\}, \{x_3, x_4, x_5, x_6\}, \{x_2, x_7\}.$$

30. For the same question body as above, how many rounds does it take for the algorithm to converge.
Answer: _____

Model answer and marking scheme for Q30
2

On how to derive it:
At the third round, three centroids are $A : (3, 9.5)$, $B : (6.5, 5.25)$, and $C : (1.5, 3.5)$. After the third round, the three clusters remain $\{x_1, x_8\}, \{x_3, x_4, x_5, x_6\}, \{x_2, x_7\}$. There is no change in clusters. The algorithm has converged.