Kunhao Ni

Apr 9, 2024

**Classifier of Hand Labeled Data: the Use of LLM**

In this document, I’ll discuss how I deployed LLM to implement the classifier function in detail.

The LLM I’m using for this task is Open AI Chat GPT 4.0. The purpose of using LLMs is to do select an appropriate method and model basing on the data set we are using. I created 2 chats: the first chat aimed to discover the data set properties and potential solutions; the second chat is aimed to generate a prototype code for this task.

In the first chat (<https://chat.openai.com/share/2ea7fe1a-3b69-4bf4-951d-fc3aaa7f9e54>), I started with conceptual questions:

**文本, 信件

描述已自动生成**

**文本, 信件

描述已自动生成**

According to the response, I realize that this problem could be solved with building a statistical model as a prediction classifier.

Then, I upload the data set code book and raw csv data to it. The purpose is to discover if there are specific features I need to take into consider before I start the next step model fitting:

图形用户界面, 文本, 应用程序, 信件

描述已自动生成

文本, 信件

描述已自动生成

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成

文本, 信件

描述已自动生成

Till here, the process of attaching this problem is clear:

1. Handle the data set, which includes dealing with the missing data. Possible methods include imputation and deletion.
2. Select a model to use for training classifier.

Then, I tried to let the LLMs generate code snippets for different method (K-nn and Missing Forest) in python to investigate which could be the potential solutions:

图形用户界面, 文本, 应用程序

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成

(I’ve attached the python file into the email)

However, the result generated by these methods are incorrect: the classifier is classifying all the missing value as Yes.

图形用户界面, 应用程序, 表格, Excel

描述已自动生成

I dive into the data and LLM’s code to find the reason. After some attemps, I realize this error could be resulted by the large numbers of missing data:

日程表

描述已自动生成

Ideally, we want to drop the columns with more than 10% missing data. However, most of the columns has missing more than 10% of data. Thus, we should increase the threshold. Otherwise, the most columns of the data with be dropped.

Also, I notice the problem of the default encoding method generated by LMMs: it is converting categorical data into numerical data by the labelEncoder() function. Machine learning models may misinterpret that there is some sort of hierarchy in them.

After reading relative essay and documents, I decide to process the data in following manner:

1. Dropping col with more than 20% missing data.
2. Splitting them into training set and prediction set basing on whether the response in ‘Directors102b7’ exists or not.
3. Using One-Hot Encoding to convert the type of data.
4. Using Random Forest method to fit the data.
5. Using cross-validation of examine for the model’s accuracy.

Basing on insights I learned from above attempts, I created another chat(https://chat.openai.com/share/c2079a02-3452-4d78-b3e4-843e3fd69cd0) to generate code prototype in r with mentioning all above problems:

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

文本

描述已自动生成

文本, 信件

描述已自动生成

文本

描述已自动生成

图形用户界面, 文本

描述已自动生成

文本

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成文本

描述已自动生成

Details steps are described in the classifer.pdf which is knitted from the original r file. The reason I’m transfer from python to R is because R is more powerful when I need to work with a loaded dataset frequently. Combining the prototype generated by LLMs, I rewrite the code in r to create a classifier for the data set.

Reference: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3S1XRN https://chat.openai.com/share/2ea7fe1a-3b69-4bf4-951d-fc3aaa7f9e54 https://chat.openai.com/share/c2079a02-3452-4d78-b3e4-843e3fd69cd0 https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00516-9#Sec32 https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/ https://www.geeksforgeeks.org/ml-one-hot-encoding/