Data Science Lab
Lab #9

Politecnico di Torino

Intro

The main objective of this laboratory is to put into practice what you have learned on regression techniques. You will work on a tabular dataset. In particular, you will try to build a regression model that is able to identify the price of an Airbnb apartment, given different information related to the listing.

Important note. For what concerns this laboratory, you are encouraged to upload your results to the competition we launched on our platform, even if the submission will not count on your final exam mark. You have to use the same personal key you already used for Lab 7. If you do not have a key yet, please write to lorenzo.vaiani@polito.it. Refer to Section 3 to read more about the competition.

Important dates

Start date: December 12, 2022 at 01:00 PM (CET)
Due date: December 20, 2022 at 00:00 AM (CET)
Open until: December 22, 2022 at 00:00 AM (CET)

Due date is a **strict deadline**.

1 Preliminary steps

1.1 Datasets

In this laboratory, you will use a publicly available dataset. Public Domain Dedication datasets constitute an extremely valuable asset for the data science community. If you want to know more about how they are distributed, refer to the CC0 licence.

1.1.1 New York City Airbnb Open Data

This public dataset is part of Airbnb, and the original source can be found on Inside Airbnb.

Each row of the dataset corresponds to an Airbnb listing in New York City, for the year 2019. As for the previous competition, the dataset has been divided in a Development set and an Evaluation one. You will find more about them later in the document.

Each file has an initial header line, containing the names of attributes at your disposal:

- id: a unique identifier of the listing
- name
- host_id: a unique identifier of the host
- host_name

- neighborhood_group: neighborhood location in the city
- neighborhood: name of the neighborhood
- latitude: coordinate expressed as floating point number
- longitude: coordinate expressed as floating point number
- room_type
- price: price per night expressed in dollars
- minimum_nights: minimum nights requested by the host
- number_of_reviews
- last_review: date of the last review expressed as YYYY-MM-DD
- reviews_per_month: average number of reviews per month
- calculated_host_listings_count: amount of listing of the host
- availability_365: number of days when the listing is available for booking

You can download the dataset at:

`https://github.com/dbdmg/data-science-lab/raw/master/datasets/NYC_Airbnb.zip`

1.1.2 Dataset tree hierarchy

The data have been distributed uniformly in two separate collections. Each collection is in a different file. The dataset archive is organized as follows:

- `development.csv` (Development set): a collection of listings **with** the price column. This collection of data has to be used during the development of the regression model.
- `evaluation.csv` (Evaluation set): a collection of listings **without** the price column. This collection of data has to be used to produce the submission file.
- `sample_submission.csv`: a sample submission file.

So far, you should be used to work, while developing your models, with training, validation and test sets. In this case, the Development data must be used to tune your hyper-parameters while you should consider the Evaluation portion as the actual test set.

2 Exercises

In this laboratory, you have a single regression task to carry out.

2.1 NYC Airbnb listing price regression

In this exercise, you will try to predict the price of an Airbnb listing in NYC, published in 2019, using several contextual information. To do so, your primary goal will be modeling, through a regression-based pipeline, the relationship between information on the listing (e.g. its geographical location, the reviews it received, or many other metrics you might figure out) and the price itself.
Once your model is complete, you will predict, for a set of listings whose price is unknown, how much would it cost to you spending one night at them.

Finally, you will be able to upload your regression results and participate to the lab competition.

1. Load the dataset from the root folder.
2. Focus now on the data preparation step. You should have noticed that the attributes that describe each listing are heterogeneous, both on the source (e.g. geographical, related to host, related to Airbnb, etc.) and on the type (e.g. numerical, categorical, date, etc.). Before continuing, take you your time to answer these questions:
   - which attribute (or set of attributes) you think could drive the price per night the most?
   - can you detect any irregularity in any attribute distribution?
   - if your regression model will fit on numerical data only, how could you handle categorical attributes?

   Transform your initial dataset following the ideas you draw out.
3. Once you have your final dataset representation, choose one regression model of those you know. Then, perform the classic training-validation pipeline on the Development dataset to identify the best set of hyper-parameters for your model. As you can read in Section 3.3, we will evaluate your results on the MSE score (Mean Squared Error). Hence, it is a reasonable option trying to optimize it on the Development set.
4. Assign a price value to each listing in the Evaluation set.
5. Define a function to generate a 2D scatterplot with the prices. The chart must be drawn as heatmap: use the latitude and longitude coordinates along the axes and the price value to assign a color to the point. Then, apply the function to the prices from the Development set and to the ones you predicted for the Evaluation set. From Section 1.1, you know that Development and Evaluation were generated with a uniform sampling on the initial listings. So, what should you expect on the map, if your regression were correct?
6. Upload your results to the submission platform. Head to Section 3 to know more about it.
7. Compile your final report and upload it to the "Portale della Didattica" as described in section 3.2.

3 Submitting you work

For this laboratory, you should upload two files to two different web sites. The first file contains the regression results, the second file contains a report on the experiments you carried out. The following sections provide further details on that.

3.1 Submit your classification results

To get your results evaluated, you have to upload a result file on our submission competition. The submission file has to be a `.csv` file formatted as follow:

```
Id,Predicted
10,120.31
123,100.00
21,523.22
345,652.02
42,225.41
...
```

As you can see, it must contain an header line and a row for each listing in the Evaluation collection. Each row must have two fields:

- the `Id` of the listing, as an integer number. Note that Ids can be in any order but they **must** match the Ids present in the Evaluation set (column `id`).
- the `Predicted` price value, as either a float or an integer number.

The submission platform is the same you used for Lab 7. Therefore, you have to use the same key. Please refer to the guide on the course website, to go through the submission procedure.
You can find the competition at http://trinidad.polito.it:8888

3.2 Upload your report (optional)

For those interested, it is possible to submit a report describing the proposed solution for this laboratory. If you would like to receive feedback about your work, please contact lorenzo.vaiani@polito.it, explicitly indicating your interest in receiving comments on the submitted report.

Please respect the following requirements:

- state clearly which pre-processing step characterized your final solution;
- describe which regression algorithm you used;
- describe which validation strategy you adopted and which are the best hyper-parameters you found on the Development set.
- comment on the heatmaps obtained in Point 5. Do they have the same "heat" distribution. If so, why? If not, why?

Please refer to the directions provided during the dedicated lecture to write your report. More specifically, you should use the IEEE conference LaTeX template. That is the template you will be using for submitting your final (graded) report, so you should get acquainted to it (and to LaTeX in general).

You can upload the file to the "Portale della Didattica", under the Homework section of the course. Please use as description: `report_lab_9`.

3.3 Evaluation

Your regression results will be evaluated on the MSE score.