دانشگاه شهید بهشتی

دانشکده علوم پایه

استاد : دکتر حسین حاجی ابوالحسن

دانشجو : محمد عضدی

پروژه درس یادگیری ماشین

مقطع کارشناسی ارشد گرایش علوم داده ها

تیر ماه ۱۳۹۹

معرفی اجمالی پروژه :

این پروژه در مورد اجرای روش های یادگیری ماشین بر روی داده هاست(داده ها از سایت kaggle.com گرفته شده است). در این پروژه مجموعه داده (Data set) در نظر گرفته شده شامل داده های یک بانک در کشور پرتغال است که یک کمپین جدید جهت سرمایه گذاری راه اندازی کرده وقصد دارد با توجه به داده هایی که از افراد دارد پیشگویی کند:

که آیا افراد در این کمپین شرکت می کنند یا خیر.

مراحل و تشریح روشها :

در طی این پروژه که در فضای ژوپیتر نوت بوک انجام میشود، ابتدا یک سری دیداری کردن و پیش پردازش داده ها را انجام میدهیم که بیشتر به صورت نمودارهای فراوانی نشان داده میشود. ودر ادامه راستای کاهش بعد قدم برمیداریم.

با بررسی بیشتر داده ها و متغیرها، تصمیم به حذف یا اضاف کردن بعضی از متغیرها میگیریم و در نهایت با اعمال مدلهای زیر بروی داده ها میپردازیم:

1: Logestic regression

2: Decision tree

3: Random Forest

بررسی نتایج:

برای انتخاب بهترین مدل خود برای بررسی و پیش بینی داده هایمان استفاده میکنیم. که در نهایت مدل random forest بهترین مدل انتخابی ما خواهد بود.

Variable ویژگی ها

این داده شامل ۲۱ ویژگی(ستون) و ۴۱۱۸۸ مشاهده(سطر) است. حال ویژگی ها را اندکی توضیح می دهیم:

Bank client data:

- Age (numeric)
- Job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
- وضعیت تاهل Marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed)
- Education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
- قرض داشتن Default : has credit in default? (categorical: 'no', 'yes', 'unknown')
- وام منزل Housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
- Loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Related with the last contact of the current campaign:

- Contact: contact communication type (categorical: 'cellular','telephone')
- Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- Day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- Duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- Previous: number of contacts performed before this campaign and for this client (numeric)
- Poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

- Emp.var.rate: employment variation rate - quarterly indicator (numeric)
- Cons.price.idx: consumer price index - monthly indicator (numeric)
- Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- Euribor3m: euribor 3 month rate - daily indicator (numeric)
- Nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

- y - has the client subscribed a term deposit? (binary: 'yes', 'no')

منابع:

Understanding Machine Learning: From Theory to Algorithms, by Shai Shalev-Shwartz and Shai Ben-David

Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd Edition) by Aurelien Geron

link:

https://www.kaggle.com/henriqueyamahata/bank-marketing