Machine learning CA 3: project

Dr. Abhishek kaushik

19th, March

Instructions

==========

1. Group Work (50 marks) : Question 1.1
2. Individual Work (35 marks): Question 1.2 and 1.1.1
3. Interview (15 marks) (Last week of Semester (Preferably))
4. Each student must submit two executable notebooks with the file names
5. firstname_KNN.ipynb (Convert them into PDF) and firstname_CA3_KMEAN.ipynb (COnvert them into PDF)(individual work)
6. You will use the datasets as indicated in subsequent questions.
7. You have to create separate code cells andor text cells for different questions.
8. Interview questions will be based on overall ML concepts used and covered in this CA.

1 Introduction

The overall lab is divided into two components: such as case study-based, and lab-based. Please answer all questions. This complete assessment is 50% of the overall module. this is a group task.

1.1 Case study Based Problem

(Group work)

==========

(50 marks)
Please read the following cases.

1. Code Smells for Machine Learning Applications (study 1)
   `https://arxiv.org/pdf/2203.13746.pdf`

2. Time and the Value of Data(Study 2)
   `https://arxiv.org/pdf/2203.09118.pdf`

These above cases illustrates, leading innovations and pioneering into areas of artificial intelligence that paved the way for creating an AI- centric world. Please answer the following questions based on your learning from the above studies.

1. How can learning from Study 1 enable us to identify and then exploit an opportunity in the Machine learning field? (500 words) **(10 marks)**
2. Write your thoughts about "Opportunities implementing the process discussed in Study 1." (500 words) **(15 marks)**
3. In the context of Study 2, what complement the theoretical results with an experiment? (500 words) **(10 marks)**
4. Briefly summarize your learning from Study 2 and how you can improve the overall process. (500 words) **(15 marks)**

1.1.1 Individual work contributions

(10 marks)
==========
Provide details of individual work contributions in Task 1 (case study problem). And submit it via Lab based Problem.

For an example, details about the specific section you have study in the case studies and which question, you have answer and how you distribute your task among your team.

1.2 Lab Based Problem

(Individual work) ==========
(25 marks)
Please choose any data from `https://data.fivethirtyeight.com/` that you believe is acceptable for KMean and KNN, describing each stage of EDA and elaborating your learning about the data set in the context of minimizing error in the machine learning model.

- Code **(15 marks)**
  1. Perform EDA and choose any one automated feature selection method.
  2. Perform meaningful observations through visualisation on the obtained k-means and KNN
  3. Clusters graph to visualise and use elbow method to select number of centroids.
  4. Visualise and use elbow method to select number of K (KNN).
- Justification and Explanation **(10 marks)**

Please share your code (Python colab notebook). Please download your colab file and convert it into PDF and upload in the moodle.
Instructions to convert the colab file into pdf. `https://stackoverflow.com/questions/52588552/google-co-laboratory-notebook-pdf-download`
Choose any appropriate method according to you.
Please justified each step with data set. Please reflect your learning from above exercise (200 words).

Please submit this Assignment as per deadline. Plagiarism is strictly forbidden. Late submissions and plagiarism will be dealt with as per institute policy.