Dataset characterization

Adrien Pavao

March 2018

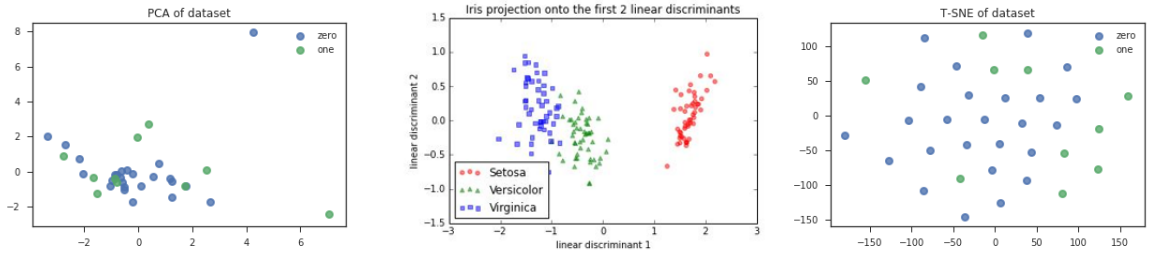
1 Change of representation

Before any computation of meta features or any visualization, we can change the data representation and reduce dimensionality.

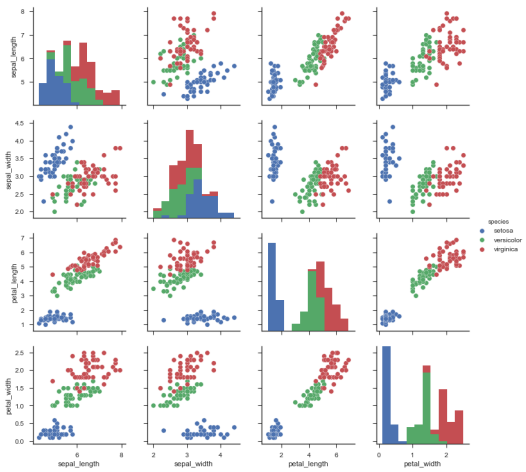
- Principal Components Analysis
- Linear Discriminant Analysis
- T-distributed stochastic neighbor embedding (t-SNE algorithm)
- Feature selection : Because we may have lots of features, it may not be possible to visualize them all. We could identity those most important or predictive of the target and eventually remove redundant features.
- Autoencoder

2 Visualization

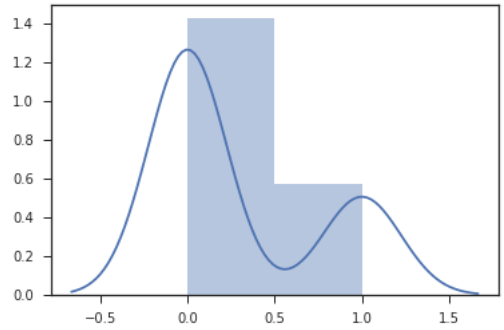
- 2D plots (after dimensionality reduction)



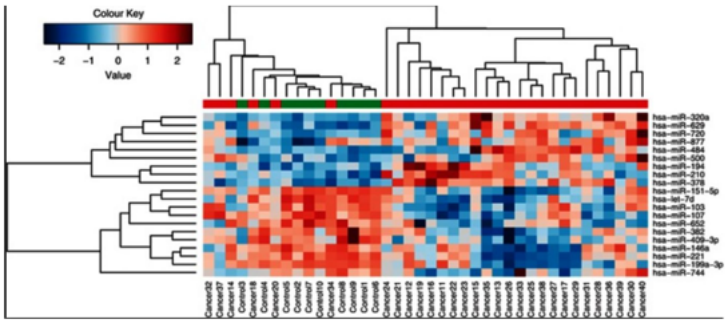
- Scatter plot matrix



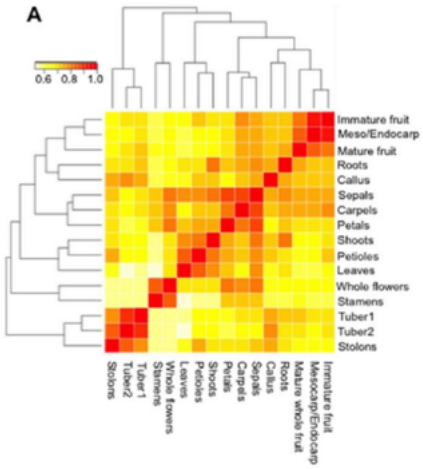
- Classes distribution



- Hierarchical clustering with heatmap matrix



- Hierarchical clustering with correlation matrix



3 Meta features

- Mean point (mean of each variable)
- Standard Deviation
- Missing values ratio
- Mean minimum distance ?
- Mesure of diversity ?
- Class Probability Min = $min_{i=1...n}(p(Class_i)) = min_{i=1...n}(\frac{NumberOfInstances\_Class_i}{TotleNumberOfInstances})$
- Class Entropy = $mean(-\sum_{i=1}^{n} p(Class_i)ln(p(Class_i)))$ where $p(Class_i)$ is the probability of having an instance of Class_i

- Dataset Ratio = $\frac{NumberOfFeatures}{NumberOfInstances}$
- Landmark[Some_Model]: accuracy of [Some_Model] applied on dataset.
- Score of some model applied on dataset (other metrics)
- Landmark Decision Node Learner & Landmark Random Node Learner: Both are decision tree with max_depth=1. 'DecisionNode' considers all features when looking for best split, and 'RandomNode' considers only 1 feature, where comes the term 'random'.
- Skewness Min: min over skewness of all features. Skewness measures the symmetry of a distribution. A skewness value > 0 means that there is more weight in the left tail of the distribution.
- Num Symbols: For each categorial feature, compute how many unique values there is ?
- Kurtosis = Fourth central moment divided by the square of the variance = $\frac{E[(x_i - E[x_i])^4]}{[E[(x_i - E[x_i])^4]]^2}$ where x_i is the ith feature.

4 User given information

In autoML format, the DataName_public.info file contains the following fields (you may supply only a subset of such fields):

- usage = 'Challenge name'.
- name = 'DataName' (dataset short nickname to be used in file names).
- task = 'regression', 'binary.classification', 'multiclass.classification', or 'multilabel.classification'.
- target_type = 'Numerical' or 'Binary' (we do not use categorical targets; multiclass problems for c classes have c binary targets).
- feat_type = 'Numerical', 'Categorical', or 'Binary'.
- metric = An AutoML metric such as'r2_metric', 'auc_metric', 'bac_metric', 'f1_metric', 'pac_metric', or your own metric.
- feat_num = number or variables (or features), i.e. number of columns of the data matrix.
- target_num = number of target values (1 for regression and same as label_num for classification problems, because we do not use categorical targets)
- label_num = number of labels for classification problems (same as target_num or NA for regression).
- train_num = number of training examples.
- valid_num = number of validation set examples.
- test_num = number of test set examples.
- has_categorical = existence of categorical variable (yes=1, no=0).
- has_missing = existence of missing values (yes=1, no=0).
- is_sparse = the data matrix is a sparse matrix (yes=1, no=0).
- time_budget = the time budget in seconds. In this version of Chalab, we impose a maximum of 500 seconds of execution time per submission.

5 Individual variable distributions

There are many tools for univariate distribution analysis.