BE-Digital

Co-funded by the European Union

Alliance for Fostering Business and Education Innovation through Digital Supply Chains

Data Analytics

Selection of the appropriate machine learning technique

ARISTOTLE UNIVERSITY OF THESSALONIKI
ATLANTIS ENGINEERING
FREDERICK UNIVERSITY
ÉCOLE ENTI CONFINDUSTRIALI LOMBARDI PER L'EDUCATION
RUSE CHAMBER OF COMMERCE AND INDUSTRY
iED institute of Entrepreneurship Development
learn
CCS Digital Education

Nature of the Problem

- **Supervised vs. Unsupervised:** Determine whether the problem is supervised (labeled data for training) or unsupervised (unlabeled data).
- **Classification vs. Regression:** Decide whether the task involves predicting categories (classification) or numerical values (regression).

Type of Data

- **Structured vs. Unstructured:** For structured data (tabular), traditional machine learning methods like decision trees, random forests, and support vector machines may be effective. For unstructured data (text, images, audio), deep learning methods such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) might be suitable.
- **Dimensionality:** If the dataset has a high dimensionality, techniques for dimensionality reduction (e.g., PCA for unsupervised learning) may be beneficial.
- **Feature Characteristics:** Depending on the characteristics of the features, different algorithms may be more suitable. For example, decision trees can handle categorical features well, while neural networks can automatically learn representations from raw data.

Other

- **Small vs. Large Datasets:** Deep learning models often require large amounts of data for effective training. If the dataset is small, traditional machine learning algorithms may be more appropriate.
- **Interpretability Requirements:** Consider the need for interpretable models. If interpretability is crucial, simpler models like decision trees or linear regression may be preferred over complex models like deep neural networks.
- **Computational Constraints:** Evaluate the available computational resources. Deep learning models can be computationally intensive, especially for training large neural networks.

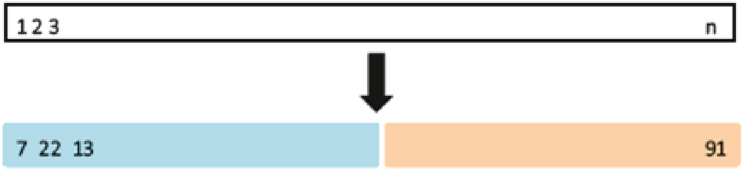
Training, Test and Validation

Co-funded by
the European Union

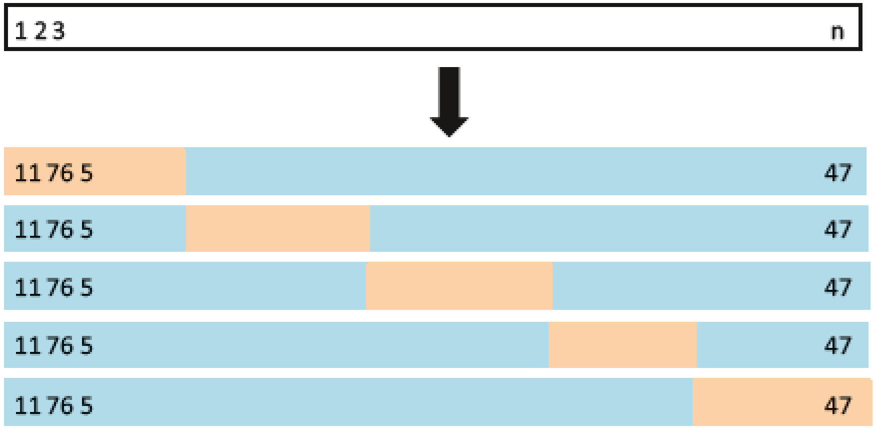
- The **training data** is the portion of the dataset used to train the model.
- The **test data**, on the other hand, is a separate subset of the dataset that is not used during the training phase. It is reserved for evaluating the performance of the trained model.
- In addition to training and test data, a third subset called **validation data** is often used during the model development process. The validation data is used to fine-tune model hyperparameters and to further evaluate model performance during training.
- **Cross-validation** is an important technique used to assess model performance, especially when the dataset is limited. It involves splitting the dataset into multiple subsets (folds), with each fold used both as a training set and as a test set in separate iterations.

Validation

Validation



Cross-validation



Source: An Introduction to Statistical Learning with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor

Regression metrics

- Mean Square Error (MSE): It measures how closely the expected response value for a certain observation matches the actual response value for that observation.
- Mean Absolute Error (MAE): It measures how closely the expected response value for a certain observation matches the actual response value for that observation in an absolute value.

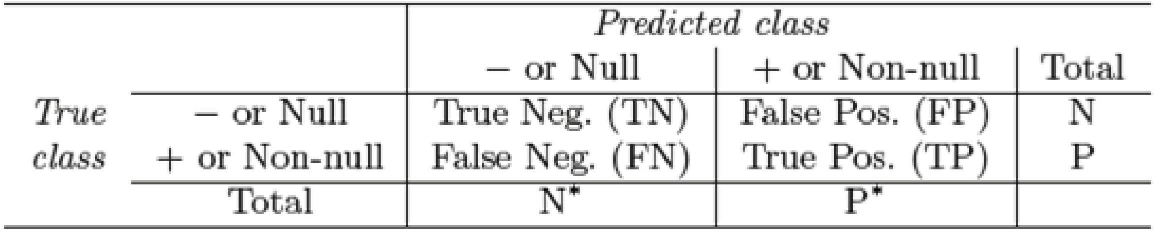
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Co-funded by
the European Union

Classification metrics

- Error rate: The percentage of errors that occur when we apply our estimate to the observations.
- Confusion matrix, precision, recall, specificity, sensitivity, accuracy and F1-score.

confusion matrix:



		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$