CluePoints - Case

Nicolas Huet

Introduction

As part of the recruitment process, two machine learning exercises have been selected to assess your skills in this field. First exercise is a binary classification use case with structured input data. Second exercise is a binary classification use case with unstructured data, i.e. text.
To solve these exercises, you can use any tool or software that you like. The goal is not to reach the best performances but to show if you are able to properly apply machine learning algorithms, write a clean code, understand the techniques that you use, etc. These exercises are just a pretext to show your expertise in machine learning so don't hesitate to use complex techniques even if they are overkill.
You should be able to

- Shortly present your findings;
- Assess the performances of your algorithm;
- Describe the mathematical concepts behind your algoritm;
- Justify your implementation choices.

Exercise 1

Description

This exercise aims to predict whether or not a person that made a medical appointment will show up or not. For this, you have some patient data with a label telling no-show or not, i.e. if the patient missed the appointment ("Yes") or came ("No"). Since some patient made several visits, you can have several records per patient. Feel free to handle each record individually or to handle each patient as a sequence. Suggestion: in the later case, make the prediction only for the last visit.

Data

The training data are stored in the file `medical_appointment_train.csv`.
Here is an explanation of the features:

- `PatientID`: Identification of a patient;
- `AppointmentID`: Identification of each appointment;
- `Gender`: Male or Female;
- `ScheduledDay`: The day someone called or registered the appointment;

- AppointmentDay : The day of the actual appointment;
- Age : Age of the patient;
- Neighbourhood : Where the appointment takes place;
- Scholarship : True or false;
- Hypertension : True or false;
- Diabetes : True or false;
- Alcoholism : True or false;
- Handicap : True or false;
- SMS_received : 1 or more messages sent to the patient;
- No-show : True or false. This is the target.

Exercise 2

Description

The second exercise focus on a common natural language processing use-case: sentiment analysis. The goal is to predict whether or not the review about a given movie is positive or negative. The training set consists of several thousands of movie reviews that have been labeled with 0 (negative review) or 1 (positive review).

Data

The training data are stored in the file movie_review_train.tsv.
Here is an explanation of the features:

- id: Unique ID of each review;
- sentiment: Sentiment of the review. 1 for positive reviews and 0 for negative reviews;
- review: Text of the review.