Module Code	Examiner	Department	Tel
INT104	Shengchen Li	INT	3077

2nd SEMESTER 24-25 SAMPLE EXAMINATION

Undergraduate

Artificial Intelligence

TIME ALLOWED: *2 hours*

INSTRUCTIONS TO CANDIDATES

1. **This is an open-book exam and the duration is 2 hours.**
2. **Total marks available are 100. This accounts for 60% of the final mark.**
3. **Answer all questions. Relevant and clear steps should be included in the answers.**
4. **Please use MCQ card delivered to answer MCQ questions. Please use answer booklet for answer other questions.**
5. **Only English solutions are accepted.**
6. **The use of calculator is allowed.**
7. **Any paper based material is allowed. NO DICTIONARIES.**

Section 1 Multiple Choice Questions (18 questions in each exam)

This section of the exam contains multiple-choice questions. Each question will be followed by four options A, B, C, and D. You are required to choose ONE answer that you deem to be the most appropriate.

1. In AI, what is meant by 'training data'? (C)

(A) Data used to measure AI performance

(B) Data used to destroy AI systems

(C) Data used to educate AI systems

(D) Data used to program AI manually

(3 Marks)

2. In supervised learning, what is the role of a 'label'? (C)

(A) It is data that the algorithm learns from autonomously.

(B) It is an error in the training data.

(C) It is the desired output for a given input.

(D) It is a type of algorithm used to process data.

(3 Marks)

3. What is the primary goal of classification in machine learning? (C)

(A) To predict a continuous value

(B) To divide data into groups based on similarity

(C) To predict the category or class of an instance

(D) To reduce the number of features in the dataset

(3 Marks)

4. What is a virtual environment in Python? (C)

(A) A type of Python interpreter

(B) A tool to manage different projects

(C) An isolated environment for Python projects

(D) A website for Python developers

(3 Marks)

5. What is feature scaling in the context of machine learning? (C)

(A) Changing the logo of the dataset

(B) Altering the importance of features according to the user's preference

(C) Bringing different features onto a similar scale

(D) Scaling up the model's complexity

(3 Marks)

6. Which of the following is not a hyper-parameter (C)

(A) The regularisation parameter in SVM

(B) The number of neighbours in kNN

(C) The distance between samples in hierarchical clustering

(D) The number of clusters in k-means

(3 Marks)

7. Which of the following statements best describes the principal components obtained in PCA? (C)

(A) They are correlated variables from the dataset

(B) They are the original features of the dataset

(C) They are new variables that are linear combinations of the original features

(D) They are values that replace missing data in the dataset

(3 Marks)

8. Which of the following is NOT a benefit of cross-validation? (C)

(A) Reducing the variance of model performance estimates

(B) Preventing overfitting

(C) Guaranteeing improved performance on independent test sets

(D) Utilizing the data effectively

(3 Marks)

9. How can lack of diversity in training data affect an AI model's performance in real-world applications? (C)

(A) It can make the model perform uniformly well across different scenarios.

(B) It reduces the overall complexity of the model.

(C) It can make the model biased toward the majority group represented in the data.

(D) It enhances the transparency of the model.

(3 Marks)

10. Which of the following is a common sign of overfitting? (C)

(A) High error on the training set

(B) Low error on both training and test sets

(C) Low error on the training set and high error on the test set

(D) High error on both training and test sets

(3 Marks)

11. How does the SVM algorithm handle multi-class classification problems? (C)

(A) By ignoring class labels that are not binary

(B) By using a single multiclass kernel

(C) By transforming them into multiple one-vs-one classification problems

(D) By converting them into regression problems

(3 Marks)

12. Which of the following is a disadvantage of using SVMs? (C)

(A) Requires large amounts of data

(B) Inefficient with high-dimensional data

(C) Sensitive to the choice of the kernel parameters

(D) Only applicable for binary classification

(3 Marks)

13. What is a decision tree in machine learning? (C)

(A) A linear regression model used for making decisions

(B) A non-linear model used for clustering

(C) A flowchart-like tree structure used for decision making

(D) A type of neural network

(3 Marks)

14. Random Forests operate by combining the results of multiple: (C)

(A) Neurons in a neural network

(B) Linear regression models

(C) Decision trees

(D) Clustering models

(3 Marks)

15. In the context of ensemble learning, what is 'voting' used for? (C)

(A) To select the best model from the ensemble

(B) To determine the weights of different models in the ensemble

(C) To combine predictions from multiple models

(D) To decide which features to use in models

(3 Marks)

16. What is hierarchical clustering primarily used for? (C)

(A) Classifying data into a fixed number of clusters

(B) Predicting future data points

(C) Identifying a hierarchy of clusters within the data

(D) Reducing the dimensionality of the data

(3 Marks)

17. Which of the following is a common method to determine the 'k' value in k-means? (B)

(A) Maximum likelihood estimation

(B) Cross-validation

(C) The elbow method

(D) Accuracy scoring

(3 Marks)

18. In GMM, what does a diagonal covariance matrix imply about the distribution of data? (C)

(A) Features are correlated

(B) Features are uncorrelated and have equal variance

(C) Features are uncorrelated and have variable variances

(D) All features have the same variance

(3 Marks)

Section 2 Computation Questions

19. Consider a dataset containing the following 2D points:

A = (2, 3), B = (3, 4), C = (10, 15), D = (15, 12), E = (10, 10), F = (10, 14)

You are required to perform one iteration of the k-means clustering algorithm manually, with K=2 starting with initial centroids as Z1 = (2, 3) and Z2 = (10, 15).

Use city block distance for easier computation.

(14 Marks)

ANSWER:

The city block distance between all samples and cluster centroids are: ***(6 Marks)***

	A	B
A	0	20
B	2	18
C	20	0
D	22	8
E	15	5
F	19	1

The sample A and B could form a cluster $Z1'$. ***(2 Marks)*** *C, D, E and F form a cluster $Z2'$* ***(2 Marks)***

The new centroid $Z1'$ is $(2.5, 3.5)$. ***(2 Marks)*** *The new centroid $Z2'$ is* $(11.25, 12.75)$ ***(2 Marks)***

20. A system classifies electronic components as **Capacitor (C)** or **Resistor (R)** using two features:

- Size (mm)
- Type (0 = Cylindrical, 1 = Rectangular)

The training data provided is:

Component	Size	Type	Label
Sample 1	15	0	C
Sample 2	17	0	C
Sample 3	16	1	R
Sample 4	18	1	R
Sample 5	14	0	C

With city block distance, please identify a component (Size = 16.5 mm, Type = 1) as either capacitor or resistor. Use kNN algorithm with the value of k = 3.

(14 Marks)

ANSWER:

The distance between samples and the component are: ***(10 Marks)***

$$\begin{aligned}
&\textit{Sample 1:} && |16.5-15|+|1-0| = 1.5+1 = 2.5\\
&\textit{Sample 2:} && |16.5-17|+|1-0| = 0.5+1 = 1.5\\
&\textit{Sample 3:} && |16.5-16|+|1-1| = 0.5+0 = 0.5\\
&\textit{Sample 4:} && |16.5-18|+|1-1| = 1.5+0 = 1.5\\
&\textit{Sample 5:} && |16.5-14|+|1-0| = 2.5+1 = 3.5
\end{aligned}$$

The nearest Nearest neighbors: Sample 3 (0.5), Sample 2 & Sample 4 (tie at 1.5) ***(2 Marks)*** *, so the component is a resistor.* ***(2 Marks)***

21. Given a dataset that consists of following points below:

A = (1, 2), B = (2, 0), C = (1, 0), D = (1, 1), E = (4, 0)

Cluster the data points by agglomerative clustering with maximum **city block** distance and draw the cluster dendrogram.

(14 Marks)

	A	B	C	D	E
A	0				
B	3	0			
C	2	1	0		
D	1	2	1	0	
E	5	2	3	4	0

ANSWER:

The city block distance between all samples are: ***(4 Marks)***

The sample C and D could form a mini-cluster F. ***(1 Mark)*** *The distance between mini-clusters are:* ***(2 Marks)***

	A	B	E	F
A	0			
B	3	0		
E	5	2	0	
F	2	2	4	0

The sample B and E could form a mini-cluster G. ***(1 Mark)*** *The distance between mini-clusters are:* ***(1 Mark)***

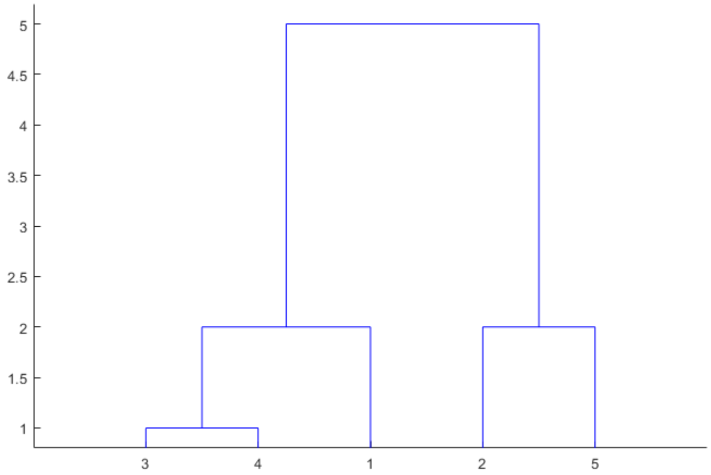
	A	F	G
A	0		
F	2	0	
G	5	5	0

The sub-cluster F and sample A could form a mini-cluster H. ***(1 Mark)*** *The remaining sub-clusters G and H would be merged together eventually with a distance of 5.* ***(1 Mark)***

The resulting dendrogram is: ***(3 Marks)***

22. The given table shows the preference of golf players for playing golf or not. Based on the information presented in the table, what is the golf player's preference under the condition of rainy weather and humidity higher than 65%? Please use Naïve Bayes to make the decision.

(14 Marks)



Outlook	Humidity	Wind Speed	Preference
Rainy	80%	0.5m/s	Yes
Rainy	40%	0.2m/s	Yes
Rainy	50%	5.0m/s	No
Rainy	50%	0.2m/s	Yes
Rainy	75%	4.0m/s	No
Sunny	70%	5.0m/s	No
Sunny	75%	0.4m/s	No
Sunny	80%	0.1m/s	No
Sunny	50%	0.2m/s	Yes
Sunny	40%	4.0m/s	Yes

ANSWER:

The class-conditioned probability table for the condition of outlook is

Outlook	Yes	No	p
Rainy	3	2	0.5
Sunny	2	3	0.5
p	0.5	0.5	

By this table we have $p(Rainy|Yes) = \frac{3}{5} = 0.6$ *and* $p(Rainy|No) = \frac{2}{5} = 0.4$. ***(4 Marks)***

The class-conditioned probability table for the condition of humidity is

Humid	Yes	No	p
$\leq 65\%$	4	1	0.5
$> 65\%$	1	4	0.5
p	0.5	0.5	

By this table we have $p(humid > 65\%|Yes) = \frac{1}{5} = 0.2$ *and* $p(humid > 65\%|No) = \frac{4}{5} = 0.8$. ***(4 Marks)***

Posterior Probability for Preference "Yes" is

$$p(Yes|Rainy,\ humid > 65\%) = \frac{p(Rainy,\ humid > 65\%|Yes)p(Yes)}{p(Rainy,\ humid > 65\%)}$$

$\propto p(Rainy,\ humid > 65\%|Yes)p(Yes) = p(Rainy|Yes)p(humid > 65\%|Yes)p(Yes) = 0.6 \times 0.2 \times 0.5 = 0.08$ ***(2 Marks)***

Posterior Probability for Preference "No" is

$$p(No|Rainy,\ humid > 65\%) = \frac{p(Rainy,\ humid > 65\%|No)p(No)}{p(Rainy,\ humid > 65\%)}$$ ***(2 Marks)***

$\propto p(Rainy,\ humid > 65\%|No)p(No) = p(Rainy|No)p(humid > 65\%|No)p(No) = 0.4 \times 0.8 \times 0.5 = 0.16$ ***(2 Marks)***

As a result, golf players prefer not to play golf under the condition of rainy weather and humidity lower than 65%. ***(2 Marks)***

23. An email service tested their spam filter on 2,000 emails:

- 320 legitimate emails **incorrectly marked as spam**
- 45 spam messages **not detected** by the filter
- 1600 emails **correctly identified** as legitimate
- 35 spam messages **properly blocked**

Calculate:

a) Precision (Spam detection rate)

b) Recall (True positive rate)

(10 Marks)

ANSWER:

The confusion matrix is: ***(6 Marks)***

$$\begin{bmatrix} & Spam & Legit \\ Spam & 35 & 45 \\ Legit & 320 & 1600 \end{bmatrix}$$

1. *Precision:* $\frac{35}{35+320} \approx 0.0986$ ***(2 Marks)***
2. *Recall:* $\frac{35}{35+45} = 0.4375$ ***(2 Marks)***

Section 3 Programming Questions (FINAL EXAM ONLY, there will be 9 blanks in 2 questions to be filled in)

24. Assume a dataset is stored in a variable `X_knn` where each column of `X_knn` represents a feature and each row of `X_knn` represents a data sample. The samples belong to a certain number of classes. A variable `label` stores the class information of each sample as a column vector where each row of `label` represents a data sample.

Both `X_knn` and `labels` are an `ndarray` in Numpy.

The Python script on the next page attempts to find the best value of `k` in kNN algorithm. A plot is generated to compare the performance of candidate systems that with different value of `k`

Please fill in the blank marked as `[#001]` to `[#010]` as appropriate in the script.

Each blank in the Python script is worth 2 marks.

A set of API of Python has been provided in the section of Appendix for your reference.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import cross_val_score, KFold

k_values = range(5,16)

# Initialize lists to store accuracy and F1 scores for
# each value of k
accuracies = [#001]

# Perform 5-fold cross validation for each value of k and
# calculate accuracy
for k in [#002]:
    knn = KNeighborsClassifier(n_neighbors=[#003])
    kf = KFold(n_splits=5, shuffle=True, random_state=42)
    acc_scores = cross_val_score(knn, X_knn, label, \
        cv=[#004], scoring=’accuracy’)
    accuracies.append([#005])

k_best = k_values[accuracies.index(max([#006]))]

plt.figure(figsize=(10, 5))
plt.plot([#007], [#008], marker=’o’, label=’Accuracy’)
plt.xlabel(’k’)
plt.ylabel(’Score’)
plt.title(’Accuracy vs. k’)
plt.xticks([#009])
plt.legend()
plt.[#010]
```

(20 Marks)

ANSWER:

The blanks of in the script should be filled as:

[#001]: `[]` *(2 Marks)*

[#002]: `k_values` *(2 Marks)*

[#003]: `k` *(2 Marks)*

[#004]: `kf` *(2 Marks)*

[#005]: `acc_scores.mean()` *(2 Marks)*

[#006]: `accuracies` *(2 Marks)*

[#007]: `k_values` *(2 Marks)*

[#008]: `accuracies` *(2 Marks)*

[#009]: `k_values` *(2 Marks)*

[#010]: `show()` *(2 Marks)*

Section 4 Programming Questions (RESIT ONLY, there will be 2 questions in the resit exam)

25. Write a Python script that trains an ensemble classifier that ensembles a SVM classifier, a kNN classifier and a decision tree classifier. Compare the performance of the ensemble classifier and each individual classifier via cross validation. You could use variable name `features` to represent the dataset and the variable name `labels` to represent the labelling information.

No data generation or import process need to be included in the Python script. It is also not necessary to show formation of matrix `features`. You could always assume each row in `features` representing a sample and each column in `features` representing a feature.

(16 Marks)

ANSWER:

```
# Required imports (5 Marks)
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import VotingClassifier
from sklearn.model_selection import cross_val_score

# Initialize individual classifiers (3 Marks)
svm_clf = SVC(kernel='rbf', probability=True, \
    random_state=42)
knn_clf = KNeighborsClassifier(n_neighbors=5)
tree_clf = DecisionTreeClassifier(max_depth=5, \
     random_state=42)

# Create ensemble classifier (soft voting) (2 Marks)
ensemble_clf = VotingClassifier(
estimators=[
('svm', svm_clf),
('knn', knn_clf),
('tree', tree_clf)
],
voting='soft'
)

# List of classifiers to evaluate (2 Marks)
classifiers = [
('SVM', svm_clf),
('kNN', knn_clf),
('Decision Tree', tree_clf),
('Ensemble', ensemble_clf)
]

# Evaluate models using 5-fold cross-validation (4 Marks)
for name, clf in classifiers:
cv_scores = cross_val_score(clf, features, labels, cv=5, \
    scoring='accuracy')
print(f"{name}:")
print(f"  Mean Accuracy: {cv_scores.mean():.4f}")
print(f"  Std Deviation: {cv_scores.std():.4f}")
print("------------------------------")
```

Over the exam, the API information for *SVC*, *KNeighborsClassifier*, *DecisionTreeClassifier*, *VotingClassifier*, *cross_val_score* ***will be provided as a part of appendix.***

Section 5 Appendix: Edited Python API being used in this exam

A series of simplified API document will be provided here in formal exam. For this sample paper, the API of the following function would be expected to be provided.

- range
- KNeighborsClassifier
- KFold
- cross_val_score
- *append* in Class List
- *index* in Class Array

The following API information shall be provided in resit exam as an accompaniment and hint for Section 4.

- VotingClassifier
- SVC
- KNeighborsClassifier
- DecisionTreeClassifier
- cross_val_score

The following show an example of simplified API information of class SVC:

sklearn.svm.SVC

`class sklearn.svm.SVC(*, C=1.0)`: C-Support Vector Classification.

Parameters

`C: float, default=1.0`

Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared $l2$ penalty.

Methods

`fit(X, y)`

Fit the SVM model according to the given training data.

Parameters

`X: array-like, sparse matrix of shape (n_samples, n_features)`

Training vectors, where n_samples is the number of samples and n_features is the number of features.

`y: array-like of shape (n_samples,)`

Target values (class labels in classification, real numbers in regression).

Returns

`self: object`

Fitted estimator.

`predict(X)`

Perform classification on samples in X.

Parameters

`X: array-like, sparse matrix of shape (n_samples, n_features)`

Sample vectors, where n_samples is the number of samples and n_features is the number of features.

Returns

`y_pred: ndarray of shape (n_samples,)`

Class labels for samples in X.

Though simplified API information provided, bring a Python handbook with you will be extremely helpful. We strongly encourage you to bring a Python handbook for your reference over the exam.

END OF EXAM PAPER
THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM