LELEC2870 Machine Learning Project: Predicting a person’s weight based on eating habits and physical activities

Friday 15th October, 2021

Introduction

Machine learning methods can be used to solve many practical problems in a wide range of applications such as weather forecast, customer clustering, medical diagnostics, spam blocking, financial time series prediction or signal de-noising, . . . In this project, you will apply machine learning to predict the weight (and bmi) of people based on their eating habits and physical activities.

Data

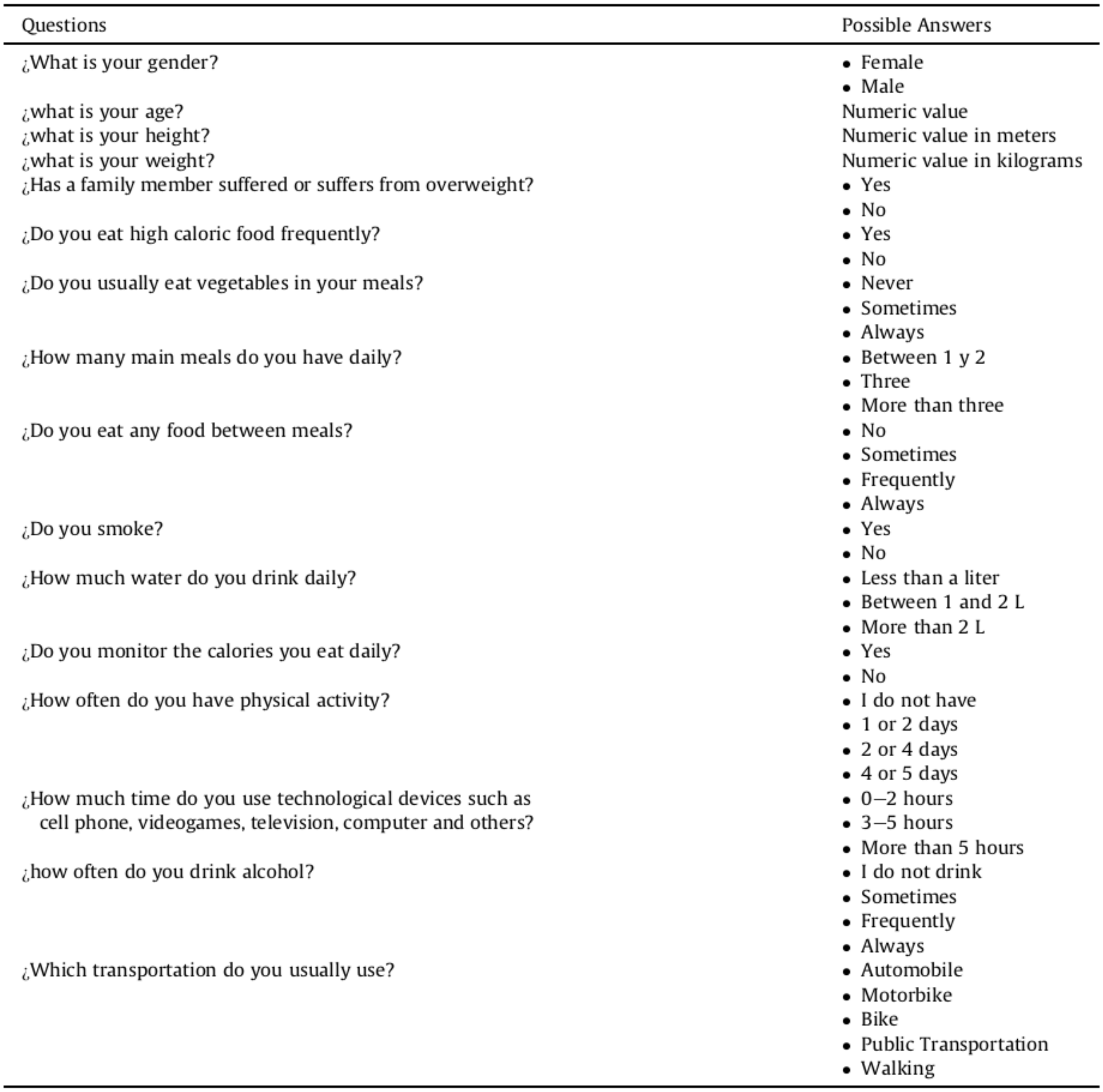
The data you’ll be working with was garnered from a study conducted in Colombia, Peru and Mexico [1]. Participants were asked a multitude of questions regarding their eating habits, their mode of transportation to work etc.

You’ll find the data on the course website in 3 csv files called X1.csv, Y1.csv and X2.csv. The first one is the labeled dataset and the second will be used to make your prediction. Paste these files in your working directory. The data can be loaded in your workspace by running the following commands:

```
import pandas as pd

# Use pandas to load into a DataFrame
#     Y1.csv doesn’t have a header so
#     add one when loading the file
X1 = pd.read_csv(”X1.csv”)
Y1 = pd.read_csv(”Y1.csv”, header=None, names=[’weight’])

# If you prefer to work with numpy arrays:
X1 = X1.values
```



Questions	Possible Answers
¿What is your gender?	• Female • Male
¿what is your age?	Numeric value
¿what is your height?	Numeric value in meters
¿what is your weight?	Numeric value in kilograms
¿Has a family member suffered or suffers from overweight?	• Yes • No
¿Do you eat high caloric food frequently?	• Yes • No
¿Do you usually eat vegetables in your meals?	• Never • Sometimes • Always
¿How many main meals do you have daily?	• Between 1 y 2 • Three • More than three
¿Do you eat any food between meals?	• No • Sometimes • Frequently • Always
¿Do you smoke?	• Yes • No
¿How much water do you drink daily?	• Less than a liter • Between 1 and 2 L • More than 2 L
¿Do you monitor the calories you eat daily?	• Yes • No
¿How often do you have physical activity?	• I do not have • 1 or 2 days • 2 or 4 days • 4 or 5 days
¿How much time do you use technological devices such as cell phone, videogames, television, computer and others?	• 0–2 hours • 3–5 hours • More than 5 hours
¿how often do you drink alcohol?	• I do not drink • Sometimes • Frequently • Always
¿Which transportation do you usually use?	• Automobile • Motorbike • Bike • Public Transportation • Walking

Figure 1: Which questions were asked to the participants? (the weight column has of course been removed and put as target in your csv files)

In the table above you can find which questions the participants got during the study. The order of the features in the dataset is the same as in this table, although some column names may look like a bizarre fit. Also do watch out for some strange encoding choices: "*How often do you perform a physical activity?*" is column "FAF", which contains integer numbers, and e.g. 0 refers to 'zero activity', while 3 refers to '4 or 5 days'. The data engineering aspect of the dataset should not be undervalued!

Instructions

The project is realized by groups of two or alone. It is composed of different aspects as specified below.

Data Engineering

You will need to re-encode some categorical and integer variables differently (as discussed above). Some features may be useless or redundant, you'll need to evaluate this as well. You may also look at augmenting the dataset in various ways (since you only dispose of 250 samples) or weighting some samples more than others, which may be useful to generate a result that'll make sense regarding the balanced metric used to evaluate your predictions (described below).

Model

You will build regression models that predict the number of shares of an online article. You can use any of the methods seen during the lectures. We expect you to, at least, implement linear regression *as a baseline*, KNN[1], an MLP and one other non-linear method (you can chose one outside those seen during the lectures).
Feature selection and model selection need also be part of your work. Once again you're allowed to use any tools available (e.g. statistical tests seen during other courses). Pay attention to the fact that the model selection can require a lot of computation time. You are advised to explore the metaparameters space according to the time available.

Prediction

Once your model is properly selected and validated, you are asked to produce predictions Y2 on the data X2 for which we have kept secret the corresponding targets. This prediction vector should be uploaded on Moodle in a csv file named "Y2.csv" that contains **one line per prediction and no header**, no quotation marks around your numbers either. Check that your format is correct by opening it with a text editor and compare it to "Y1.csv". At the *tail end* of this file you will add an additional number which is the *estimated performance of your model on the unseen data.* (For the first deadline this value may be set to zero)

While predicting the weight of people is inherently a regression problem, it is also important to look at the different weight-classes with respect to bmi (= weight/height^2):

- 00.0-18.5 : underweight
- 18.5-25.0 : normal
- 25.0-30.0 : overweight
- $\geq$30.0 : obese

[1] K-Nearest Neighbour : Regression model with metaparameter K that predicts the output of a sample as the mean of output of the K nearest neighbours in the features space.

To spice things up, your prediction criterion (lower is better) will be balanced according to the different weight-classes (see code below). This evaluation was used over RMSE to add a classification element to your work and forcing you to use a non-aligned loss. This also allows you to train your models to directly predict the bmi. Do you think this would work better?

In addition, **you will provide in your report, a confusion matrix of your best model (at least) on your validation data** with respect to these bmi categories and discuss the results. The following code can be copy-pasted into your working python script:

```
import numpy as np
def score_weight_class(bmi_pred, bmi_true, low, high):
    tol = 1
    vpred = (bmi_pred>=low-tol) & (bmi_pred<high+tol)
    vtrue = (bmi_true>=low) & (bmi_true<high)
    if vtrue.sum()==0:
        print(”no true samples here”)
        return 0
    rmse = np.sqrt(((bmi_true[vtrue]-bmi_pred[vtrue])**2).mean())
    rmse = rmse/(high-low+tol) # normalize rmse in interval
    acc = (vpred&vtrue).sum()/vtrue.sum()
    return rmse*(1-acc)

def score_regression(ytrue, ypred, height):
    bmi_pred = ypred/(height*height)
    bmi_true = ytrue/(height*height)

    scores = []
    for bmi_low, bmi_high in zip([0,18.5,25,30], [18.5,25,30,100]):
        scores.append(score_weight_class(bmi_pred, bmi_true,
                                         low=bmi_low, high=bmi_high))
    return np.mean(scores)
```

First deadline New this year, there will be an intermediate deadline to force you to start this project as soon as possible. For this first deadline you'll only need to submit a prediction file in the correct format (described above).
The point of this submission is to test that your framework at least partially works. You shouldn't try to obtain the best performance, even a baseline result with a basic linear regressor is ok! But carefully looking at the data and parsing it properly before this point is highly encouraged.

Report

You will produce a report documenting your technical choices and experimental results. **We do not need a course on the methods you use. We are more interested in what you did and why. Be straight and to the point!** Try to illustrate your results with graphics (with *legends* and *labeled axes*) and comment them. Be critical about what you observe and try to give a possible justification of the obtained results. Summarize your results and observations in a conclusion. A strict maximum of 7 pages (font of size 11 or larger) will be observed. Appendices might be included in the digital version that you'll submit on Moodle, but shouldn't be resorted to unless something really interesting was found. Remember to make your main observations very clear! All your figures and computation need to be reproducible by us running your implementation code on the provided data.

Programming languages

The programming language you will use is Python. You can use any toolbox/library available on-line. In particular, we strongly recommend using the `scikit-learn` library as it provides many useful implementations of standard machine learning approaches. For the MLP we recommend you use the one integrated into sklearn or `pyTorch`. `skorch` is a python package that links `pyTorch` with the `sklearn` syntax, some of you might find it handy, but don't hesitate to come to us with questions if you encounter bugs with these packages.

Agenda

- As soon as possible: Register your group (maximum two people) on the course website.
- Thursday 21/10 at 08h30: Q/A session #1
- Thursday 16/11 at 23h55: **first deadline** where you only submit a prediction file
  - A csv file called "Y2.csv" **no header line: one line per prediction values**.

  This submission is mandatory ! You have ample time to submit something functional, so *no* extensions will be given.
- Thursday 02/12 at 08h30: Q/A session #2
- Sunday 19/12 at 23h55: **final deadline** where you submit your work as 3 separate files
  - Your report (.pdf)
  - A csv file called "Y2.csv" **no header line: one line per prediction values**.
  - A *compressed* folder containing all scripts you wrote for the project. The code should be commented well enough and installation instructions about non-standard packages you used should be provided.

  This submission is mandatory ! You have ample time to submit something functional, so *no* extensions will be given.

Evaluation Criteria

- Respect of the instructions and deadlines
- Quality of the report and its defense (discussion)
- Your machine learning approach (data engineering, model choices and validation)
- Reproducibility of your results
- Consistency between the report, your implementation and your predictions

Please note that the performance of your models is not the most critical part in the evaluation. Your report + your points on the project question at the exam, will encompass 10 of the 20 points of your final mark.

Tips

Here is a list of advices for the project.

Before any analysis:

- Visualizing the data is always useful.
- Evaluate your model correctly
- Normalize or Standardize your data if necessary.
- Some outliers might be excluded from the learning set (if you decide to remove some observations, explain why you removed them).
- Categorical data has been encoded as integers. What does that mean for your algorithms? Is there perhaps a better way of encoding this?

You can discuss the project with other students, in fact, it is a great idea! You could compare your results to those obtained by other groups, but remember that it is not allowed to copy what others did...

We will be happy to answer your questions during the Q/A sessions or on appointment. Good luck !

References

[1] Fabio Mendoza Palechor and Alexis de la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, 25:104344, 2019.