Pattern Classification And Machine Learning : Project One, Group 16

Marc Chachuat, Lie He, Alfonso Peterssen

Due date : 31/10/2016

1 Introduction

In this report, we explain the strategy we chosed to deal with this first project. This strategy can be divided in three parts answering to three questions :

1. How to pre-process the data ?
2. How to train each model ?
3. How to choose the best model for the prediction ?

In the following, a section is dedicated to each of these questions.

2 Data pre-processing

Before even thinking to models and prediction, the first task of the data-scientist is to study his data and to pre-process them. The idea is essentially to remove the non relevant data, and to put the data set under a shape suitable for training the models.

2.1 Data cleaning

One of the first things we noticed in our training set, is that a lot of data were missing, replaced by -999 values.
To avoid a bias, we decided to replace all the -999 values by the mean over the clean values of the current feature.
Yet, after a lot of work we realized that this wasn't optimal, because some information was reliying behind the fact that some features were missing for a given sample.
We decided to add a binary feature indicating wether or not at least one feature was originally missing for the given sample.
This was concluding, as it increased our score.

2.2 Standardization

After the cleaning of the data, we used the classical method of standardization seen in the course in order to reduce the scale differences between the different values.

2.3 Selecting the relevant data

The data set of the project contained a large number of features for each sample.
This was a problem because it was considerably slowing our algorithm.
Furthermore, we were suspecting that not all the data were relevant. We then had to select some. Yet, we hadn't the necessary background in physic to do it arbitrarily.
We decided then to implement our own Principal Component Analysis. Running it on the training data set, it gives us a projector that projects all the data (validation, training and test sets) over a linear subspace of the data. Moreover, it enables us to choose the percentage of variance we wanted to keep.

2.4 Using polynomial basis

Finally, after some work, we concluded that the linear model was too poor to predict labels with a good accuracy. Then we decided to enrich it with a method seen in the course : the polynomial basis.
With the PCA we had reduced the dimension of each sample, which enabled us to make the label depends not only of the features, but also of power of the features.
Then, an other problem was to choose the right degree for polynomial basis. But this was related to model training.

3 Training of the models

As least squares method is a particular case of ridge regression, and logistic regression is a particular case of regularized logistic regression we decided to focus only on the ridge regression and on the regularized logistic regression.
For each of these models we had implemented training methods in the course. Yet, this was for a given lambda (regularization parameter) and a given degree for the polynomial basis.
We realized then a grid search over a wide range of lambdas and degrees, training our models for each couple of parameters, and measuring its performance over the validation set.
Finally, we kept for each model the couple of parameters ensuring the best performance.

4 Choice of the model and prediction

In the above section, we mean by "performance", the accuracy over the validation set of the model trained over a training set.
This was for us the best way to evaluate the good behaviour of a model, as comparing the losses of the different models wasn't very relevant. Moreover, taking the accuracy over the validation set (and not the training set) decreases the risk of overfitting.

Finally, to predict the labels for the submissions we just had to select the most performant of our two models by comparing the two optimal accuracy.