Xi'an Jiaotong-Liverpool University
西交利物浦大学

Module Code	Examiner	Department	Tel
INT104	Shengchen Li	INT	3077

2nd SEMESTER 23-24 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. **Besides lecture notes and hand writing notes, only books (with an ISBN) are allowed. NO DICTIONARIES.**

Section 1 Multiple Choice Questions (54 Marks)

本试卷MCQ答案均为C

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'?

(A) Data used to measure AI performance test

(B) Data used to destroy AI systems

(C) Data used to educate AI systems 使用训练集训练模型，也可以说是教育AI**模型**
不能说是编程或手动操作或测量

(D) Data used to program AI manually

(3 Marks)

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

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

(B) It is an error in the training data. 在监督学习中，数据的标签"label"**就是数据正确的类别，模型**
需要通过学习尽量输出/预测数据正确的类别

(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?

(A) To predict a continuous value 这个是回归算法做的事情，尽管也可用来分类

(B) To divide data into groups based on similarity 这个更类似于无监督的聚类，根据数据相似度分组

(C) To predict the category or class of an instance 机器学习中的分类任务就是预测样本的类别

(D) To reduce the number of features in the dataset
这个是降维

(3 Marks)

4. What is a virtual environment in Python?

(A) A type of Python interpreter

(B) A tool to manage different projects

(C) An isolated environment for Python projects Python解释器及其周围的工具、库和配置

(D) A website for Python developers

(3 Marks)

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

(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 做标准化归一化等把不同规模的数据变为相同，消除数据规模不同带来的影响，并不会根据用户偏好调整特征权重（只是统一把权重归一）

(D) Scaling up the model's complexity

(3 Marks)

机器学习中，超参数指的是在学习前可以人为预先设置的参数；参数指的是模型在机器学习过程中学得的参数

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

(A) The regularisation parameter in SVM 正则化防止过拟合的参数是人为控制的，属于超参数

(B) The number of neighbours in kNN 最近邻个数k是人为手动设置的

(C) The distance between samples in hierarchical clustering 距离的计算方法可以指定，但不是距离大小

(D) The number of clusters in k-means
聚类数k显然是超参数

(3 Marks)

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

(A) They are correlated variables from the dataset

(B) They are the original features of the dataset 显然已经不是原始特征了

(C) They are new variables that are linear combinations of the original features PCA得到的新特征都是原特征集的线性组合，特征值代表方差大小，特征向量就是组合的系数

(D) They are values that replace missing data in the dataset 这么就把数据集中缺失的数据补完了？

(3 Marks)

交叉验证的好处

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

(A) Reducing the variance of model performance estimates 可以降低模型表现的方差，毕竟使用不同的多个训练集

(B) Preventing overfitting 典型的防止过拟合的方法

(C) Guaranteeing improved performance on independent test sets 不保证可以在测试集中表现更佳有时甚至更差，但总的来说提高模型泛化能力

(D) Utilizing the data effectively 一个训练集可以用n折，确实更有效

(3 Marks)

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

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

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

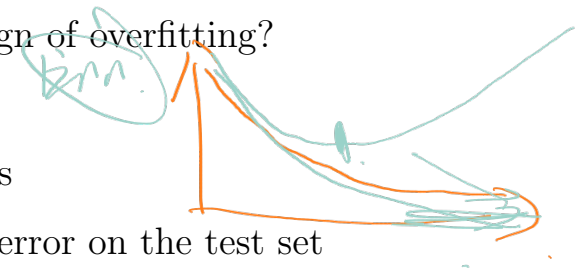
(C) It can make the model biased toward the majority group represented in the data. 训练集中不存在或较少的内容，模型练得少学到的少，以后遇到的时候性能就不好

(D) It enhances the transparency of the model.

(3 Marks)

随着训练的次数增加，模型在训练集上的错误逐渐下降，在测试集中的错误率会先下降（在拟合中）后上升（过拟合了）

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



(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

众所周知SVM的原理只是二分类，多分类策略有
一对一OvO：任意两个类别之间构造分类器，对n个类别共需构造n(n-1)/2个分类器。最后投票得出结果
一对多OvR：训练时依次把某个类别的样本归为一类，其他剩余的样本归为另一类，构造出多个SVM。最后分给分数最大的一类

(3 Marks)

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

(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?

(A) Requires large amounts of data 理论上只需要几个支持向量（样本点）所以适合小样本

(B) Inefficient with high-dimensional data 恰恰是避免维度灾难的一种方法

(C) Sensitive to the choice of the kernel parameters 核函数的选择确实很重要，决定了是否能进行高维非线性分类

(D) Only applicable for binary classification 你要说原版SVM也没错，但有OvO，OvR等多分类策略

(3 Marks)

13. What is a decision tree in machine learning?

(A) A linear regression model used for making decisions 决策树是非线性的模型，每次使用不同的特征判断

(B) A non-linear model used for clustering 不怎么是用来聚类的

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

(D) A type of neural network 不是NN

(3 Marks)

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

(A) Neurons in a neural network 不是NN，只是有很多不同的树，使用不同的样本和特征集

(B) Linear regression models 一般不回归

(C) Decision trees 许多不同的决策树，一起做出回答

(D) Clustering models 太离谱

(3 Marks)

集成学习

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

(A) To select the best model from the ensemble 大家一起上

(B) To determine the weights of different models in the ensemble 这个一般是手动控制不同模型重要性或者做Stacking，但不是目的

(C) To combine predictions from multiple models 三个臭皮匠，赛过诸葛亮

(D) To decide which features to use in models

(3 Marks)

16. What is hierarchical clustering primarily used for?

(A) Classifying data into a fixed number of clusters 根据距离不同可以有不同粒度的聚类，带来不同数量的簇

(B) Predicting future data points 聚类就别做预测了

(C) Identifying a hierarchy of clusters within the data

(D) Reducing the dimensionality of the data 不能降维

(3 Marks)

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

(A) Maximum likelihood estimation 最大似然估计凑什么热闹

(B) Cross-validation 只是防止过拟合，无监督没这个毛病

(C) The elbow method 劳大！肘！

(D) Accuracy scoring 如果只是看内聚+分离的话不用看这个，而且没有惩罚最后分出来一万个簇，可以用ARI

(3 Marks)

对角协方差矩阵

18. In GMM, what does a diagonal covariance matrix imply about the distribution of data? 协方差矩阵一般不是对角阵，如果只有对角线上有元素，代表是维度不相关
多元正态分布，每个特征之间相互独立；至于每个feature方差是否相同，一般不相同

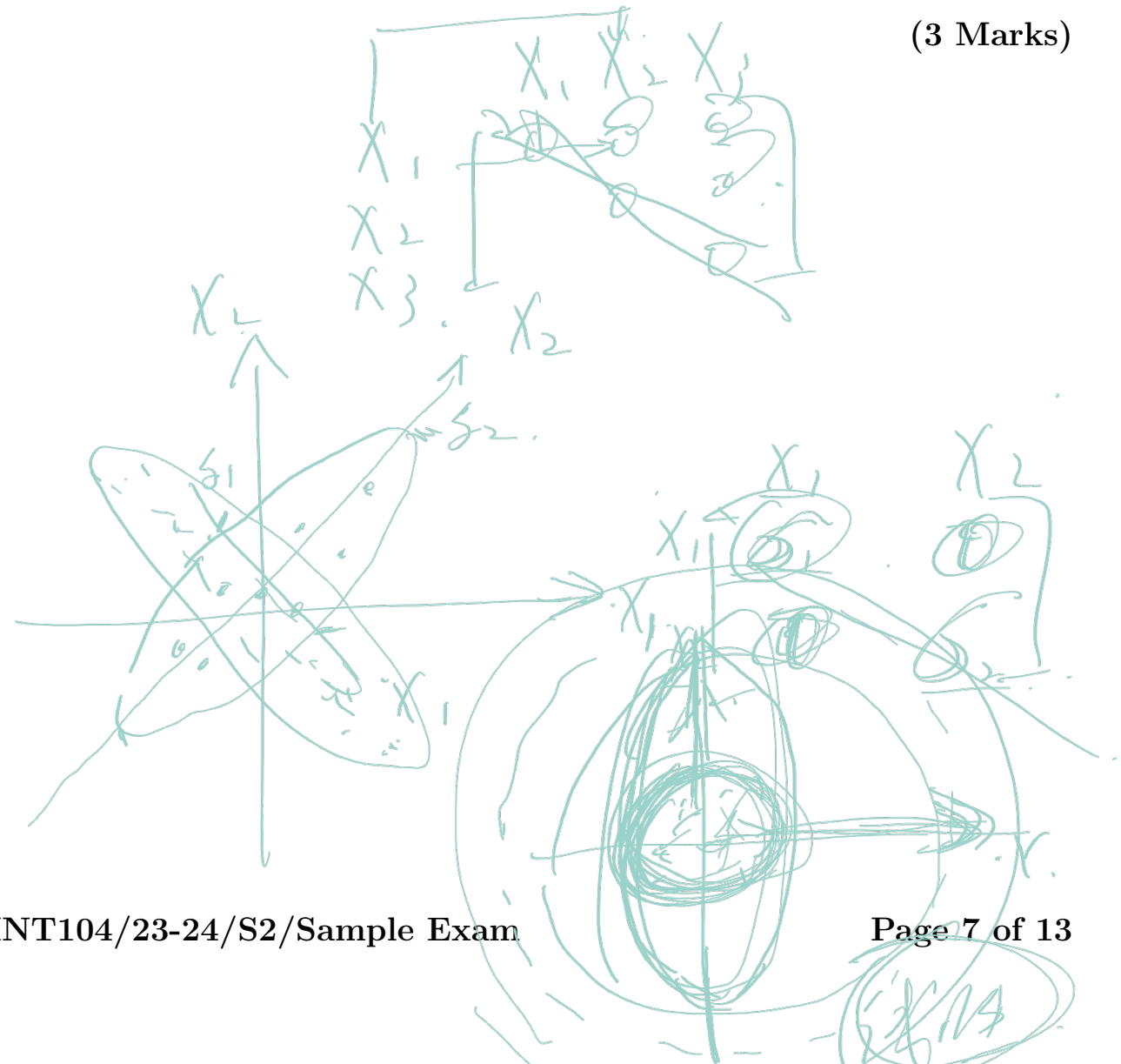
(A) Features are correlated

(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 (14 Marks)

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).

曼哈顿距离 $d = \sum_{j=1}^{k} |a_j - z_j|$

Use city block distance for easier computation.

(14 Marks)

①	z_1	z_2	cluster
A	0	20	1
B	2	18	1
C	20	7	2
D	22	8	2
E	15	5	2
F	19	4	2

$\Rightarrow z_1' = (2.5, 3.5)$
$z_2' = (11.75, 6.25)$
This is the End of iteration 1.

②	z_1'	z_2'	cluster
A	1	16.5	1
B	1	14.5	1
C	19	6	2
D	21	5.5	2
E	14	1.5	2
F	18	5	2

如果要求是"1 iteration"而是 devide into，则要继续迭代直到 centroids 不再变化(坐标)

Section 3 Programming Questions (32 Marks)

20. 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 and then answer the following question:

highest accuracy score under cross validation.

- According to the script, how is the best value of `k` determined? As accuracy is used as an indicator in the script, comment that in which case the accuracy cannot perform well for model evaluation.

① class imbalance → use f1-score/ARI instead.

- Given the range of value `k` tested, do believe the best value of `k` can be found? Why? Propose your own way to find the best value of `k`. (No codes need to be written)

① need wider grid search, test smaller/bigger k value, ② use more indexes to find best k, ARI, AUC-ROC

Each blank in the Python script is worth 2 marks. The question you are asked to answer is worth 6 marks.

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

```
1  import numpy as np
2  import matplotlib.pyplot as plt
3  from sklearn.neighbors import KNeighborsClassifier
4  from sklearn.model_selection import cross_val_score, KFold
5
6  k_values = range(5,16)
7
8  # Initialize lists to store accuracy and F1 scores for
9  # each value of k
10 accuracies = [#001]
11
12 # Perform 5-fold cross validation for each value of k and
13 # calculate accuracy
14 for k in [#002]:
15     knn = KNeighborsClassifier(n_neighbors=[#003])
16     kf = KFold(n_splits=5, shuffle=True, random_state=42)
17     acc_scores = cross_val_score(knn, X_knn, label, \
18         cv=[#004], scoring='accuracy')
19     accuracies.append([#005])
20
21 k_best = k_values[accuracies.index(max([#006]))]
22
23 plt.figure(figsize=(10, 5))
24 plt.plot([#007], [#008], marker='o', label='Accuracy')
25 plt.xlabel('k')
26 plt.ylabel('Score')
27 plt.title('Accuracy vs. k')
28 plt.xticks([#009])
29 plt.legend()
30 plt.[#010]
```

(32 Marks)

Section 4 The following type of question will ONLY appeared in RESIT exam

21. 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)

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