ML Use case for Practice

Abstract:

To understand and revise some useful machine learning concepts and tricks, Here we are going to solve a simple machine learning use case. To develop an end to end ML cycle, we will start with data EDA -> data preprocessing -> model creation -> model deployment

This is a four week assignment need to be completed individually.

**Week 1:**

1. Data understanding and EDA
2. Create the dataset from raw arff file.
3. Locate features and labels
4. Perform classical statistical and graphical analysis on the data
5. Understand the distribution and correlation between various features (heatmap and corr matrix)
6. Record the insights from the data
7. Create a comprehensive document of understanding.

**Week 2:**

1. Data preprocessing
2. Model creation – scikit learing/TF/pytorch (use any)
3. Feature importance
4. Hyperparameter tuning
5. Creating a final version of the model
6. Model prediction script
7. Documentation

**Week 3:**

1. Creating flask api for rest endpoint
2. Creating simple UI to deploy the model as a web app in localhost
3. Model deployment and testing

**Week 4:**

1. Deploying the model using docker image
2. Creating the dockerfile
3. Running the docker container and running the same webapp on the localhost
4. Testing
5. Documentation

**Dataset description:**

Citation Request:

This dataset is public available for research. The details are described

in [Moro et al., 2011]. Please include this citation if you plan to use this database:

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct

Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.),

Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121,

Guimarães, Portugal, October, 2011. EUROSIS.

Bank Marketing

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

The classification goal is to predict if the client will subscribe a term deposit (variable y).

Available at: [pdf] http://hdl.handle.net/1822/14838

[bib] http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt

1. Title: Bank Marketing

2. Sources

Created by: Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) @ 2012

3. Past Usage:

The full dataset was described and analyzed in:

S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application

of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation

and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

4. Relevant Information:

The data is related with direct marketing campaigns of a Portuguese banking institution.

The marketing campaigns were based on phone calls. Often, more than one contact to the same client

was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

There are two datasets:

1) bank-full.csv with all examples, ordered by date (from May 2008 to November 2010).

2) bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv.

The smallest dataset is provided to test more computationally demanding machine learning algorithms

(e.g. SVM).

The classification goal is to predict if the client will subscribe a term deposit (variable y).

5. Number of Instances: 45211 for bank-full.csv (4521 for bank.csv)

6. Number of Attributes: 16 + output attribute.

7. Attribute information:

For more information, read [Moro et al., 2011].

Input variables:

# bank client data:

1 - age (numeric)

2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur",

"student","blue-collar","self-employed","retired","technician","services")

3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced"

means divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign

(numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

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

8. Missing Attribute Values: None