Challenge Otto from Kaggle

Robin Duraz - Jiaxin Gao
TC1 - Apprentissage

November 5, 2018

1 Challenge presentation

This challenge was proposed by the Otto Group, one of the worlds biggest e-commerce companies, selling products in multiple countries around the world. Analyzing and classifying products is proven be a difficult task, and in fact, identical products from different global infrastructures were classified differently.
Therefore, the ability to correctly analyze products depends on accurately clustering similar products.

The competition's dataset is composed of more than 200 000 products, each having 93 features. The goal of the challenge is to correctly classify products in 9 classes. Of this dataset, a bit more than 60 000 products are used as train set, while the rest is used as a test set for which we have to compute probabilites for each class. These probabilities are what is tested by the challenge.

The error on the challenge is calculated with the formula :
$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij})$

2 Methodology

First, we performed an analysis on the data. We saw class division, and small summaries of feature values, including minimum, maximum, and mean.
Then, we simply tried different basic models from sci-kit learn to see what performances they would achieve on raw data, then normalized and/or scaled data. From that, we chose to normalize and scale data for every algorithm we used.
We then tried simple improvements on the models we used, and chose three that had good results. The three algorithms we kept are XGBClassifier from xgboost, RandomForestClassifier and MLPClassifier from sklearn.
We then tried to further tune our models. We mostly did grid search in order to find good values for their parameters, and we also found out that using CalibratedClassifierCV from sklearn on the Random Forest algorithm improved massively its results.
In the end, we tried to average our three models with weights to find better solutions than for single models. We succeeded in slightly improving our final results.

3 Results