

FACULTY of SCIENCE and ENGINEERING

Department of Computer Science and Information Systems

**End-of-Semester** **Exam**

|  |  |  |
| --- | --- | --- |
| **Academic** **Year:** | 2021/2022 | **Semester:** Spring |
| **Module** **Title:** | Data Mining | **Module** **Code:** CS4168 |
| **Exam** **Duration:** | 2 Hours | **Total** **Marks:** 50 |
| **Lecturer:** | Dr. N. S. Nikolov | (Equal to 50% of the final grade) |

**Instructions** **to** **Candidates:**

This exam consists of 2 sections.

**Section** **A:**

• Section A consists of 30 multiple-choice questions, each having exactly one correct answer.

• You **must** **use** **the** **bubble** **sheets** to answer the questions in this section.

• Answer ALL questions. Each question will be marked as follows:

- Correctly answered: 1 mark

- Incorrectly answered: -0.25 marks

- Unanswered: 0 marks

**Section** **B:**

• Choose to answer either Q31 or Q32.

• Write your answers in the **answer** **booklet**.

• If both questions are answered, the one with higher marks will be considered.

**SECTION** **A:** **Multiple** **Choice** **Questions** **–** **answer** **ALL** **questions** **30** **marks**

**Q1.** According to Witten and Frank, data mining is the extraction of implicit, \_\_已知但之前未被明确表达的信息\_\_ and potentially useful information from data.

**Q2.** The input of a machine learning algorithm typically consists of a large number of examples which are: 数据点和实例

**Q3.** When treating missing values in a dataset, it is recommended to drop a column if more than \_超过数据总数的25-30%\_\_\_\_\_\_\_ of the values in it are missing.

**Q4.** In scikit-learn, \_\_\_\_\_StandardScaler\_\_\_\_\_ results in a distribution with a standard deviation equal to 1 and a mean approximately 0.

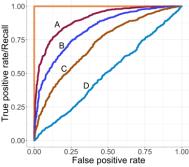
**Q5.** Not removing outliers from a dataset can increase the probability for training a/an \_inaccurate or less robust\_\_\_\_\_\_ predictive model.

**不从数据集中移除异常值会增加训练不准确或不太稳健的预测模型的可能性。 异常值是与数据集中其他数据点明显不同的数据点，它们会扭曲数据集的统计特性并影响机器学习算法的性能。 离群值会导致过度拟合，当模型过于复杂并且适合数据中的噪声而不是潜在模式时，就会发生这种情况。 过度拟合会导致泛化能力差，并降低对新的、看不见的数据的预测性能。 因此，通常建议在训练预测模型之前从数据集中识别并删除异常值**

**Q6.** The confidence in the accuracy of a predictive model is higher when the training, validation and test datasets are: diverse, representative, and independent.

**Q7.** Cross-validation and repeated holdout are methods used in the evaluation of: machine learning models

**Q8.** In *k*-fold cross validation, the \_\_\_the examples or data point\_folds\_\_\_\_ in the training dataset are split into *k* subsets of equal size.



**Q9.** In classification, the harmonic mean of the precision and recall is known as: F1score

**Q10.** The prime use of a validation dataset is for: evaluate and tune the performance of a machine learning model during the training process.

**Q11.** Assume a classification model does not perform well on a well-designed test dataset. Which one of the following actions is inappropriate?

不适当的操作是更改测试数据集或根据测试数据集上的性能调整模型。

适当的操作可能包括收集更多数据、平衡类、提高数据质量、更改模型架构或调整超参数。

**Q12****.** A regression model can be evaluated by calculating the \_\_预测误差或损失\_\_\_ on a test dataset.

Commonly used metrics to evaluate the prediction error of regression models include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R2).

常用的回归模型预测误差评估指标包括均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）和R平方（R2）等。

**Q13.** The plot below contains the ROC curves of four alternative classification models for the same classification problem. Which one of them is the best model?

A

**Q14.** You have trained four classifiers that determine if a

certain mushroom is edible (positive class) or poisonous

(negative class). Which one of the classifiers, represented by

their confusion matrices, would use to select the

mushrooms for your risotto?

对负类高召回，对正类高精度。必须确保所选分类器对两个类别都具有高准确度、精确度和召回率，以避免因食用有毒蘑菇而带来的任何健康风险。

**Q15.** Scikit-learn pipelines help avoid \_\_\_数据泄露\_\_\_\_\_ in cross-validation, by ensuring that all steps in the pipeline are applied separately to the training and test datasets.

数据泄露问题是指在交叉验证中，测试数据集中的信息可能泄露到训练数据集中，导致模型在测试数据集上表现良好，但在实际情况下表现较差。

**Q16.** All steps in a scikit-learn pipeline, except the last one, must be \_\_\_\_\_\_\_\_\_transformers. \_

**Q17.** If the examples in a training dataset are not linearly separable, it may be possible to make them linearly separable by:映射到高维空间，从而使它们在高维空间中线性可分

使用SVM（Support Vector Machines）算法来实现核方法。SVM算法可以使用不同的核函数进行分类，如线性核函数、多项式核函数、高斯核函数等，从而实现非线性分类。

**Q18.** The \_\_\_\_Perceptron\_\_\_\_\_\_ algorithm trains a single-layer neural network. **Q19.** ID3 employs \_\_\_a top-down, greedy approach/the entropy measure\_\_\_ to build a decision tree.

**Q20.** One of the parameters of kNN is the \_\_\_\_\_\_the number of neighbors (k)/the value of k\_\_\_ .

**Q21.** Logistic regression and Perceptron work best if the examples are \_\_\_linearly\_\_ separable. **Q22.** The parameter k in the clustering algorithm k-means is the: the number of clusters to be formed

**Q23.** The optimal value of the parameter k in the k-means clustering algorithm can be determined with the \_\_\_\_elbow\_\_\_\_ method.

**Q24.** \_\_\_Feature selection \_\_\_\_\_\_\_ methods learn which features best contribute to the accuracy of the model while the model is being trained.

There are two main types of feature selection methods: filter methods and wrapper methods.

**Q25.** \_\_\_Filter\_\_\_ feature selection methods apply a statistical measure to assign a scoring to each feature.

**Q26.** \_\_wrapper\_\_\_\_ methods consider the selection of a set of features as a search problem, where different combinations of features are prepared, evaluated and compared to other combinations.

**Q27.** The Hughes effect refers to the observation that with a fixed number of training examples ,the accuracy of a classifier first:

休斯效应是指观察到在训练样例数量固定的情况下，分类器的准确率随着特征空间维数的增加先增加后减少。

**Q28.** In collaborative filtering, the decomposition of the user-item interaction \_\_\_\_matrix\_\_\_\_ into the product of two lower dimensionality \_\_\_\_matrices\_\_\_\_\_ is known as matrix factorization

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ .

**Q29.** Content-based filtering and collaborative filtering are \_\_\_recommendation \_\_ approaches.

**Q30.** In text mining, the tf-idf weighting scheme is used for reducing the weight of words occurring:

在文本挖掘中，tf-idf（term frequency-inverse document frequency）加权方案用于降低语料库中频繁出现的词的权重，但对于区分文档没有信息。 tf-idf 加权方案是针对语料库中每个文档中的每个单词计算的，并同时考虑了单词在文档中的频率 (tf) 和单词在整个语料库中的频率 (idf)。

具体来说，文档中单词的 tf-idf 权重计算为其词频 (tf) 和逆文档频率 (idf) 的乘积。 tf 衡量单词在文档中出现的频率，而 idf 衡量单词在语料库中的稀有度。 因此，在文档中频繁出现但在语料库中很少出现的词被赋予高 tf-idf 权重，而在文档和语料库中频繁出现的词被赋予低 tf-idf 权重。

tf-idf 加权方案用于文本分类和聚类任务中的特征选择，它有助于识别最具辨别力的词以区分类别或聚类。

**END** **OF** **SECTION** **A**

**SECTION** **B:** **Problem** **Solving** **–** **choose** **to** **answer** **either** **Q31** **or** **Q32** **20** **marks**

**Q31.** **(20** **marks)** Consider the **training** and **test** **datasets** in Tables 1 and 2, respectively.

**Table** **1.** Training dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Education** | **Income** **(in** **thousands)** | **Property** | **Loan\_Status** |
| Primary | 41 | Rural | Yes |
| Primary | 20 | Rural | Yes |
| Primary | 25 | Urban | No |
| Secondary | 50 | Rural | No |
| Secondary | 31 | Urban | Yes |
| Secondary | 150 | Urban | Yes |
| Tertiary | 45 | Rural | Yes |
| Tertiary | 30 | Urban | Yes |

**Table** **2.** Test dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Education** | **Income** **(in** **thousands)** | **Property** | **Loan\_Status** |
| Secondary | 58 | Urban | Yes |
| Secondary | 35 | Rural | Yes |
| Primary | 39 | Rural | Yes |
| Tertiary | 80 | Urban | Yes |
| Tertiary | 20 | Rural | Yes |
| Primary | 60 | Urban | No |

a. **(5** **marks)** Draw a boxplot for all values in column **Income** in both Table 1 and Table 2, i.e., the array of values [20, 20, 25, 30, 31, 35, 39, 41, 45, 50, 58, 60, 80, 150]. Calculate *Q1*, *Q2*, *Q3* and IQR. Are there any outliers in the array and why?

b. **(12** **marks)** Replace the values in column **Income** by two categories:

**High** (for Income >= 40) and **Normal** (for Income < 40). Then apply the ID3 algorithm to the **training** **dataset** to build a decision tree for predicting the value of **Loan\_Status**.

c. **(3** **marks)** Calculate the accuracy of the decision tree you have built on the **test** **dataset**

in Table 2. Also calculate the precision and the recall for class **Yes** on the **test** **dataset**. **Hints:**

**Boxplot:** The box in a boxplot extends from Q1 to Q3, known as interquartile range (IQR). Any values smaller than Q1 − 1.5 ∗ (Q3−Q1) or larger than Q3 + 1.5 ∗ (Q3−Q1) are considered outliers and are drawn as separate dots. The two whiskers extend from Q1 and Q3 to the smallest and the largest values within 1.5 ∗ (Q3−Q1), respectively.

**Info** **function:** Let (a + b + c) = p and (d + e) = q . Then:

info[a, b, c] = − () × log2 () − () × log2 () − () × log2 ()

info([a, b, c], [d, e]) = () info[a, b, c] + () info[d, e]

**accuracy** = (TP + TN)/(TP + FP + TN + FN), **precision** = TP/(TP+FP), **recall** = TP/(TP+FN)

**Q32.** **(20** **marks)** Consider the user-movie rating matrix in Table 3, as well as the user-user similarity matrix in Table 4 and the movie-movie similarity matrix in Table 5.

**Table** **3.** User-movie rating matrix.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Movie1** | **Movie2** | **Movie3** | **Movie4** | **Movie5** | **Movie6** |
| **User1** | 4 |  | 3 | 5 |  |  |
| **User2** |  |  | 3 |  | 5 |  |
| **User3** | 5 | 4 |  | 4 |  |  |
| **User4** |  | 2 |  |  | 5 | 4 |
| **User5** |  |  | 4 |  | 4 |  |
| **User6** | 4 | 5 |  | 4 |  | 3 |

**Table** **4.** User-user similarity matrix.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **User1** | **User2** | **User3** | **User4** | **User5** | **User6** |
| **User1** | 1.00 | 0.22 | 0.75 | 0.00 | 0.30 | 0.63 |
| **User2** | 0.22 | 1.00 | 0.00 | 0.64 | 0.97 | 0.00 |
| **User3** | 0.75 | 0.00 | 1.00 | 0.16 | 0.00 | 0.91 |
| **User4** | 0.00 | 0.64 | 0.16 | 1.00 | 0.53 | 0.40 |
| **User5** | 0.30 | 0.97 | 0.00 | 0.53 | 1.00 | 0.00 |
| **User6** | 0.63 | 0.00 | 0.91 | 0.40 | 0.00 | 1.00 |

**Table** **5.** Movie-movie similarity matrix.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Movie1** | **Movie2** | **Movie3** | **Movie4** | **Movie5** | **Movie6** |
| **Movie1** | 1.00 | 0.79 | 0.27 | 0.98 | 0.00 | 0.32 |
| **Movie2** | 0.79 | 1.00 | 0.00 | 0.71 | 0.18 | 0.69 |
| **Movie3** | 0.27 | 0.00 | 1.00 | 0.34 | 0.65 | 0.00 |
| **Movie4** | 0.98 | 0.71 | 0.34 | 1.00 | 0.00 | 0.32 |
| **Movie5** | 0.00 | 0.18 | 0.65 | 0.00 | 1.00 | 0.49 |
| **Movie6** | 0.32 | 0.69 | 0.00 | 0.32 | 0.49 | 1.00 |

In both user-based and item-based neighborhood collaborative filtering below, consider **only** **the** **top** **2** most similar users to **User1** and the **only** **top** **2** most similar movies to the movies previously rated by **User1**, respectively.

a. **(10** **marks)** Apply the user-based neighbourhood collaborative filtering method to calculate movie recommendation scores for **User1**. Which one of the movies not rated by **User1** yet would you recommend to **User1** first?

b. **(10** **marks)** Apply the item-based neighbourhood collaborative filtering method to calculate movie recommendation scores for **User1**. Which one of the movies not rated by **User1** yet would you recommend to **User1** first?

**END** **OF** **EXAM**