

EE2211 - Introduction to Machine Learning

(Semester 2 : AY2021/2022)

1hr 15min

Time Allowed : 2 Hours

INSTRUCTIONS TO STUDENTS

1. This assessment paper contains Three Types of questions: True/False Questions, Multiple Option Questions, and Fill in Blanks.
2. There are 16 True/False Questions, each with 1.5 marks; 17 Multiple Option Questions, each with 2 marks; 9 Fill in Blanks with in total 26 blanks, each blank with 2.5 marks. In total there are 123 marks.
3. Students are required to answer ALL questions.
4. This is an open-book assessment.

True/False Questions (1.5 marks * 16 Questions)

1. If we would like to use imputation for data cleaning, we may consider replacing the missing value using the average from other samples.
a) True
b) False

2. Data integrity concerns the maintenance and the assurance of data accuracy and consistency.
a) True
b) False

3. Assume we have two urns. The first contains 6 white and 5 black balls, and the second contains 4 white and 5 black balls. We flip a fair coin, and pick the first urn if the coin shows a head while we pick the second urn if the coin shows a tail. We then draw a ball from the chosen urn. The probability of eventually getting a white ball is greater than a black ball.
a) True
b) False

4. The One Hot Encoding (OHE) can be applied to a binary target.
a) True
b) False

5. The polynomial model can approximate any continuous real-valued function on a closed and bounded interval to any degree of accuracy.
a) True
b) False

6. There are three samples of two-dimensional data points $\mathbf{X} = \begin{bmatrix} 1 & 3 \\ 4 & -1 \\ 3 & 3 \end{bmatrix}$ with the corresponding single output target vector $\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$. Suppose you want to use a full third-order polynomial model to fit these data, the polynomial model has 9 parameters to learn.
a) True
b) False

7. Suppose the model parameters are given by $\mathbf{w} = [w_1, \dots, w_d]^T$. Then an inclusion of the L2-regularization during training reduces the value of $\sum_{i=1}^d w_i^6$.
a) True

b) False

8. Suppose we are minimizing a cost function $C(w)$ with respect to w using a gradient descent algorithm. We observe the following: At iteration 1, $C(w)$ is -15. At iteration 2, $C(w)$ becomes 15. At iteration 3, $C(w)$ becomes -7. At iteration 4, $C(w)$ becomes 12. At iteration 5, $C(w)$ becomes -13. To improve our algorithm, we should consider decreasing the learning rate.

a) True

b) False

9. Compared with a decision tree, a random forest has lower squared bias and the same variance

a) True

b) False

10. To enhance the model generalization capability, we should wisely select samples from the test set and utilize them in the training stage, which will potentially improve the performances.

a) True

b) False

11. A higher training accuracy not necessarily leads to a higher test accuracy; however, in most cases, a higher training accuracy indeed leads to a higher validation accuracy.

a) True

b) False

12. Suppose we would like to develop a human face detector, where the positive class corresponds to face and negative class corresponds to non-face. In this case, *false positive* of the positive class denotes the face(s) missed by the detector.

a) True

b) False

13. We have built a binary classifier. During testing, it is possible that the sum of the false-positive-rate and the false-negative-rate is greater than 100%.

a) True

b) False

14. In the naïve K-means method, the optimal K is learned automatically.

a) True

b) False

15. Confusion matrix is a popular metric for evaluating regression tasks.
 a) True
 b) False
16. Recent research in neural networks mainly focus on manually setting the parameters of the networks smartly, so that they can generalize well to unseen data.
 a) True
 b) False

Multiple Option Questions (1.5 marks * 16 Questions)

1. Which of the following statement(s) is/are true about machine learning? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
- a) Jack went hiking with his friends and collected a bag of colorful stones. He would like to group these stones by color. As such, Jack is doing a supervised classification on the stones.
 - b) Jack went hiking with his friends and took 100 pictures of birds. One of his friends David, who is a scientist working on bird species, looked at all the pictures and told Jack there are in total five bird species. He picked one picture from each species, showed it to Jack, and told Jack the name of the species. Jack went home and categorized the rest of the pictures based on Jack's annotation. As such, Jack is doing a supervised classification on the bird pictures.
 - c) Jack went hiking with his friends and collected a bag of colorful stones. After he went home, he measured the weight, size, and color of each stone, before grouping them into different categories. The weight, size, and color measurement step can be thought as the feature extraction step.
 - d) None of the others.
2. Which of the following statement(s) is/are true about data wrangling? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
- a) Data wrangling concerns transforming "raw" data into an appropriate form for downstream purposes such as analytics.
 - b) Assume we have three classes, *car*, *truck*, *bicycle*, and we would like to use one-hot encoding to denote the three classes. One possible representation is as follows: 'car'=[0,0,1], 'truck'=[1,0,0], 'bicycle'=[1,1,1].
 - c) There is no need to conduct data wrangling if we know a priori that, there are no noisy samples in the data records.
 - d) None of the others.

3. The discrete random variable X has the following probability mass function,

X	1	2	3	4
$P[X]$	0.1	$2k$	$2k$	$5k$

What is value of k ?

- a) $k=0.1$
- b) $k= 0.2$
- c) $k=0.3$
- d) $k=0.02$

4. This question is related to the understanding of linear systems and partial derivatives. Which of the following statements below is/are correct? **There can be more than one answer.**

a) In under-determined linear systems, the number of parameters is greater than the number of unknown equations.

b) The system $\begin{bmatrix} 1 & 4 \\ 2 & 7 \\ -3 & 11 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2.5 \\ 4 \end{bmatrix}$ has no exact solution but an approximated solution is available using the left inverse.

c) If $f(\mathbf{x})$ is a vector-valued function of size $p \times 1$ and \mathbf{x} is an $m \times 1$ vector, then differentiation of $f(\mathbf{x})$ with respect to \mathbf{x} is an $m \times p$ matrix.

d) Consider the linear system $\mathbf{X}\mathbf{w} = \mathbf{y}$, $\mathbf{X} \in \mathbb{R}^{m \times d}$ is the input data matrix, $\mathbf{w} \in \mathbb{R}^{d \times 1}$ is the parameter vector, and $\mathbf{y} \in \mathbb{R}^{m \times 1}$ is the target vector. If $d > m$, the system has more equations than parameters.

e) None of the others.

5. This question is related to the understanding of linear systems, ridge regression, and polynomial regression. Which of the following statements below is/are correct? **There can be more than one answer.**

a) A set of linear equations is written as $\mathbf{X}\mathbf{w} = \mathbf{y}$ where $\mathbf{X} \in \mathbb{R}^{5 \times 3}$ is the input data matrix, $\mathbf{w} \in \mathbb{R}^{3 \times 1}$ is the parameter vector, and $\mathbf{y} \in \mathbb{R}^{5 \times 1}$ is the target vector. There are three simultaneous equations in this set of equations.

b) The ridge regression can be applied to multi-target regression.

c) The solution for a ridge regression problem with feature vectors in \mathbf{X} and targets in \mathbf{y} can be written as $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$ for $\lambda > 0$. As λ increases, $\hat{\mathbf{w}}^T \hat{\mathbf{w}}$ increases.

d) Given four samples of two-dimensional data points \mathbf{X} and the corresponding target output \mathbf{y} , the 2nd order polynomial regression system is an under-determined system.

e) None of the others.

6. You are given a collection of 5 training data points of three features (x_1, x_2, x_3) and their corresponding class labels which are stacked as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 3 & -2 \\ -4 & 0 & -1 \\ 3 & 1 & 8 \\ 2 & 1 & 6 \\ 8 & 4 & 6 \end{bmatrix}, \mathbf{Y \text{ class}} = \begin{bmatrix} \text{class 1} \\ \text{class 1} \\ \text{class 2} \\ \text{class 3} \\ \text{class 3} \end{bmatrix}.$$

We would like to predict the class labels of new test samples. Which of the following statements below is/are correct? **There can be more than one answer.**

a) If we would like to perform multi-class classification using the optimal linear regression model with an offset term, the estimated parameter matrix \mathbf{W} is of dimension 3 x 4.

b) Given a new sample $(x_1, x_2, x_3) = (1, -2, 4)$, the predicted class label using the optimal linear regression model with an offset term is class 3.

c) If we would like to perform multi-class classification using the optimal second-order polynomial model, the corresponding polynomial expansion matrix \mathbf{P} (with respect to \mathbf{X}) is of dimension 5 x 10. Therefore, it is an under-determined system.

(d) For two new samples \mathbf{X}_{new} , after learning and testing using some polynomial model, the prediction is $\mathbf{Y}_{new} = \mathbf{P}_{new}^T \hat{\mathbf{W}} = \begin{bmatrix} -0.58 & 0.40 & -0.67 \\ 1.25 & 0.10 & -0.20 \end{bmatrix}$. Therefore, the class label prediction for the first sample is class 3 and the second sample is class 1.

e) None of the others.

7. Suppose your machine learning algorithm achieves 95% accuracy on your training set, but only 10% accuracy on your test set. Which of the following modification might potentially improve your algorithm's test accuracy? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

- (1) Decrease Regularization
- (2) Increase Regularization
- (3) Decrease Model Complexity
- (4) Increase Model Complexity
- (5) Decrease Number of Features
- (6) Increase Number of Features

8. Suppose we are minimizing a cost function $C(w)$ with respect to w using a gradient descent algorithm. We observe the following: At iteration 1, $C(w)$ is 11. At iteration 2, $C(w)$ becomes 10.9. At iteration 3, $C(w)$ becomes 10.7. At iteration 4, $C(w)$ becomes 10.6. At iteration 5, $C(w)$ becomes 10.55. Which of the following is true? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
- (1) Decreasing the learning rate will speed up the optimization further
 - (2) Increasing the learning rate will speed up the optimization further
 - (3) Adding a regularization to the cost function will speed up the optimization further
 - (4) Insufficient information in the question to tell which option is true
9. Consider the cost function $C(w) = \text{DataLoss}(w) + \lambda \text{Regularization}(w)$. Assume that the global minimum of $C(w)$ when $\lambda = 2$ is 12. Now we change λ to be equal to 20 and again minimize $C(w)$. Assuming we attain the global minimum, the new optimal cost function value C
- (1) will be higher than before
 - (2) will be lower than before
 - (3) will stay the same
 - (4) Insufficient information in the question to tell which option is true.
10. Consider the cost function $C(w) = \text{DataLoss}(w) + \lambda \text{Regularization}(w)$. When $\lambda = 3$, the minimum of $C(w)$ is achieved at $w = w^*$ and $C(w^*) = 25$. Now we change λ to be equal to 30 and minimize $C(w)$. The minimum now is achieved at $w = w^{**}$. Assume that we always achieve the global minimum. Which of the following is true? Note that there might be more than one true option. If so, you should select all the correct options.
- (1) In general, $\text{Regularization}(w^*) > \text{Regularization}(w^{**})$
 - (2) In general, $\text{Regularization}(w^*) < \text{Regularization}(w^{**})$
 - (3) In general, $\text{DataLoss}(w^*) > \text{DataLoss}(w^{**})$
 - (4) In general, $\text{DataLoss}(w^*) < \text{DataLoss}(w^{**})$
 - (5) Insufficient information in the question to tell which option is true.
11. Suppose in a 2-class classification problem, our decision tree algorithm achieves 52% accuracy on our training set, and also 52% accuracy on our test set. Which of the following modification might potentially improve our algorithm's test accuracy? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.
- (1) Increase maximum depth of tree

- (2) Decrease maximum depth of tree
- (3) Increase minimum number of samples for splitting a leaf node
- (4) Decrease minimum number of samples for splitting a leaf node
- (5) Try random forest instead of decision tree

12. Which of the following statement(s) about the training-validation-test is/are true? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

- (a) We should always set the number of training samples to be equal to the number of test samples.
- (b) We should always set the number of test samples to be equal to the number of validation samples.
- (c) If no validation data is available, we should test the performances of different models on the test set, and choose the model that performs the best to be our model for deployment.
- (d) None of the others

13. Which of the following statement(s) is/are true about the confusion matrix in a classification task? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

- (a) A reasonable binary classifier should always lead to the following result: $\text{True Negative} + \text{True Positive} > \text{False Negative} + \text{False Positive}$.
- (b) A reasonable binary classifier should always leads to the following result: $\text{False Negative} > \text{False Positive}$
- (c) A reasonable binary classifier should always leads to the following result: $\text{True Positive} > \text{True Negative}$
- (d) None of the others.

14. Which of the following statement(s) is/are true about ROC curve? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

- (a) ROC curve is a widely used evaluation metric.
- (b) Area Under the Curve (AUC) is lower-bounded by 0 and upper-bounded by 1.
- (c) Given a fixed dataset, a higher Gini coefficient of ROC curve usually indicates better performance.
- (d) None of the others

15. Which of the following statement(s) is/are true about K-means? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

- (a) Before convergence, K-means is guaranteed to not increase the objective values.
- (b) K-means always produces the same clusters and centroids for different initializations.
- (c) Given a fixed K value, a set of samples, and the objective of K-means, i.e., $J = \sum_{i=1}^m \sum_{k=1}^K w_{ik} \|\mathbf{x}_i - \mathbf{c}_k\|^2$, if we traverse all possible values of w_{ik} for all i and k , we will find the global optimal solution, which is guaranteed to be no worse than the result of K-means.
- (d) None of the others

16. Which of the following statement(s) is/are true about unsupervised learning? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

- (a) In unsupervised learning, we may use the labels of the validation data for reference.
- (b) K-means always yields better results than fuzzy C-means.
- (c) K-means never finds the global optimal solutions.
- (d) None of the others.

17. Which of the following statement(s) is/are true about deep neural networks? Note that there might be more than one true option. If so, you should select all the correct options in order to get all the marks.

- (a) If we build a multilayer perceptron network, we do not need non-linear activation functions.
- (b) Backpropagation involves a backward stage but not a forward stage.
- (c) Parameters of the neural networks are chosen manually in a smart way, in order to produce the outputs as similar as the labels.
- (d) None of the others.

Fill in the Blank (In total 9 Questions with 26 blanks, each 2.5 questions, in total 2.5*26 = 65 marks)

1. You are given a collection of 6 training data points of one feature (x) and their class label (y) as follows:

$x \in \{4, 7, 10\}$ are labelled as $y = -1$, $x \in \{2, 3, 9\}$ are labelled as $y = +1$.

Answer each of the following subquestions regarding the training outcome or prediction. Note that the training outcome is in terms of the error count, i.e., the total number of data points being incorrectly classified. **[4 Blanks]**

2

- (i) What is the best training error count that you can achieve on these data from a linear classifier, i.e., $f(x) = \text{sign}(ax + b)$, on the original input features?
- (ii) You need to predict the class label of a new data point given by $x_t = 6$ when a linear classifier, i.e., $f(x) = \text{sign}(ax + b)$ is used for prediction. What is the class label (-1 or 1) for x_t ?

-1

- (iii) What is the best training error count that you can achieve from a 4th order polynomial model, e.g., $f(x) = \text{sign}(\sum_{k=0}^4 c_k x^k)$?

0

- (iv) You need to predict the class label of a new data point given by $x_t = 6$ when a 4th order polynomial model, e.g., $f(x) = \text{sign}(\sum_{k=0}^4 c_k x^k)$ is used for prediction. What is the class label (-1 or 1) for x_t ?

-1

2. Suppose our training set comprises 5 data samples shown in the table below. Each data point consists of 3 features (x_1, x_2, x_3) and a regression target y .

x_1	3.3459	1.0893	3.2103	1.744	1.6762
x_2	2.7435	2.9113	1.4706	1.2895	2.1366
x_3	-1.7253	-0.7804	-0.9944	0.5307	-1.0502
y	2.9972	1.1399	2.228	0.3387	2.5042

On the other hand, our test set comprises 4 data samples shown in the table below.

x_1	0.9478	1.4931	2.5809	2.0607
x_2	1.1619	1.7705	2.9569	1.1695
x_3	0.7779	-0.6149	-1.2087	0.4972
y	0.4168	1.4707	2.1037	0.9177

Our goal is to train a quadratic regression model to predict y . Suppose we utilize Pearson's correlation to perform feature selection.

For the purpose of feature selection, the correlation between feature x_1 and target y is equal to BLANK1 (your answer should be up to 3 decimal places)

For the purpose of feature selection, the correlation between feature x_2 and target y is equal to BLANK2 (your answer should be up to 3 decimal places)

For the purpose of feature selection, the correlation between feature x_3 and target y is equal to BLANK3 (your answer should be up to 3 decimal places)

Based on the correlations, the best feature is feature BLANK4 (Your answer should be an integer 1, 2 or 3 corresponding to feature 1, feature 2 or feature 3).

[4 Blanks]

0.651

0.372

-0.931

featurer 3

3. We would like to minimize the cost function $C(w) = \sin^2(w)$ using gradient descent. Suppose we initialize our algorithm with $w = 3$. The gradient of the cost function at $w = 3$ is BLANK1 (your answer should contain up to 3 decimal places). Suppose the step size is 0.1. After the first round of gradient descent, the updated value of w is BLANK2 (your answer should be up to 3 decimal places). **Note that w is in radians.**

[2 Blanks]

$$c' = 2\sin w \cos w$$

$$= 2\sin 3 \cos 3 = -0.279$$

$$3.0279 = 3.028$$

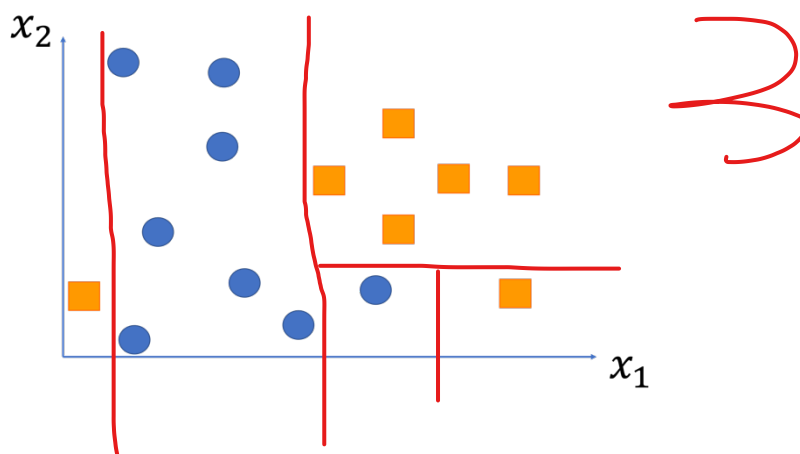
4. We would like to minimize the cost function $C(x, y) = x^2 + xy^2$ using gradient descent. Suppose we initialize our algorithm with $x = 3$, $y = 2$. Suppose the step size is 0.2. After the first round of gradient descent, the updated value of x is BLANK1 (your answer should be up to 2 decimal places) and the updated value of y is BLANK2 (your answer should be up to 2 decimal places).

$$1, -0.4$$

[2 Blanks]

5. To classify the data points in the plot below with 100% accuracy, a decision tree needs to have a minimum depth of BLANK (your answer should be a positive integer)

[1 Blank]



6. Consider the following dataset comprising 10 datapoints: $\{x, y\} = \{0.2, 2.1\}, \{0.7, 1.5\}, \{1.8, 5.8\}, \{2.2, 6.1\}, \{3.7, 9.1\}, \{4.1, 9.5\}, \{4.5, 9.8\}, \{5.1, 12.7\}, \{6.3, 13.8\}, \{7.4, 15.9\}$. Our goal is to use a regression tree to predict y from x . Suppose at depth 1, we consider a decision threshold of 3.

The MSE at the root is BLANK1 (your answer should be up to 3 decimal places).

The overall MSE at depth 1 is BLANK2 (your answer should be up to 3 decimal places).

[2 Blanks]

20.638

5.568

7.

In a binary classification task, we use a dataset that contains in total 1000 samples, where we take out 200 samples as the test set. We then conduct a standard 4-fold cross validation. In each fold, we set the number of validation samples to be the same as that of test samples. As such, in each fold, the binary classifier is trained on BLANK1 samples.

600

Now we have three classifier candidates, where their average training, validation, and number of parameters are given as follows. We should adopt classifier # BLANK2 for the test set. #2

Classifiers	Average Training Accuracy	Average Validation Accuracy	Number of parameters
Classifier #1	90.5%	92.1%	5000
Classifier #2	91.5%	92.1%	3000
Classifier #3	93.5%	90.5%	3000

[2 Blanks]

8.

Jack developed a classifier to detect spam emails. The classifier conducts binary classification, and identifies an email to be spam (positive class) or not spam (negative class). The detected spam emails go into the 'Spam' folder and non-spam emails go into the 'Inbox' folder.

In an unseen dataset of 100 emails, the classifier achieves the following performances

	Spam (Predicted)	Not Spam (Predicted)
--	------------------	----------------------

Spam (Actual)	40	x
Not Spam (Actual)	15	35

The number x is BLANK1. 10

We can also derive that, among the 100 emails, BLANK2 emails have the actually label of being Spam, and in total BLANK3 emails are correctly classified.

55, 75

Assume we have the following cost matrix

	Spam (Predicted)	Not Spam (Predicted)
Spam (Actual)	$C_{s,s} * 40$	$C_{s,n} * x$
Not Spam (Actual)	$C_{n,s} * 15$	$C_{n,n} * 35$

We assume that, if an email is correctly classified, we impose no penalty, i.e., $C_{s,s} = C_{n,n} = 0$. On the other hand, we know that customers will be very unhappy if the spam folder contains non-spam emails. But customers find it acceptable that some spam emails go into the inbox folder. In this case, when comparing two classifiers, we should ensure that

$C_{s,n}$ BLANK4 $C_{n,s}$ (choose '=', '>' or '<' and fill in the blank).

<

With the same dataset, David also developed a classifier that achieves the following performance

	Spam (Predicted)	Not Spam (Predicted)
Spam (Actual)	50	y
Not Spam (Actual)	10	25

With the same set of $\{C_{n,n}, C_{s,s}, C_{n,s}, C_{s,n}\}$ as above, David's classifier is considered to be BLANK5 (1:Better or 2:Worse; choose '1' or '2' and fill in the blank) than the one of Jack's.

1

[5 Blanks]

9.

We have a collect of 10 books. We measure their thickness in millimetres and summarize them as follows

Book ID	01	02	03	04	05	06	07	08	09	10
Thickness (mm)	50	60	66	68	71	72	75	82	90	99

70.5

71.25

We'd like to group the books into two groups according to their thickness.

1) Assume we pick book 03 as the initial centroid for Group A, book 07 for Group B, and assign the books to the two groups using Euclidean distance. Before updating the new centroid, we will have BLANK1 books in Group A (please enter an integer here)

4

2) With the above group assignments, we re-estimate the new centroids for the two groups. The new centroid of Group A is BLANK2 , and the new centroid of Group B is BLANK3 . (please specify your answer with two decimal places)

61.00, 81.50

3) With updated centroid, we run group assignment again. In this second round, the number of books in the Group A (the group with lower centroid) is BLANK4.

5

[4 Blanks]

END OF PAPER