Statistical Machine Learning

Hilary Term 2023

Group-Assessed Practical

Classification of Stars, Galaxies and Quasars

Description. This project aims at automatically classifying astronomical objects (stars, galaxies and quasars), based on a number of imaging and spectral measurements on that object. The dataset, which originates from the Sloan Digital Sky Survey, consists of measurements of 200,000 astronomical objects. Each object i is represented by an input vector $x_i = (x_{i1}, \ldots, x_{ip})$ where $x_{ij} \in \mathbb{R}$ represents the j'th feature/measurement of object i. $p = 63$ corresponds to the number of measurements per object. Each object may be of three different classes: GALAXY, STAR, or QSO (quasar). The objective is to (i) construct a classifier which, based on the features of an object, predicts its class, and (ii) to estimate its generalisation error under the 0–1 loss.

The dataset is split into a training set and a test set, each with 100,000 observations. For the training observations, you have access to both the inputs (X_train) and outputs (y_train). For the test set, you have only access to the inputs (X_test), and the objective is to predict their class.

Methods. You are free to use any machine learning technique you wish, as long as you describe clearly in the report all the steps and choices that you have made. While getting a good predictive performance for your method will be important, remember that you will be assessed based on the quality of your report; so explaining your steps and choices clearly and discussing all the issues you have faced in this practical will be essential. Besides explaining your final classifier, you should also describe some of the other techniques you have tried and include a brief description of the more computational aspects of your work. It is particularly important to discuss the potential advantages/disadvantages of the different methods considered, in terms of interpretability, computational cost, etc. You can use any programming language you wish (Python, R, etc.), and any available library/toolbox, as long as you understand and can describe the methods used. In Python, most of the methods covered in the course (except convolutional neural networks) are implemented in Scikit-learn. In R, many machine learning methods are implemented in the (meta)-package caret.

Report. The report has a limit of 2,500 words. Please be as concise as you can. You should work in teams of 4 participants. Remember to place your team name, which consists of the collated anonymous IDs of all group members, on the cover page of the report. Please name the pdf file of your submitted report using the list of anonymous IDs of all team members, e.g. P001-P002-P003-P004.pdf. Please include the code you used to get your final score as an appendix (this is not counting towards the 2,500 words limit). Make sure the code is readable (i.e. it contains comments explaining what you are doing). Only one student from each group is required to make the submission.

François Caron: caron@stats.ox.ac.uk

Submissions. Together with your report, you should also submit a csv file, containing the predicted class for the 100,000 observations in the test set. Your report should include an estimate of the classification accuracy on the test set of your prediction. The submission file (csv format) should contain two columns: Index and Class. The file should contain a header, followed by the 100,000 class predictions, and have the following format:

```
Index,Class
0,STAR
1,GALAXY
2,GALAXY
...
```

A sample submission file is available.

Files available:

- This pdf with the instructions
- Training inputs: `X_train.csv`
- Training outputs: `y_train.csv`
- Test Inputs: `X_test.csv`
- Sample submission file: `myprediction.csv`
- Sample Python code: `AssessedPracticalSamplePythonCode.ipynb`

Evaluation metric. The metric used is the classification accuracy (proportion of well-classified examples in the test set).

Sample Python code. A sample Python notebook is provided. The code loads the data, fits a 1-nearest neighbour on the training set and predicts the class in the test set. It then exports a csv file of the correct format. For your information, the 1-nearest neighbour classifier has a classification accuracy of about 96.6% on the test set. You can use this value as a benchmark, and a lower bound for the performance of your classifier; you should be able to achieve higher performances with other methods.

Deadline. The deadline to submit your pdf report and csv file is Wednesday 22 March noon (week 10).

Additional information about the features.

This section is just provided for your information. There is no expectation that you should have a good understanding of the physical quantities involved. For more information about the different measurements, see:

`https://skyserver.sdss.org/dr16/en/help/browser/browser.aspx#&&history=description+PhotoObjAll+U`

`https://skyserver.sdss.org/dr16/en/help/browser/browser.aspx#&&history=description+SpecObjAll+U`

The photometric letters U (ultraviolet), G (green), R (red), I (infrared) and Z correspond to a section of light of the electromagnetic spectrum, and are used as subscripts for the features (e.g. `sky_u`, `sky_g`, `sky_r`, `sky_i`, `sky_z`).

Name	Description
ra	Right Ascension
dec	Declination
rowv	Row-component of object's velocity
colv	Row-component of object's velocity error
sky	Sky flux at the center of object
psfMag	PSF magnitude
fiberMag	Fiber magnitude
fiberFlux	Fiber flux
petroMag	Petrosian magnitude
petroRad	Petrosian radius
psfFlux	PSF flux
airmass	Airmass at time of observation
extinction	extinction
psffwhm	FWHM
redshift	Final redshift
waveMin	Minimum observed (vacuum) wavelength
waveMax	Maximum observed (vacuum) wavelength
wCoverage	Coverage in wavelength
spectroFlux	Spectrum projected onto filter

François Caron: caron@stats.ox.ac.uk