ANALYST

Co-funded by the European Union

MODULE 4/Unsupervised and Supervised Learning

Carlo Drago PhD

University Niccolo Cusano, Rome

MASSIVE OPEN ONLINE COURSE (MOOC)

Project N. 2023-1-IT02-KA220-HED-000161770

ANALYST - A New Advanced Level for Your Specialised Training

HOZINT HORIZON INTELLIGENCE

fyg consultores

UNIVERSITY OF WESTERN MACEDONIA

Centro Studi AMIStaDeS APS

UNIVERSITÀ CUSANO

European Institute for Local Development

UNIVERSIDAD DE ALMERÍA

UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning that helps us find patterns in data without using labeled examples. Unlike supervised learning, which relies on data that's already been labeled and defined, unsupervised learning works on raw, unlabeled data to uncover hidden relationships between observations.

Two common techniques are **clustering**, where similar data points are grouped together, and **dimensionality reduction**, which simplifies data by focusing on the most important features without losing key information.

In this sense these techniques are very relevant in data analytics.

See Witten et al (2013).

PREPROCESSING FOR UNSUPERVISED LEARNING

Data may be preprocessed before the unsupervised learning approaches

- Data with outliers can be problematic for unsupervised learning
- Standardization and normalization can be useful
- Feature engineering can be used to reduce data dimensionality (using principal component analysis).
- At the same time recoding of variables can be useful to convert categorical variables into numerical ones or create new features that better capture the characteristics of the observations considered

DISTANCES

In unsupervised learning, the **distance is a key concept**. It helps algorithms to group data points that are similar to each other.

Since there are no labels to guide the process, measuring how close or far data points are from each other is essential for finding patterns or creating clusters.

K-MEANS

K-Means is a popular clustering algorithm used in data analysis. It aims to group a set of data points into a certain number of clusters defined by their centers. This is how the algorithm works:

Initialization: The algorithm randomly selects 'k' initial centroids from the data points. The number 'k' is predefined and stands for the number of clusters you want to find.

Assignment step: Each data point is assigned to the closest centroid, forming clusters. The distance between the data points and the centroids is usually measured using the Euclidean distance.

See Witten et al. (2013)

K-MEANS

Update step: After all points have been assigned, the centroids are recalculated as the average points in each cluster. These new centers become the new centers of the respective clusters.

Repeat: The assignment and update steps are repeated until the centers no longer change significantly. This means that the algorithm converges and the clusters are stable.

K-Means is efficient and easy to implement, but it requires that you specify the number of clusters in advance. K-Means works best when the clusters are approximately the same size and spherical.

See Witten et al. (2013)

I

HIERARCHICAL CLUSTERING

Hierarchical clustering is a method of cluster analysis that aims to build a hierarchy of clusters. It can be carried out in agglomerative (bottom-up approach) or divisive (top-down approach). In the agglomerative method, each data point starts as a separate cluster, and the cluster pairs merge as they move up the hierarchy.

In contrast, the divisive method starts with all data points in a single cluster and divides them into smaller clusters.

See Witten et al. (2013)

HIERARCHICAL CLUSTERING

This process continues until each data point is in its own individual cluster. Hierarchical clustering is often visualized using a dendrogram that illustrates the arrangement of the clusters formed at each step.

This method helps identify the natural grouping within a dataset without predetermining the number of clusters.

ANALYST

Co-funded by the European Union

CLUSTERING VALIDATION

Clustering validation is the process of evaluating the **quality and effectiveness of a clustering** algorithm. It assesses how well the results meet the intended objectives and how accurately the data points are grouped.

This process can be carried out using internal, external, and relative validation measures. Internal validation focuses on the dataset's intrinsic properties, such as clusters' compactness and separation.

Hassan, et al. (2024)

CLUSTERING VALIDATION

External validation compares the clustering results with existing ground truth, often using metrics such as the Rand Index or the F-measure.

Relative validation also compares clustering algorithms or configurations to determine the best approach for a particular data set. Overall, clustering validation is crucial to ensure the reliability and accuracy of the clustering process.

Hassan, et al. (2024)

SUPERVISED LEARNING

Supervised learning is a type of machine learning where an algorithm is trained using a dataset where the correct answers are already provided.

In this sense there are relevant differences with unsupervised learning in which algorithms work with data which are not labelled.

See Bishop (2006), Witten et al. (2013)

SUPERVISED LEARNING

The supervised learning algorithms learns by finding patterns in the data (the training set) and matching inputs to the correct outputs. Once it's trained, it can use what it has learned to make predictions or classify new data it hasn't seen before (the test set).

This approach is used in tasks like filtering spam, recognizing images, and predicting analytics application which are very useful.

See Bishop (2006), Witten et al. (2013)

PREPROCESSING FOR SUPERVISED LEARNING

Outliers are very relevant in this context.

In supervised learning, preprocessing the data is crucial to ensure the model can learn effectively. When dealing with outliers, preprocessing is about identifying and managing these extreme values that can distort the results.

PREPROCESSING FOR SUPERVISED LEARNING

The first step is identifying outliers using statistical methods such as the Z-score or IQR (Interquartile Range). Once they are identified, there are several ways to deal with them. Removing these outliers if they are errors or not representative of the population is possible. Alternatively, the data analyst can transform them, through a logarithmic transformation or normalization, to minimize their impact.

In some cases, it is also possible to replace outliers with the median or mean of the data set. These steps help create a more robust model by reducing noise and ensuring that the learning algorithm focuses on the patterns that represent the data.

STEPWISE LOGISTIC REGRESSION

Stepwise regression is a method used in supervised learning to select a subset of relevant features by systematically adding or removing predictors. In stepwise logistic regression, this approach is applied to models where the dependent variable is categorical, often binary.

Initialization: Start with an initial model, which can be empty (no predictors) or a complete model (all potential predictors).

Forward selection: Start with no variables in the model. Add the predictor that contributes the most to the model based on a chosen criterion, such as the Akaike Information Criterion (AIC). Add one predictor at a time, selecting the one that most improves the model.

STEPWISE LOGISTIC REGRESSION

Backward elimination: Start with all potential predictors in the model. Remove the least significant predictor, again based on a specific criterion. Continue removing predictors until further elimination would degrade the model's performance.

Bidirectional elimination: This is a combination of forward selection and backward elimination. After each step in which a predictor is added, the algorithm checks whether existing predictors can be removed without significantly affecting the model's performance.

Model evaluation: Evaluate the model at each step using cross-validation or other performance metrics to ensure that the model generalizes well to unseen data.

STEPWISE REGRESSION

Termination: Terminate if adding or removing predictors does not significantly improve the model or if a predefined termination criterion is met.

Stepwise logistic regression helps streamline the modeling process by focusing on the most meaningful predictors, simplifying the model, and improving its predictive power.

ANALYST

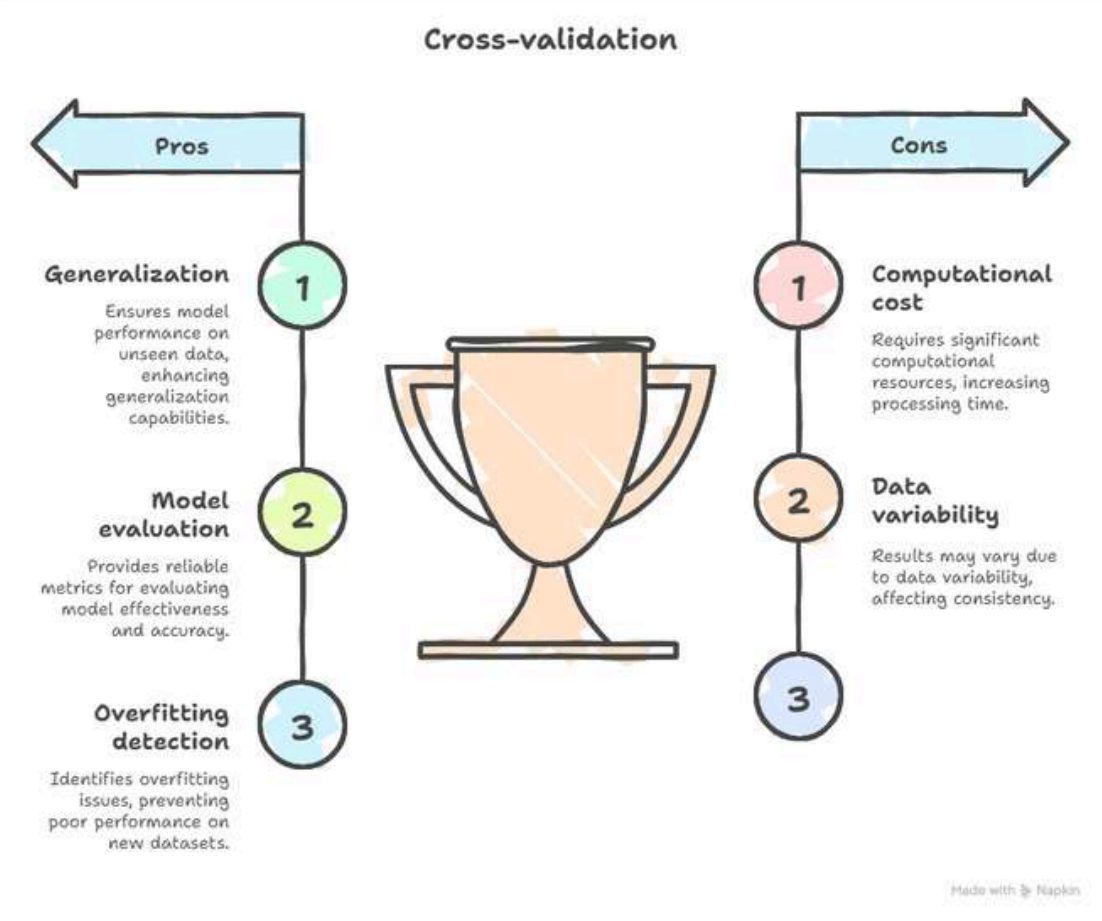
Co-funded by the European Union

CROSS-VALIDATION

Cross-validation is a technique used in supervised learning to evaluate the performance and robustness of a model. It involves dividing the data set into several subsets or "folds"

The "rule" for splitting the data set is: 80% of the observations in the training set and 20% of the observations in the test set. The observations are randomly assigned to the subset.

CROSS-VALIDATION



CROSS-VALIDATION

The model is then trained with the training set, while the remaining subset is used for testing.

The results provide an estimate of the model's performance. Cross-validation helps to detect possible overfitting and ensures that the model generalizes well to unseen data.

NEURAL NETWORKS

Exists various algorithms used in supervised learning. Here we present Neural Networks and decision trees two of the most used algorithms used in literature.

Neural networks are supervised algorithms inspired by the human brain. They consist of interconnected nodes, called neurons, that work together to process data and learn patterns. These networks are organized in layers: an input layer, one or more hidden layers and an output layer.

See Bishop (2006)

NEURAL NETWORKS

Each neuron receives input, processes it, and passes it on to the next layer, allowing the network to recognize complex patterns and make predictions.

Neural networks are used in many areas, such as image and speech recognition, natural language processing, and many others, as they can learn from large amounts of data and improve over time.

See Witten et al. (2013)

ANALYST

Co-funded by the European Union

DECISION TREES

Decision trees are a data analysis tool used for decision-making and predictive modeling. They divide data into branches based on specific criteria, visually representing possible outcomes and decisions.

Decision trees are beneficial for dealing with complex data sets, including those with errors, as they can help identify patterns and relationships.

See Witten et al. (2013)

HYPERPARAMETER TUNING

Hyperparameter tuning is **optimizing the parameters of a machine learning model** that are not learned from the data but are set before the training process. These parameters, known as hyperparameters, can significantly influence the performance of a model.

Examples include the learning rate, the number of layers in a neural network, or the number of trees in a random forest. Tuning involves systematically searching through hyperparameters to find the best combination that produces the most accurate model. This can be done using techniques such as grid search, random search, or more advanced methods such as Bayesian optimization.

See Yang, & Shami (2020)

ANALYST

Co-funded by the European Union

ENSEMBLE LEARNING

Ensemble learning is a Supervised Learning approach useful to identify the best classification of the predictive approach considered using different algorithms.

These algorithms are considered jointly and then the final valuation come from a rule which combine the different results of the different algorithms. For instance it is possible to use a "Majority Rule" using as a response the most "voted" by the different algorithms.

In this way it is increased the robustness of the results considering different algorithms or approaches.

See Witten et al. (2013)

REFERENCES

- Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2
- Hassan, B. A., Tayfor, N. B., Hassan, A. A., Ahmed, A. M., Rashid, T. A., & Abdalla, N. N. (2024). From A-to-Z review of clustering validation indices. Neurocomputing, 601, 128198.
- Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data. John Wiley & Sons.
- Tufte, E. R (1983). The visual display of quantitative information (Vol. 2, No. 9). Cheshire, CT: Graphics press.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, No. 1). New York: springer.

REFERENCES

- Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295-316.