Project Proposal – Milestone 1

Qi Li

li.qi8@northeastern.edu

Percentage of Effort Contributed by Student : 100%

Signature of Student : Qi Li

Submission Date: Jan. 22. 2022

● Introduction

1) What is the business problem you are going to work on in this course?

From this class, I'm going to grasp the knowledge of how to use python to do machine learning and data mining work.

2) Explain a few sentences about the business problem

Firstly, I'm going to grasp the basic knowledge from class to learn how to use python (for example some packages) to do data mining work. I checked the course syllabus and I found that the most challenge for me in this course may be the *Discriminant analysis* and *Association rules & collaborative filtering* in chapter 12 and 14. Since I have learned linear regression, k-means regression, and Naive Bayes classifier from class IE5374 but these two chapters are totally new knowledge for me.
Secondly, I'm going to finish the homework every time to make sure I can catch up with professor Arasu.
Also I'll finish my own project to use the knowledge that the professor taught to make sure I fully grasped the knowledge and can finish the project by myself.

3) Explain one or two sentences on the data set

I'm going to use Letter Recognition Data Set to finish my project. This data set is about classifying 26 English capital letters. The character images were based on 20 different fonts and each letter was randomly distorted to 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes. This data set has 17 columns (the first column is the capital letter from A to Z and the rest 15 columns are the numerical attributes). They trained on the first 16000 items and then use the resulting model to predict the letter category for the remaining 4000 items.

● Problem statement

1) Write down your hypothesis.

I'm going to use Letter Recognition Data Set from UCI to complete my project. By testing different models to the data set, I can find out which model applied to this data set best.

2) How your machine approach or automation is going to solve the business use case?

I'll separate the data set into training set and validation set: 16000 for training set and 4000 for validation set. Then build a variety of classification models (such as

SVM, MLP, KNN models) using the training data set. After then, I'll use the validation data set to assess performance of the models and pick the best one.

3) What does the business get out of this solution?

By classify these 20000 unique stimuli from 20 different fonts, we can better distinguish the 26 capital letters in the English alphabet.