Some super concise and informative title
Data Analysis Project for
Machine Learning: Basic Principles

November 30, 2017

Abstract

Precise summary of the whole report, previews the contents and results. Must be a single paragraph between 100 and 200 words.

1 Introduction

Background, problem statement, motivation, many references, description of contents. Introduces the reader to the topic and the broad context within which your research/project fits

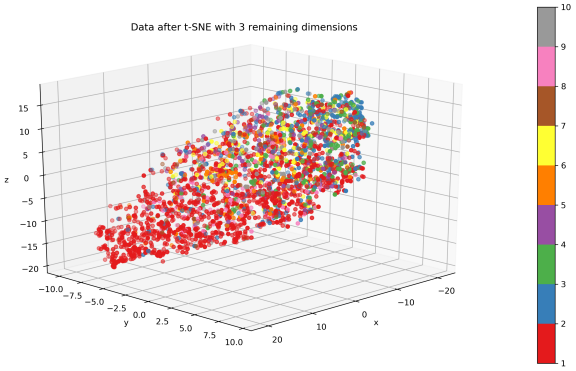
- *What do you hope to learn from the project?*
- *What question is being addressed?*
- *Why is this task important? (motivation)*

Keep it short (half to 1 page).

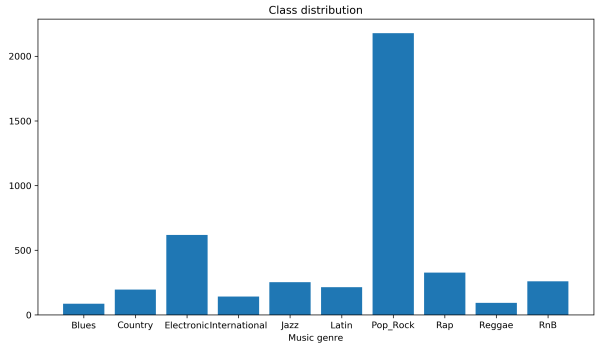
2 Data analysis

This competition is performed on two datasets, a training and a test dataset with 4.363 resp. 6.544 songs. Each dataset has a total of 264 features, which will be used for predicting one of 10 classes. The features con be grouped into the 3 main components of music: timbre, pitch and rhythm. The 10 classes are: Pop Rock, Electronic, Rap, Jazz, Latin, R&B, International, Country, Reggae and Blues.

In order to better visualise the training data we performed the *t-Distributed Stochastic Neighbor Embedding* (**t-SNE**) with 3 remaining dimensions. The result of this award-winning embedding is shown below:



It is also important to know the distribution of the given training dataset. The distribution is shown in the following picture:



This distribution shows very clearly that the training data is skewed, which means that the predictor will be able to generalise better for the majority classes and worse for those classes that have not many samples representing them.

3 Methods and experiments

3.1 Overall approach

To achieve best overall results we tried various machine learning techniques ranging from logistic regression (LogReg) and support vector machines (SVM) to naïve Bayes classifiers (NB) and neural networks (NN). Common for all machine learning techniques was, that we first standardised the training as well as the test data prior to performing any analysis. This step is crucial as it helps to reduce multicollinearity within the data and helps to improve the generalisation. This behaviour is backed up by comparing the accuracy of the prediction of the training data without standardisation (p) and with standardisation (p_{st}), which is summarised in table 1.

ML technique	p	p_{st}
LogReg	0.66	0.74
SVM	0.03	0.22
NB	0.46	0.52
NN	0.55	0.73

Table 1: Comparison of accuracy of training data without standardisation and with standardisation

In order to prevent our analysis against heavy overfitting we chose to implement cross-validation for all machine learning techniques. We associated randomly 20% of the training set to be the validation set. We then trained the model with the remaining 80% of the training set and validating the analysis on the validation set. This is done multiple times and averaged.

Explain your whole approach (you can include a block diagram showing the steps in your process).

- *What methods/algorithms, why were the methods chosen.*
- *What evaluation methodology (cross CV, etc.).*

4 Results

Summarize the results of the experiments without discussing their implications.

- *Include both performance measures (accuracy and LogLoss).*
- *How does it perform on kaggle compared to the train data.*
- *Include a confusion matrix.*

5 Discussion/Conclusions

Interpret and explain your results

- *Discuss the relevance of the performance measures (accuracy and LogLoss) for imbalanced multiclass datasets.*
- *How the results relate to the literature.*
- *Suggestions for future research/improvement.*
- *Did the study answer your questions?*

6 References

List of all the references cited in the document

7 Appendices

Additional information that is not essential to explain your findings, but supports your work. For example, source code, additional images, mathematical derivations, etc. If you include source code, don't include the whole code, focus only on the most important parts, for example, a function implementing a specific algorithm