國立成功大學
111學年度碩士班招生考試試題

編　號：261

系　所：數據科學研究所

科　目：計算機概論

日　期：0219

節　次：第 2 節

備　註：不可使用計算機

編號： 261

國立成功大學 111 學年度碩士班招生考試試題

系　　所：數據科學研究所

考試科目：計算機概論

考試日期：0219，節次：2

※ 考生請注意：本試題不可使用計算機。 請於答案卷(卡)作答，於本試題紙上作答者，不予計分。

1. (10pt) What is the definition of regularization in machine learning and deep learning? Please give a specific example to explain the regularization.

2. (5pt) What are the major differences between Depth-First Search and Breadth-First Search in the graph?

3. (10pt) In the following Python code, explain
   3.1. (5pt) Please specify the name of such operation (the function within a function) in Python.
   3.2. (5pt) Please give the output of two function-calls.

```
def print_func_name(func):
  def wrap():
      print("Function Name:'{}'".format(func.__name__))
      func()
  return wrap
@print_func_name
def add():
    print("Note")
@print_func_name
def sub():
    print("Stop")
if __name__ == "__main__":
    add()
    sub()
```

4. (10pt) Shortest path. Please use the Dijkstra algorithm to find the shortest path with start point (a) for Figure 1. Note that the edge weight should be taken into account in this problem.
5. (10pt) Write a function to find the longest possible "simple" path between any two vertexes in that graph (A simple path is a non-cycle path) for Figure 1. It is unnecessary to consider edge weight, only path length.

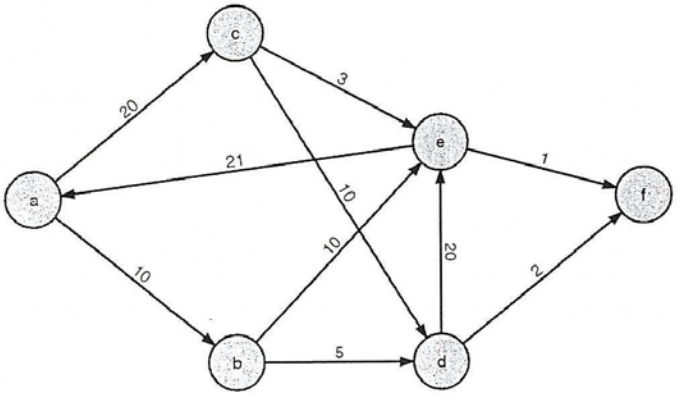

Fig. 1. Sample Graph.

6. (22pt) Answer the following questions on data science.
   6.1. (4pt) What is logistic regression?
   6.2. (4pt) How to store the graph structure in data structure? Please give at least two solutions for storing the graph structure.
   6.3. (4pt) For the class imbalance issue (say, relatively few samples in some classes), give the solution to resolve this issue.
   6.4. (10pt) Consider a time series with multivariant data. Usually, we need to partition the data into the training, validation, and testing parts. Please answer the following questions
      6.4.1. (5pt) What are the purposes of training, validation, and testing data, respectively?
      6.4.2. (5pt) What is the best way for data split? Which axis do we need to split? Time or Spatial?

國立成功大學 111 學年度碩士班招生考試試題

系　　所：數據科學研究所

考試科目：計算機概論

考試日期：0219，節次：2

7. (23pt) In the following statements, please specify if the statement is **True** or **False**. If the statement is **True**, explain why it is True. If it is **False**, give the correct answer or explain **why**.
   - 7.1. (2pt) The only imputation method for missing data is deletion.
   - 7.2. (3pt) Outlier detection is usually used to remove the missing values.
   - 7.3. (3pt) It is well-known that the overfitting issue will cause performance degradation in the training phase.
   - 7.4. (3pt) A decision tree can be regarded as an explainable classifier/regressor.
   - 7.5. (3pt) All the categorical data should be transformed into a one-hot-encoding form for the decision tree.
   - 7.6. (3pt) XGBoost, CatBoost, and LightGBM have similar time and space complexity.
   - 7.7. (3pt) Huffman coding can be used on data compression.
   - 7.8. (3pt) K-means clustering is an unsupervised learning method.

8. (10pt) What is the Closure in Python programming language? Please define it and give a pseudo-code for the Closure form.