Project protocol
Big Data Science
Fabienne Haot, Juliet van Rosendaal
2022-04-08

1 The classification task - Fabienne

The goal is to allocate fingerprints to five possible classes. We were given different datasets created by different data generating processes. For this reason we will create models for each dataset separately. This choice means we can not predict the class of fingerprints if the origin of the new data is not known.

1.1 Exploration of the data

The data includes real observations, so there will be issues with the data.
During the exploration, we will obtain general information about the distribution of the response variable and the datatype of the features. We will also form an idea of which issues are present in the data, and how big these issues in order to deal with them properly. We will check for missing values, concentrations, univariate and we will evaluate if there are continuous highly non normal variables present in the data.

1.2 Train-test split

The models will be trained, tuned and tested. We will first take away 20% from the data as test set, the remaining records will be used for 10 fold validation, assuming the number of observations is large enough. If the validation contains less than 20 observations, we will decrease the number of folds. Since we cannot assume that the dataset is balanced we will use stratified folding both for test as for train-validation splits.

1.3 Preprocessing

The identification and remediation of existing problems with respect to missing values, concentration points and outliers take place on the level of the (inner) training set. All remediations and transformations on the validation set will be performed based on the information obtained from the training set.
Missing We exclude the observations for which the majority of the columns is missing. For columns with missing data, we will use simple regression imputation.
Concentrations Concentration points representing more than 5% of the data will be flagged with a one hot feature. If more than 5% of a feature was missing, the imputation will give rise to a concentration.
Outliers Extreme outliers will be deleted. Even if they are part of the data, they should be classified by a human. Extreme outliers have distance $> 10\sigma$ from the center.
Tranformations Strongly skewed features will undergo a boxcox transformation. All features will be centered and standardized.

1.4 Determining different models

Unsupervised dimension reduction will be applied through principal components truncation, initially we will keep 95% of the variance on the trainingset. This will also remove highly correlated variables.
We will start with elastic net for the logistic model on principal components. For the tree model we use random forests, the nonlinear model will be a kernel discriminant analysis.

1.5 Evaluation of the best model on the validation set

Underlying scores will be evaluated through AUROC and BIC. Classification itself is based on softmax. Misclassification will be the deciding criterion since we assume each misclassification represents the same loss.

1.6 Final model

Application of the chosen trainingmodel on the test set. A confusion matrix will inform us about the misclassications made by the chosen model.

2 The regression task - Juliet

Given a benchmark dataset of 106,574 records of music tracks, consisting of 518 features, two supervised regression models are fit. The first model **m1** predicts **Y1**: track listens (between 0 and 543252) and the second model **m2** predicts **Y2**: album date released (between 1902-01-01 and 2021-03-01).

2.1 Data exploration

An exploratory analysis is performed using descriptive statistics and data visualization. Features are grouped according to their column names (chroma_cens, chroma_cqt, chroma_stft, mfcc, rmse, spectral_bandwidth, spectral_centroid, spectral_contrast, spectral_rolloff, tonnetz, zcr).

- For each target, an overview of a subsample of the data is visualized using an **interactive PCA plot**.
- **Mean**, **median**, **standard deviation**, and **quantiles** are reported for features and targets.
- The **distributions** of the features are visualized with **violin plots**. The **distributions** of the targets are visualized with **histograms**, on which a **density estimate** is plotted.
- The features are assessed on **covariance** and **correlation** using **heatmaps**.
- **Missing data** is reported for each variable and assessed on randomness.

2.2 Preprocessing

All data is preprocessed according to the following steps.

- Data is **split** in a **training** set of 80% of the data and a **test** set of 20% of the data.
- **Missing feature data** is imputed using the mean. **Missing target data** is inferred from other available metadata.
- **Outliers** are removed, data is **normalized** and **centered**. Target **Y1** is **binned** per 10 listenings and target **Y2** is **binned** per year.

Next to feature set **f1**, which contains all features, two more feature sets are created with **PCA** dimensionality reduction. For feature set **f2** PCA is applied per column name group, and for feature set **f3** PCA is applied on the total of features.

2.3 Models and evaluation

For both **m1** and **m2**, the same types of models are fit on each feature set. A simple **linear regression** will function as a point of reference. Furthermore, experiments with **polynomial regression**, **tree regression** and **singular value regression** are performed. **Gridsearch** is applied to tune **hyperparameters**. In case training turns out to be too computationally expensive, **DASK-ML** is implemented.

Each model is assesed using **5-fold cross validation**. Measures of **least squared error**, **sum squared error** and **mean squared error** are produced, as well as **accuracy** and **AUROC**.

2.4 Final model

The held out **test set** is fit on the best performing models **m1** and **m2**. Results are reported using the same measures as the ones used for training.