Feature Selection

Bruno Iochins Grisci

Generative AI Academy

Part 2: Benchmark datasets

Overview

1. Tabular Data

2. Synthetic Data

3. Real-World Data

Tabular Data

Why Tabular Data?

- Common format in many real-world domains: healthcare, finance, bioinformatics.
- Each row is a sample; each column is a feature.
- Fixed-length, interpretable, and structured.
- Suitable for statistical and machine learning models.

Feature Selection is most often applied to tabular data.

Tabular data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*

Tabular data

1. Receives less attention than image and text data from the interpretability community [Molnar, 2019].
2. One of the domains not dominated by deep learning.
3. But this is changing [Borisov et al., 2021] [Kadra et al., 2021].
4. Much of the available data in science and business is tabular [Borisov et al., 2021].
5. Surrogate models (LIME) [Ribeiro et al., 2016] require the definition of a neighborhood, which is not well defined for tabular data [Molnar, 2019].
6. The attention layer is restricted to categorical features and is not applied to continuous features [Huang et al., 2020].

Synthetic Data

XOR Problem

Motivating example for feature interactions.

- Binary classification problem with non-linear separability.
- Cannot be solved with univariate feature selection.
- Requires multivariate or interaction-aware methods.

$$\mathrm{XOR}(x_1, x_2) = x_1 \oplus x_2$$

Generating XOR with Random Features

Steps to create an XOR dataset:

- Sample two binary features $x_1, x_2 \sim \text{Bernoulli}(0.5)$
- Label is $y = x_1 \oplus x_2$ (XOR logic)
- Add random noise features (irrelevant)

XOR

Class	Informative 1	Informative 2	Noisy 1	Noisy 2	Noisy 3	...	Noisy 48
0	0	0	0	1	0	...	1
1	0	1	0	1	0	...	1
1	1	0	1	0	1	...	1
1	1	0	0	0	0	...	1
0	0	0	1	0	0	...	0
1	1	0	0	0	1	...	0
0	0	0	0	1	0	...	0
0	1	1	0	1	1	...	0

Synthetic Data

Why use synthetic data for feature selection?

- Controlled experiments with known ground truth.
- Evaluate selection accuracy (PIFS, PSFI).
- Test scalability and robustness.

Common designs:

- Informative + redundant + irrelevant features
- Binary or continuous variables

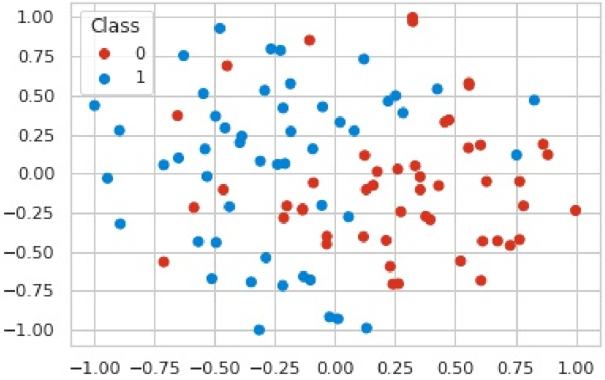
Creating Synthetic Datasets in Python

Strategy for reproducible and interpretable datasets:

- Define relevant features (informative to the label)
- Generate redundant features (e.g., linear combinations)
- Add irrelevant noise features

```
sklearn.datasets.make_classification(n_samples=100,
n_features=20, *, n_informative=2, n_redundant=2,
n_repeated=0, n_classes=2, n_clusters_per_class=2,
weights=None, flip_y=0.01, class_sep=1.0, hypercube=True,
shift=0.0, scale=1.0, shuffle=True, random_state=None)
```

Synthetic Datasets



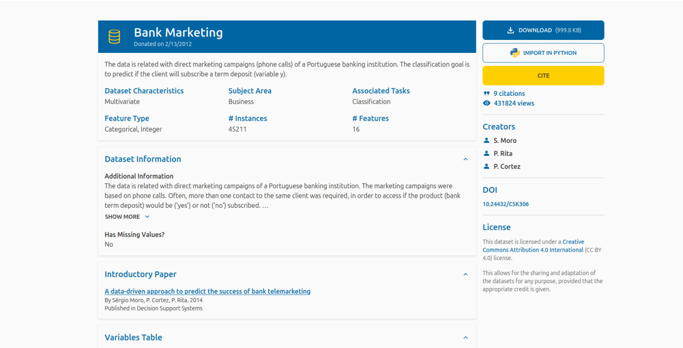
Real-World Data

UCI Machine Learning Repository

A classic benchmark collection for ML and FS.

- Curated dataset repository since 1987.
- Contains tabular datasets for classification, regression, clustering.
- Well-documented and widely used in research.
- Examples: Iris, Wine, Breast Cancer, Spam, Heart Disease.
- Drawbacks: age and size.
- `https://archive.ics.uci.edu/`

UCI Machine Learning Repository

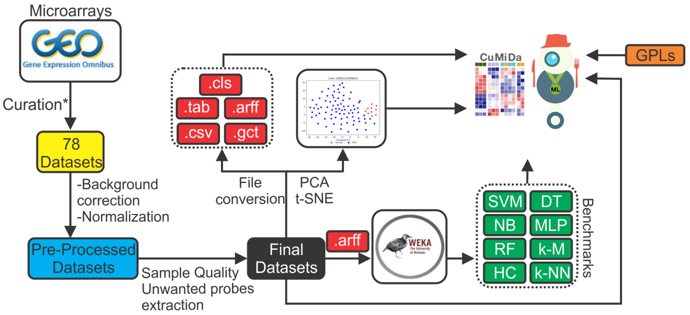


CuMiDa (Curated Microarray Database)

Focus on gene expression microarray data for FS research.

- Designed for reproducibility and reliability.
- Addresses issues in common microarray datasets.
- Contains metadata: sample origin, platform, preprocessing.
- Facilitates fair comparison of FS algorithms.
- `https://sbcb.inf.ufrgs.br/cumida`

CuMiDa


Microarrays
GEO
Gene Expression Omnibus
Curation*
78 Datasets
-Background correction
-Normalization
Pre-Processed Datasets
Sample Quality
Unwanted probes extraction
Final Datasets
File conversion
.cls
.tab
.arff
.csv
.gct
PCA
t-SNE
.arff
WEKA
CuMiDa
ML
GPLs
SVM
DT
NB
MLP
RF
k-M
HC
k-NN
Benchmarks

Being Cautious with Datasets

Why careful selection and documentation matter.

- Dataset bias can lead to misleading FS conclusions.
- Many popular datasets contain mislabeled or unbalanced data.
- Some are outdated or no longer representative.
- Always document preprocessing, sources, and limitations.

Critical for reliable FS research and reproducibility.

References I

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2021).
Deep neural networks and tabular data: A survey.
arXiv preprint arXiv:2110.01889.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. (2020).
Tabtransformer: Tabular data modeling using contextual embeddings.
arXiv preprint arXiv:2012.06678.

Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J. (2021).
Regularization is all you need: Simple neural nets can excel on tabular data.
arXiv preprint arXiv:2106.11189.

References II

Molnar, C. (2019).
Interpretable Machine Learning.
https://christophm.github.io/interpretable-ml-book/.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).
"why should i trust you?" explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco California USA. ACM.