

Aprendizado de Máquina e Reconhecimento de Padrões 2021.2



Machine Learning Concepts

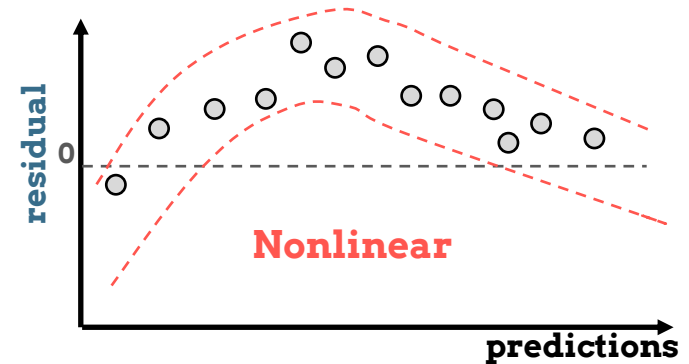
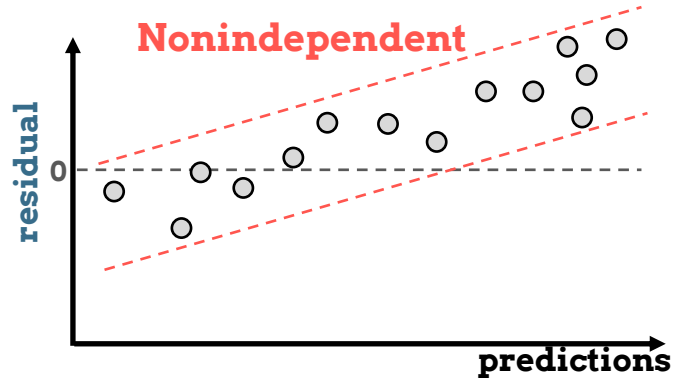
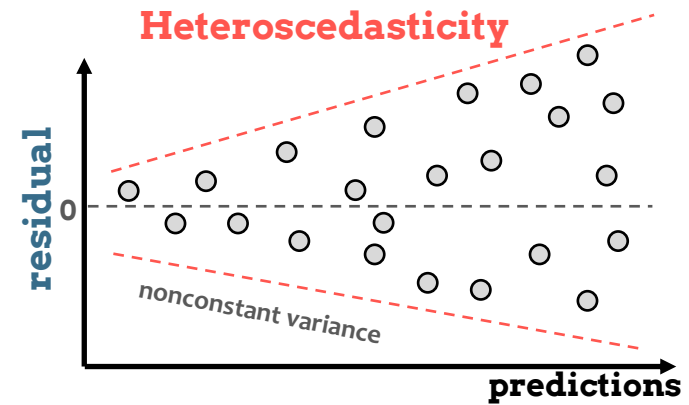
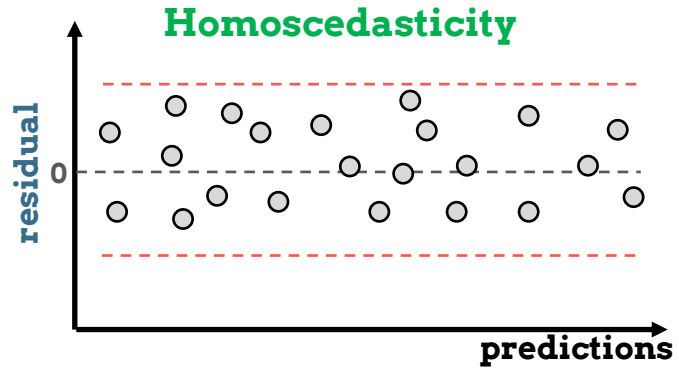
Prof. Samuel Martins (Samuka)

samuel.martins@ifsp.edu.br

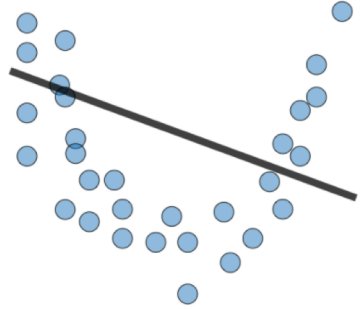
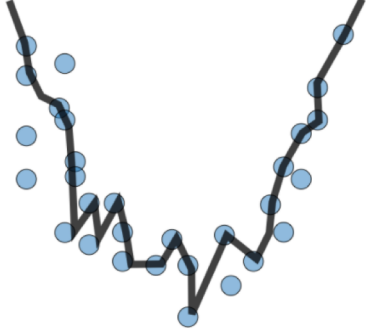
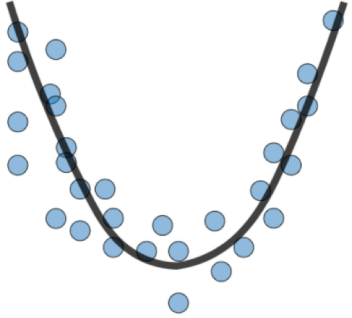
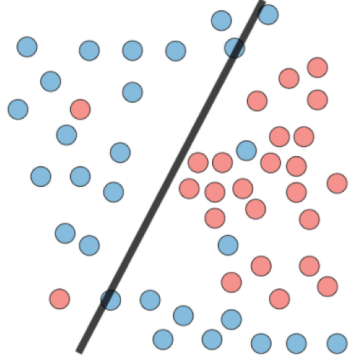
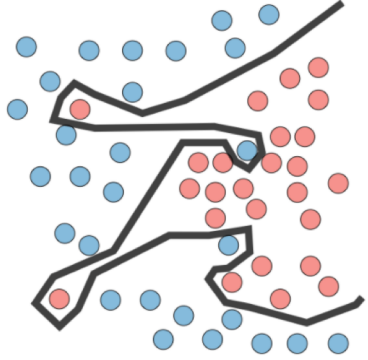
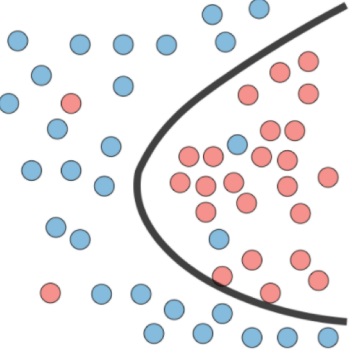


Checking Homoscedasticity Visually

Checking Homoscedasticity Visually



Overfitting vs Underfitting

	Underfitting	Overfitting	Just right
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance 	<ul style="list-style-type: none"> • Training error slightly lower than test error
Regression illustration			
Classification illustration			
Possible remedies	<ul style="list-style-type: none"> • Complexify model • Add more features • Train longer 	<ul style="list-style-type: none"> • Perform regularization • Get more data 	

Feature Scaling

Normalization

$$X_{\text{new}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

Standardization

$$X_{\text{new}} = (X - \text{mean}) / \text{Std}$$

Robust Scaler

$$X_{\text{new}} = (X - \text{median}) / \text{IQR}$$

Normalization

$$X_{\text{new}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

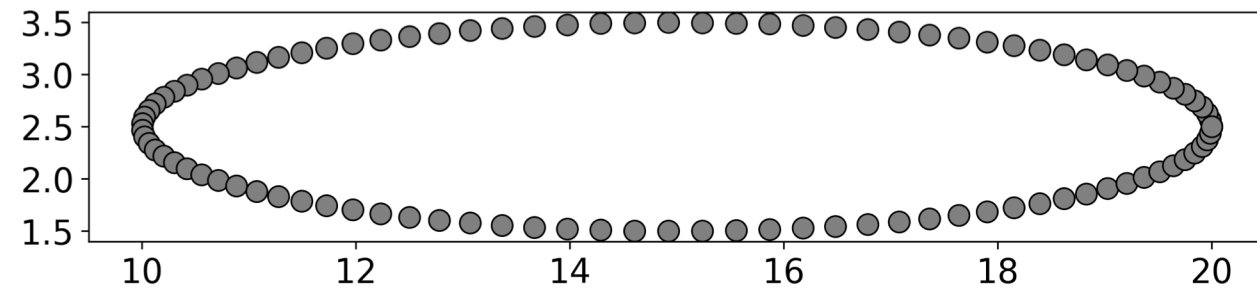
Standardization

$$X_{\text{new}} = (X - \text{mean}) / \text{Std}$$

Robust Scaler

$$X_{\text{new}} = (X - \text{median}) / \text{IQR}$$

Data



Normalization

$$X_{\text{new}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

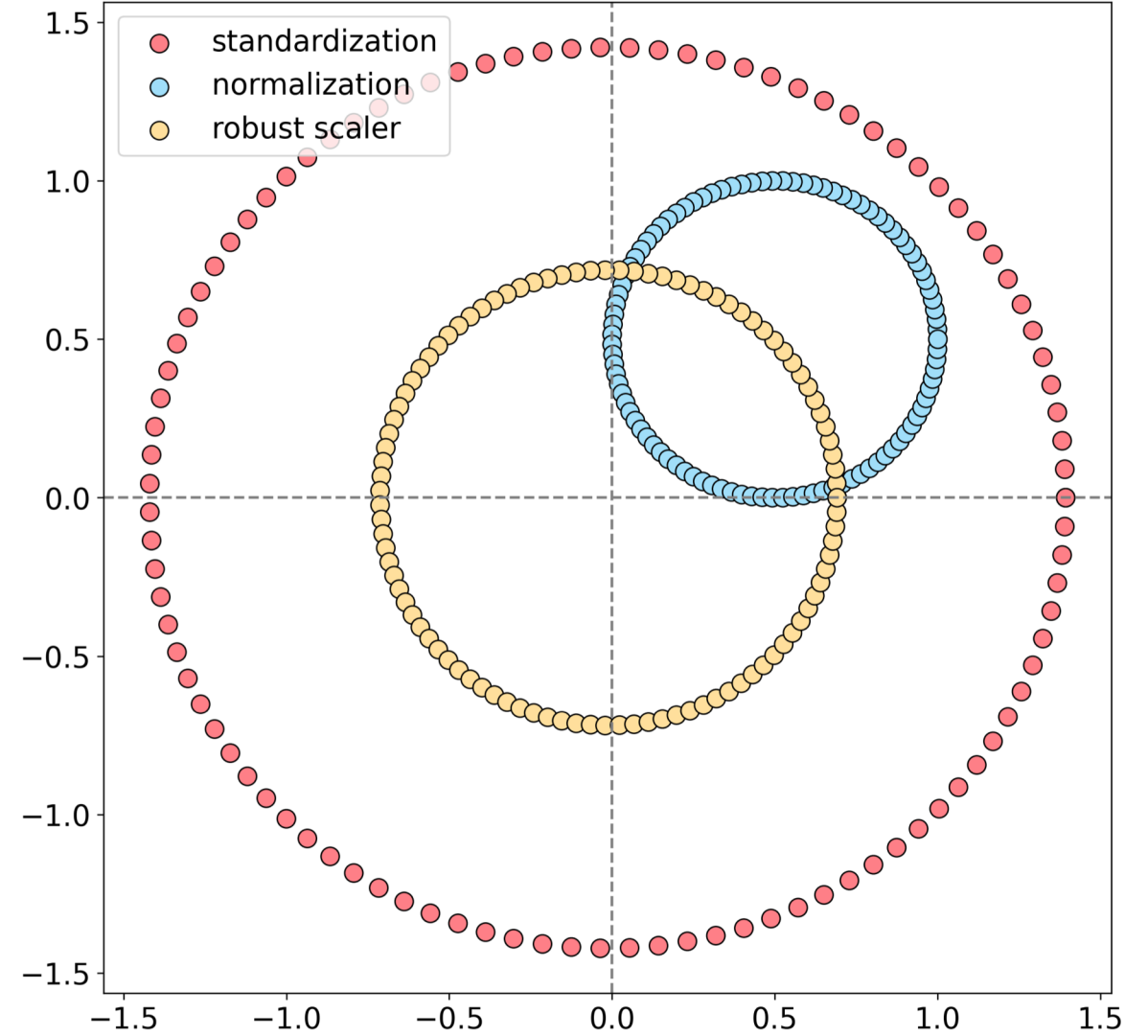
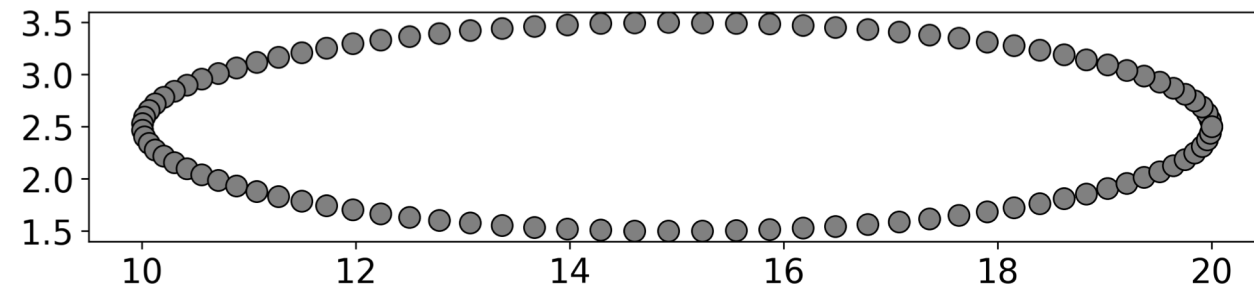
Standardization

$$X_{\text{new}} = (X - \text{mean}) / \text{Std}$$

Robust Scaler

$$X_{\text{new}} = (X - \text{median}) / \text{IQR}$$

Data



Normalization

$$X_{\text{new}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

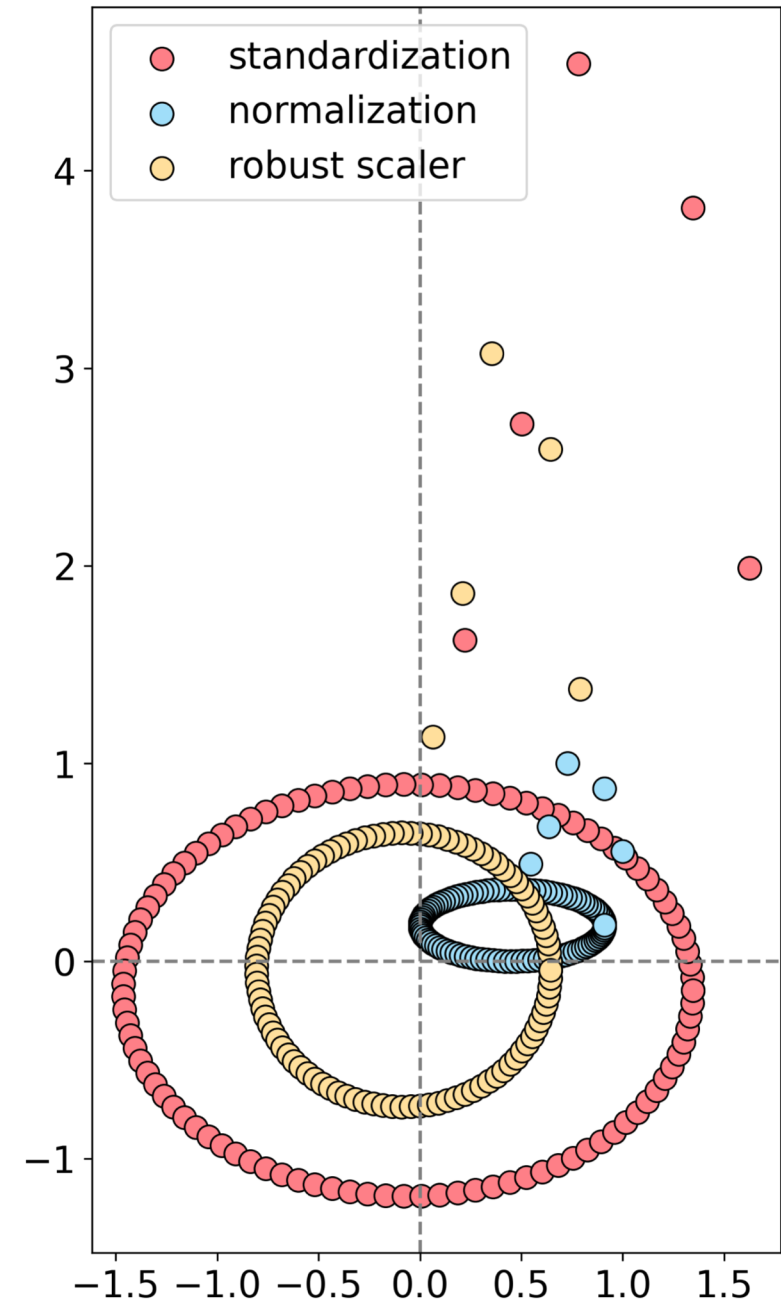
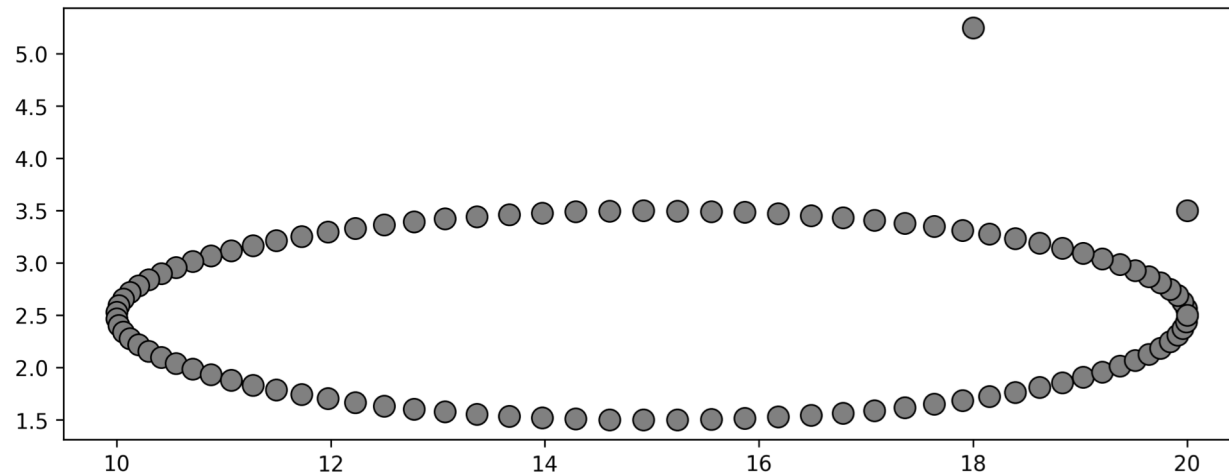
Standardization

$$X_{\text{new}} = (X - \text{mean}) / \text{Std}$$

Robust Scaler

$$X_{\text{new}} = (X - \text{median}) / \text{IQR}$$

Data with Outliers



Normalization (Min-Max Scaling)

$$X_{\text{new}} = (X - X_{\text{min}}) / (X_{\text{max}} - X_{\text{min}})$$

Minimum and maximum value of features are used for scaling

It is used when features are of different scales.

Scales values between [0, 1] or [-1, 1].

It is really affected by outliers.

Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.

This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.

It is useful when we don't know about the distribution

It is often called as Scaling Normalization

`sklearn.preprocessing.MinMaxScaler`

Standardization (Z-Score Normalization)

$$X_{\text{new}} = (X - \text{mean}) / \text{Std}$$

Mean and standard deviation is used for scaling.

It is used when we want to ensure zero mean and unit standard deviation.

It is not bounded to a certain range.

It is much less affected by outliers.

Scikit-Learn provides a transformer called `StandardScaler` for standardization.

It translates the data to the mean vector of original data to the origin and squishes or expands.

It is useful when the feature distribution is Normal or Gaussian.

It is often called as Z-Score Normalization.

`sklearn.preprocessing.StandardScaler`

Robust Scaler

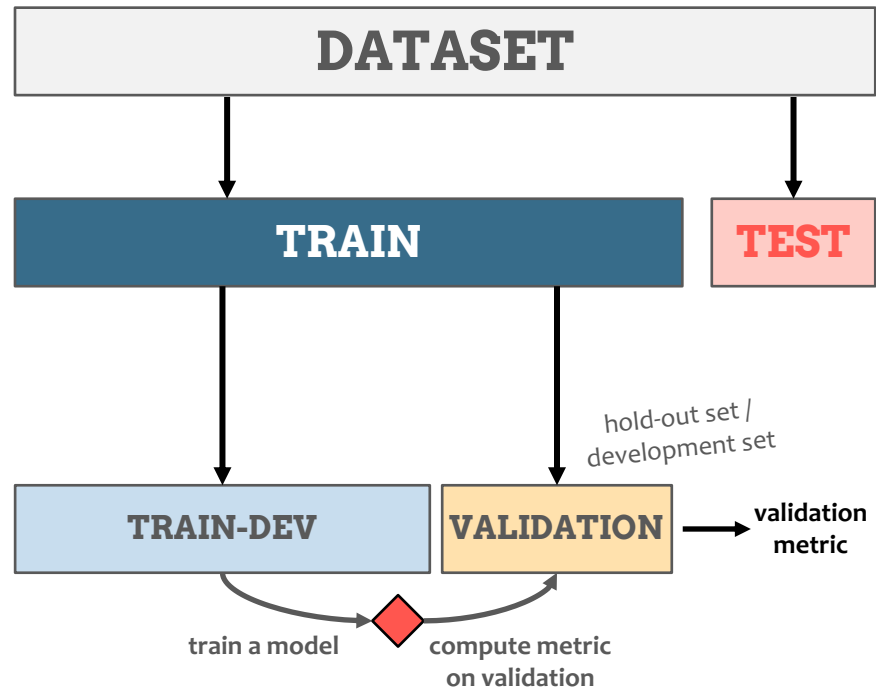
$$X_{\text{new}} = (X - \text{median}) / \text{IQR}$$

Alternative to Standardization to be yet much less affected by **outliers**.

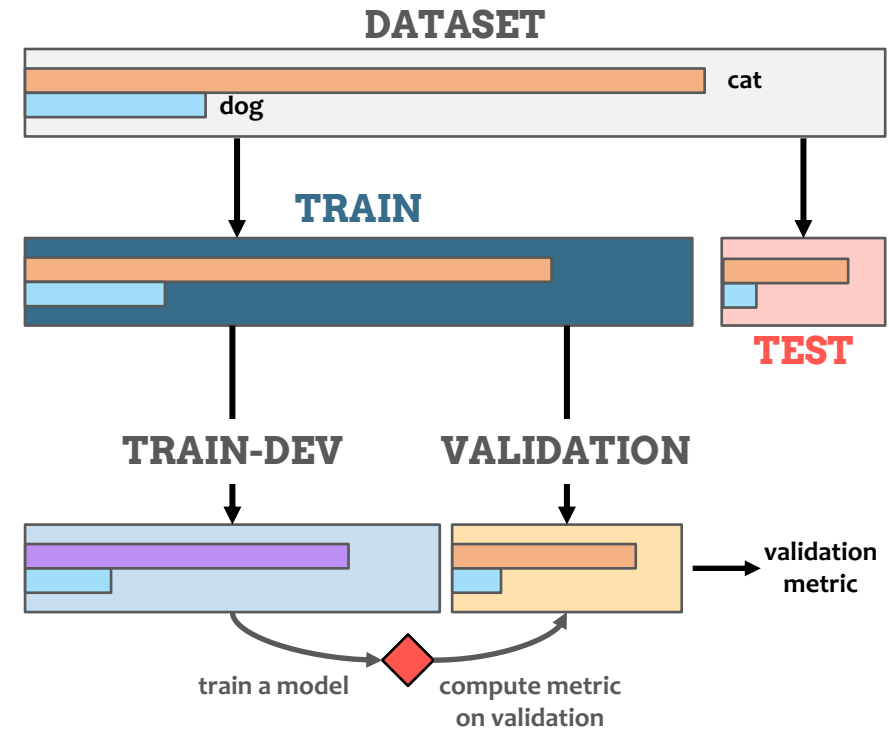
`sklearn.preprocessing.RobustScaler`

Model Validation

Holdout Validation



Stratified Holdout Validation

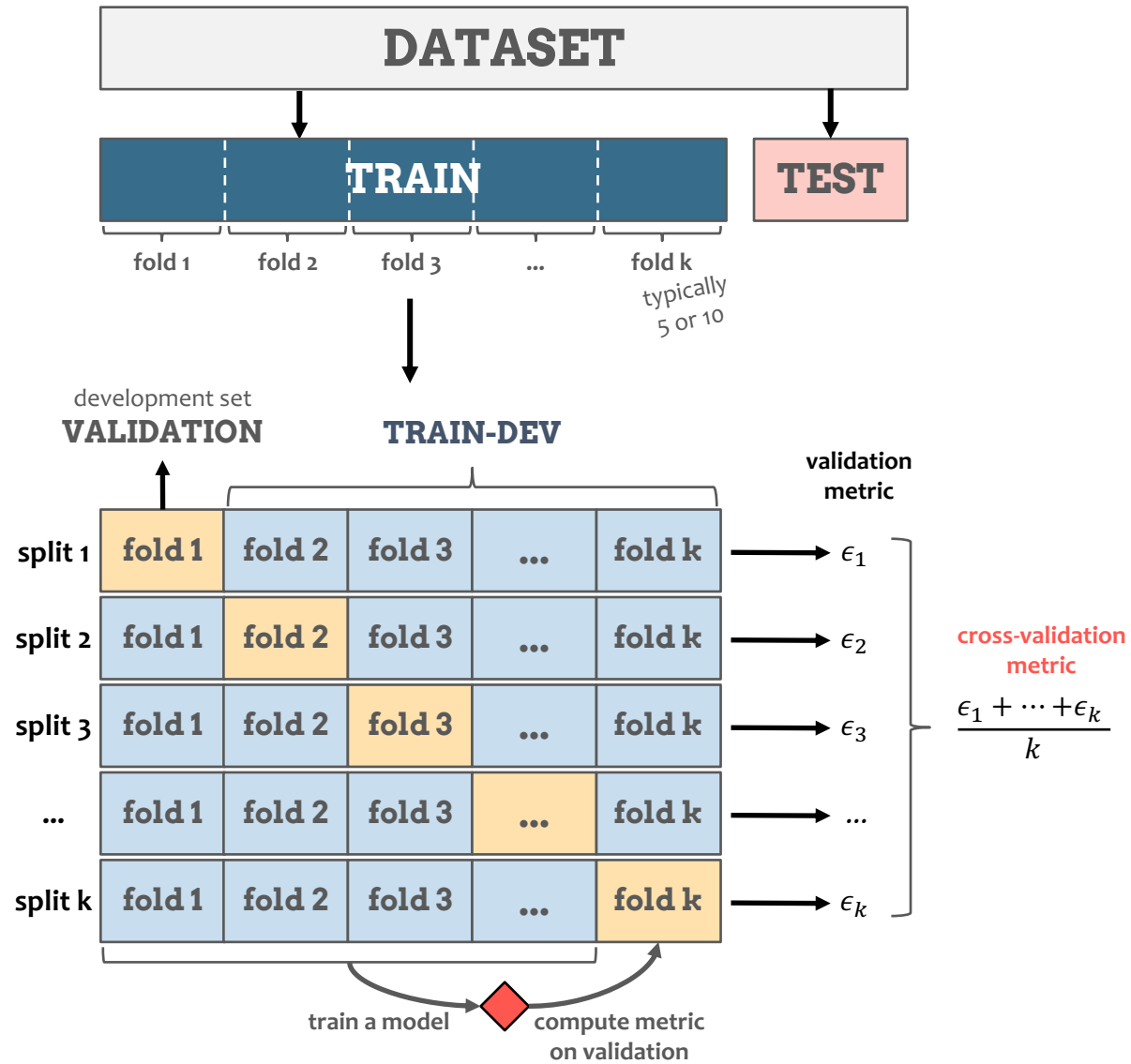


Rule of Thumb

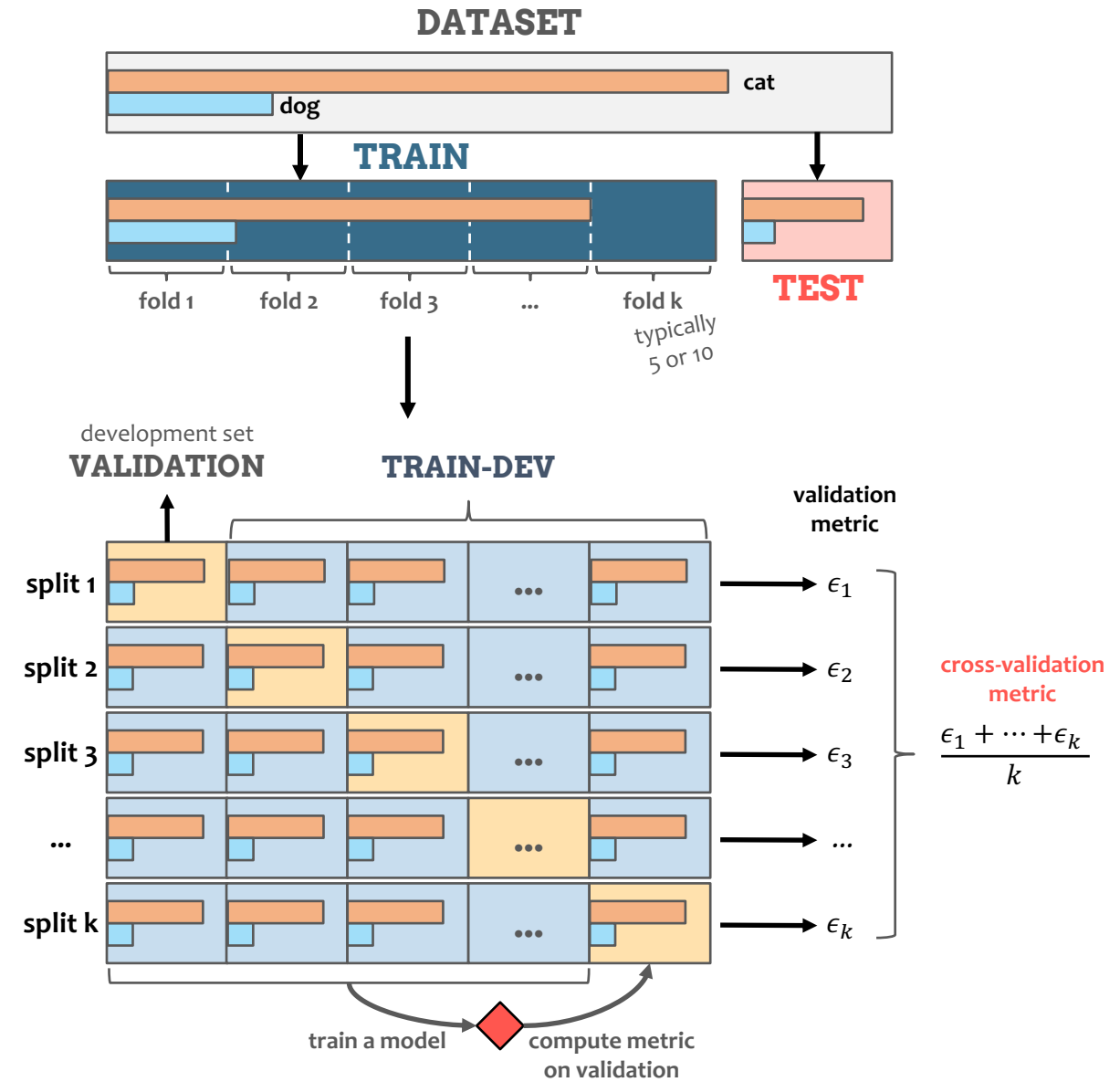
Train set: 80% of the dataset = 60% (**train-dev set**) + 20% (**validation set**)

Test set: 20% of the dataset

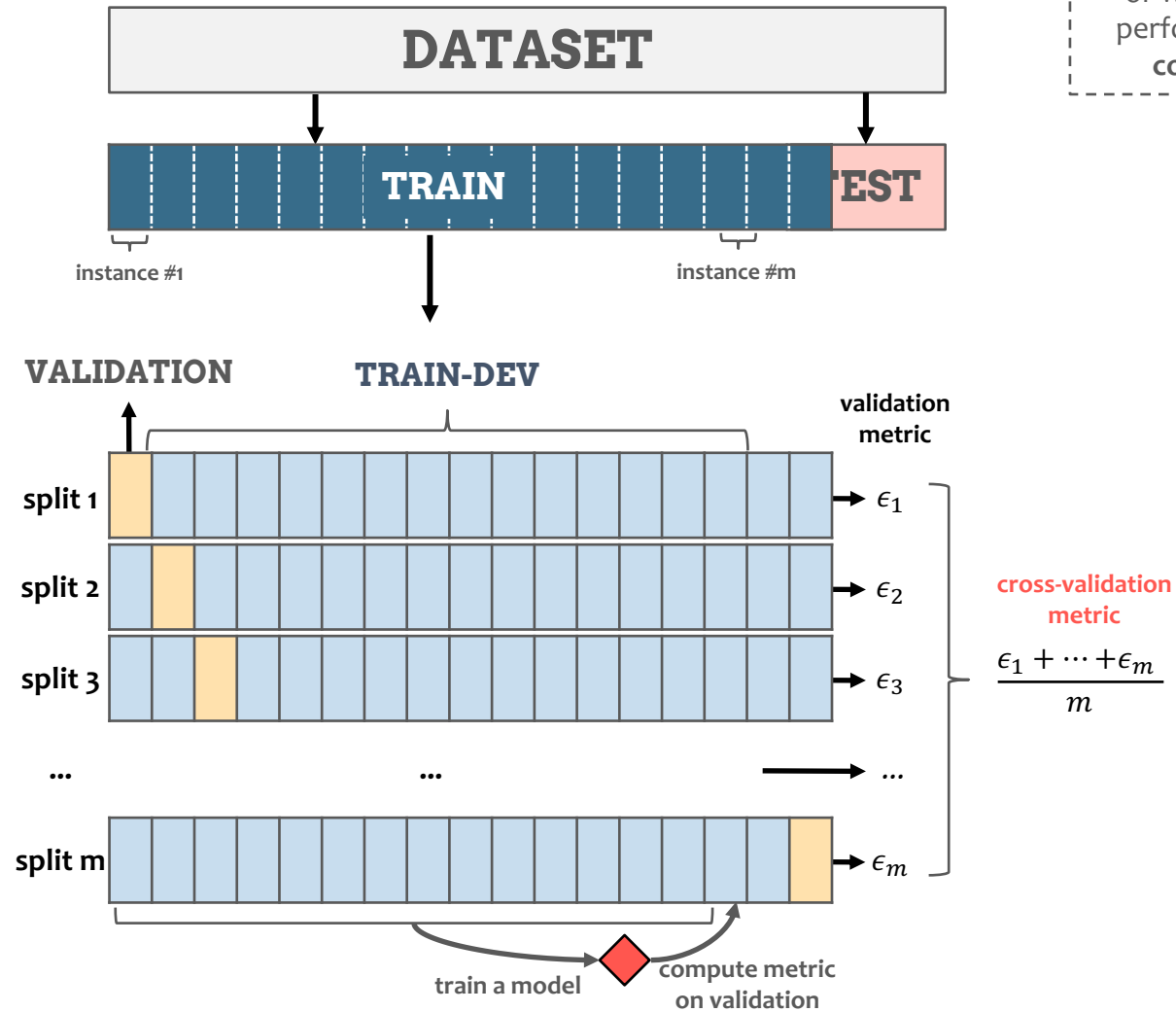
k-Fold Cross-Validation



Stratified k-Fold Cross-Validation



Leave one out Cross-Validation



Appropriate when you have a **small dataset** or when an **accurate estimate** of model performance is **more important** than the **computational cost** of the method.

Aprendizado de Máquina e Reconhecimento de Padrões 2021.2



Machine Learning Concepts

Prof. Samuel Martins (Samuka)

samuel.martins@ifsp.edu.br

