AI Concepts and Definitions

AI & SUSTAINABILITY IN VET EDUCATION

ERASMUS 2023-1-LT01-KA220-VET-000155506

Kristina Sutiene and Liepa Bikulciene

Kaunas University of Technology

Darica, Türkiye
Sep 16-20, 2024

Outline

Table of Contents

Concept and Types of ML models

Roughly speaking, ML aims at learning to predict the label of a data point based solely on the features of this datapoint.
Formally, the prediction is obtained as the function value of a hypothesis map $h : X \to Y$ whose input argument are the features of a datapoint, i.e.

$$y \approx \underbrace{h(\mathbf{x})}_{\hat{y}}, \text{where } \mathbf{x} \in X - \text{ a feature set}, y \in Y - \text{a label}, \hat{y} - \text{a prediction.}$$

- Based on how ML methods assess the quality of different hypothesis maps we distinguish three main types of ML: supervised, unsupervised and reinforcement learning.

Concept and Types of ML models. Supervised Learning

- These methods use a data set that consists of labeled data points (for which we know the correct label values).
- Labeled data points might be obtained from human experts that annotate ("label") data points with their label values.
- Supervised ML searches for a hypothesis (ML method) that can imitate the human annotator and allows to predict the label solely from the features of a data point.

Concept and Types of ML models. Supervised Learning

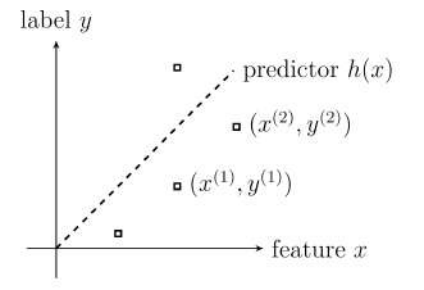


Figure: Supervised ML methods fit a (typically highly non-linear) curve to a (typically large) set of data points

Supervised ML methods learn a hypothesis (model) with minimum discrepancy between its predictions and the true labels of the datapoint in the data set. For the actual implementing of this curve fitting we need a loss function that quantifies the fitting error.

Concept and Types of ML models. Supervised Learning

- Supervised ML methods differ in their choice for a loss function to measure the discrepancy between predicted label and true label of a data point.
- ML methods must process billions of data points with each single data point characterized by a potentially vast number of features.
- Besides the size and complexity of datasets, another challenge for modern ML methods is that they must be able to fit highly non-linear predictor maps. For example, deep learning methods address this challenge by using a computationally convenient representation of non-linear maps via artificial neural networks.

Concept and Types of ML models. Unsupervised Learning

- Some ML methods do not require knowing the label value of any data point and are therefore referred to as unsupervised ML methods.
- Unsupervised methods must rely solely on the intrinsic structure of data points to learn a good hypothesis. Thus, unsupervised methods do not need a teacher or domain expert who provides labels for data points.

Concept and Types of ML models. Unsupervised Learning

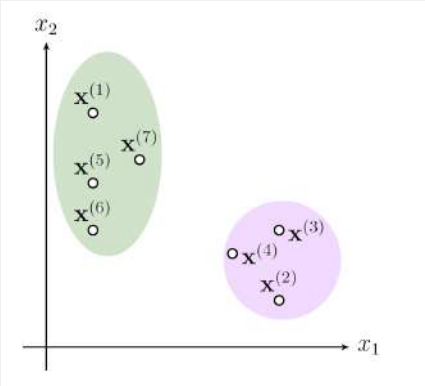


Figure: Clustering methods learn to predict the cluster (or group) memberships of data points based solely on their features

Concept and Types of ML models. Unsupervised Learning

- Clustering methods group data points into few subsets such that data points within the same subset or cluster are more similar with each other than with data points outside the cluster.
- Feature learning methods determine numeric features such that data points can be processed efficiently using these features. Two important applications of feature learning are dimensionality reduction and data visualization.

Concept and Types of ML models. Reinforcement Learning

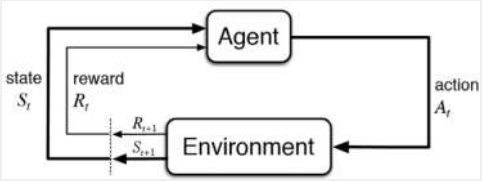
- Reinforcement learning (RL) studies applications where the predictions obtained by a hypothesis influences the generation of future data points.
- RL applications involve data points that represent the states of a (programmable) system (an artificial intelligence agent) at different time instants.
- The label of such a data point has the meaning of an optimal action that the agent should take in a given state. Similar to unsupervised ML, RL methods must learn a hypothesis without having access to labeled data points.

Concept and Types of ML models. Reinforcement Learning

- What sets RL methods apart from supervised and unsupervised methods is that it not possible for them to evaluate the loss function for different choices of a hypothesis.
- Mathematically speaking, RL methods can evaluate the loss function only pointwise, i.e., for the current hypothesis that has been used to obtain the most recent prediction.
- These point-wise evaluations of the loss function are typically implemented by using some reward signal. Such a reward signal might be obtained from some input device (sensor) or system and allows to quantify the usefulness of the current hypothesis.

Concept and Types of ML models. Reinforcement Learning

RL algorithm (agent) evaluates a current situation (state), takes an action, and receives feedback (reward) from the environment after each act. Positive feedback is a reward, and negative feedback is punishment for making a mistake



RL algorithm learns how to act best through many attempts and failures. Trial-and-error learning is connected with the so-called long-term reward. This reward is the ultimate goal the agent learns while interacting with an environment. The algorithm gets short-term rewards that together lead to the cumulative, long-term one.

Concept and Types of ML models. Reinforcement Learning

- RL in the healthcare sector could be applied in case of DTRs (Dynamic Treatment Regimes) to supportd medical professionals in handling patients' health. DTRs use a sequence of decisions to come up with a final solution. This sequential process may involve the following steps:
  - Determine the patient's live status
  - Decide the treatment type
  - Discover the appropriate medication dosage based on the patient's state
  - Decide dosage timings, and so on

  With this sequence of decisions, doctors can fine-tune their treatment strategy and diagnose complex diseases such as mental fatigue, diabetes, cancer, etc.

Table of Contents

Supervised Learning

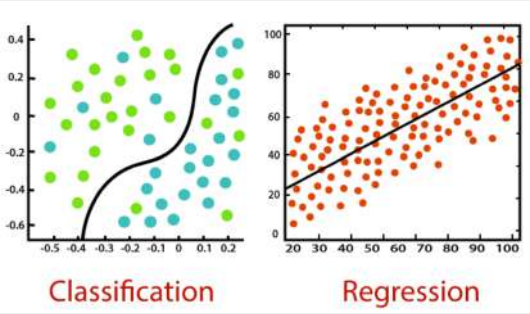
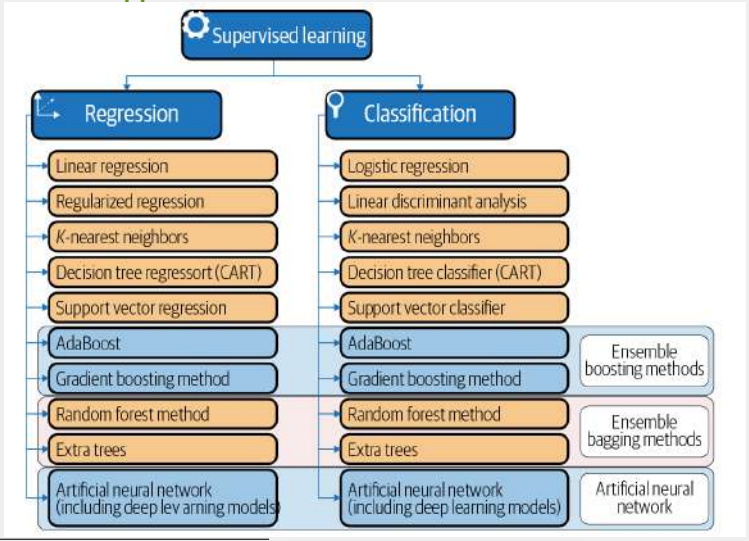


Figure: Classification is a process of finding a function which helps in dividing the dataset into categories based on different parameters, while regression is a process of finding the relations between labels (continuous) and independent variables (features) in order to predict labels

[0] https://www.javatpoint.com/regression-vs-classification-in-machine-learning

Supervised Learning



[0] https://www.oreilly.com/library/view/machine-learning-and/9781492073048/ch04.html

Supervised Learning. Classification

- Classification is predicting categories or classes.
- In classification tasks, models use characteristics or features of an input data point to determine which specific category the data point belongs to.
- In medical diagnostics, a classification model might predict whether a tumor is cancerous or benign based on features such as a patient's age, tumor size, and tobacco use. This is an example of binary classification—the special case in which models predict one of two categories.
- Multi-class classification, on the other hand, involves predicting one of multiple categories.
- An image classification model might classify an image as belonging to one of multiple different classes such as tumour or not. Computer vision often applies these methods to enable computers to interpret and understand visual data from the world.

Supervised Learning. Regression

- Regression is predicting numbers.
- In regression tasks, models use features of input data to predict numerical outputs.
- A real estate company might use a regression model to predict house prices from a dataset with features such as location, square footage, and number of bedrooms.
- While classification models produce discrete outputs that place inputs into a finite set of categories, regression models produce continuous outputs that can assume any value within a range.

Supervised Learning. Classification. Decision Tree

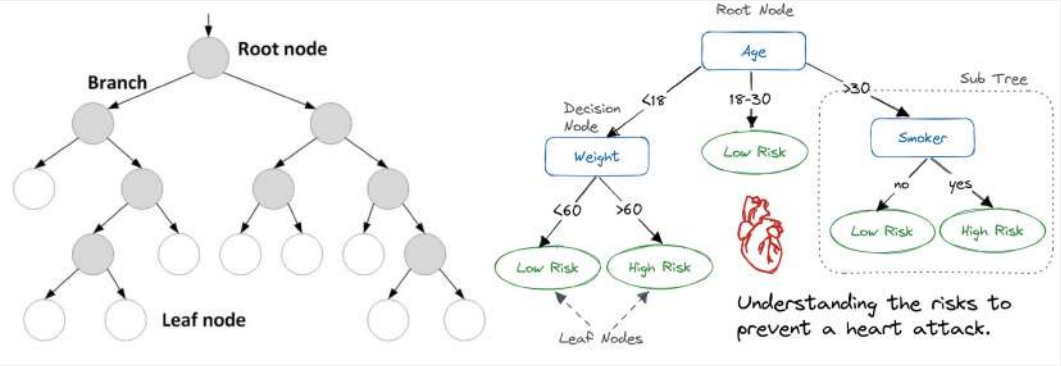
- The decision tree is a well known and popular ML approach.
- Can be used for both classification and regression tasks. In case of classification, the method is called a decision tree. For regression problem, the decision tree is typically called regression tree.
- It also serves in other more advanced ML methods, mainly ensemble models, as the base learner.
- They provide a clear and intuitive way to make decisions based on data by modeling the relationships between different variables.
- Capable of capturing non-linear relationships between features and target variables.
- Decision trees do not require normalization or scaling of the data.

Supervised Learning. Classification. Decision Tree

Structure:

- Root Node: Represents the entire dataset and the initial decision to be made.
- Internal Nodes: Represent decisions or tests on attributes. Each internal node has one or more branches.
- Branches: Represent the outcome of a decision or test, leading to another node.
- Leaf Nodes: Represent the final decision or prediction. No further splits occur at these nodes.

Supervised Learning. Classification. Decision Tree



Supervised Learning. Classification. Decision Tree

The process of creating a decision tree involves:

- Selecting the Best Attribute: Using a metric like Gini impurity, entropy, or information gain, the best attribute to split the data is selected.
- Splitting the Dataset: The dataset is split into subsets based on the selected attribute.
- Repeating the Process: The process is repeated recursively for each subset, creating a new internal node or leaf node until a stopping criterion is met (e.g., all instances in a node belong to the same class or a predefined depth is reached).

Supervised Learning. Classification. Decision Tree

Metrics for Splitting

- Gini Impurity: Measures the likelihood of an incorrect classification of a new data point if it was randomly classified according to the distribution of classes in the dataset

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2,$$

where p_i is the probability of an instance being classified into a particular class.

- Entropy: Measures the amount of uncertainty or impurity in the dataset

$$\text{Entropy} = -\sum_{i=1}^{n} (p_i) log_2(p_i).$$

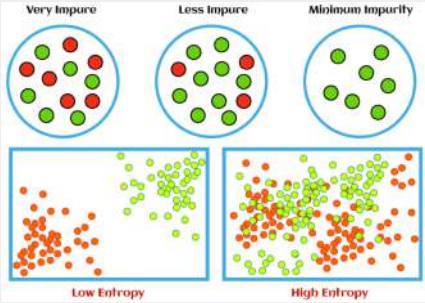
Supervised Learning. Classification. Decision Tree

Metrics for Splitting

- Information Gain: Measures the reduction in Entropy or Gini impurity after a dataset is split by some feature

$$\text{InformationGain} = Entropy_{Parent} - \sum_{i=1}^{n} \frac{D_i}{D} \cdot Entropy(D_i),$$

where D_i is the subset of D after splitting by an splitting by some feature

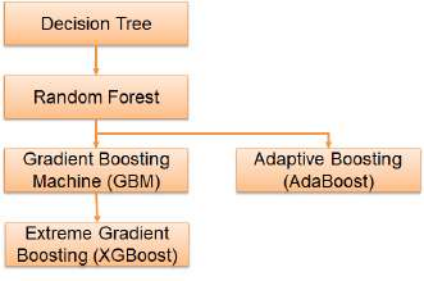


Supervised Learning. Classification. Decision Tree

Strategies to Overcome Overfitting in Decision Tree Models

- Pruning involves removing parts of the decision tree that do not contribute significantly to its predictive power. This helps simplify the model and prevent it from memorizing noise in the training data.
- Setting a maximum depth for the decision tree restricts the number of levels or branches it can have. The model becomes more generalized and less likely to capture noise or outliers.
- Specifying a minimum number of samples required to create a leaf node ensures that each leaf contains a sufficient amount of data to make meaningful predictions.
- Ensemble methods such as Random Forests and Gradient Boosting combine multiple decision trees to reduce overfitting.

Supervised Learning. Decision Tree based models



Random Forests: An ensemble method that builds multiple decision trees and merges their results to improve accuracy and control overfitting.
AdaBoost and Gradient Boosting (GB): Sequentially builds models where each new model corrects errors made by previous models, improving overall accuracy.
Extreme Gradient Boosting (XGBoost): An optimized version of GB that includes regularization to reduce overfitting and improve performance.

Supervised Learning. Decision Tree: example

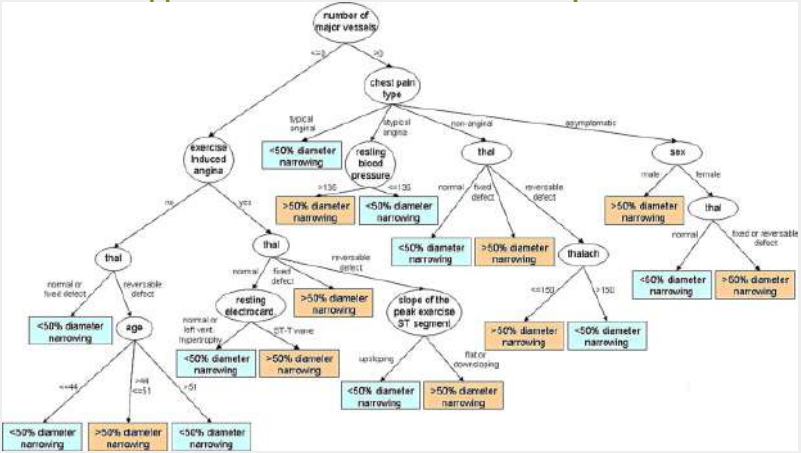


Figure: Decision tree to identify patients with heart disease

[0] https://www-sciencedirect-com.ezproxy.ktu.edu/science/article/pii/S0957417412006707

Supervised Learning. ANN

ANNs (artificial neural networks) are used for regression or classification problems and they consists of two basic architecture:

- Single-Layer Artificial Neural Network (Perceptron)
- Multi-Layer Artificial Neural Network

Supervised Learning. ANN

Single-Layer Artificial Neural Network (Perceptron)

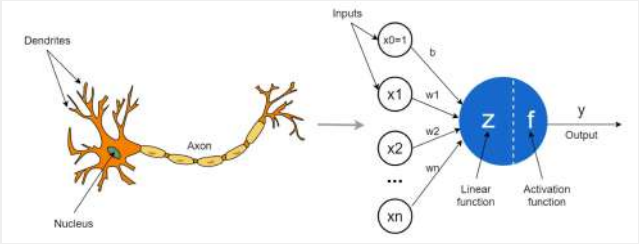


Figure: Artificial neurons (Perceptrons, Units or Nodes) are the simplest elements or building blocks in a neural network. They are inspired by biological neurons that are found in the human brain.

[0] https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-networks-fab22249cbfc

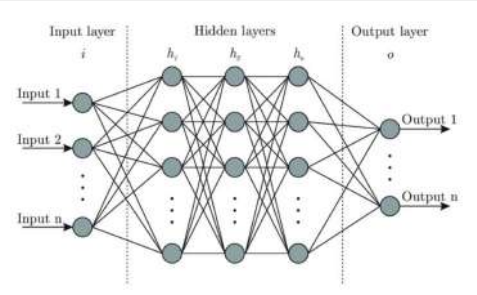
Supervised Learning. ANN

The perceptron consists of five components:

- Inputs: These are the features that we have.
- Weights: Weight parameters (w) control the strength of the connection between inputs and neurons. It can also be said to represent the effect of an independent variable on the result.
- Bias value(b): It is a constant value that allows to control the output value.
- Activation Functions: The activation function (f) defines the output of the neuron according to certain conditions.
- Output: The predicted value (y) is the result we want to find.

Supervised Learning. ANN

Multi-Layer Artificial Neural Network



Multi-Layer Artificial Neural Network consists layers more than one. Beside of the perceptrons, they can be used for non-linearly separable problems. They achieve this with the activation functions they use in their layers. The activation functions make the output of neurons nonlinear, which enables to solve more complex problems.

Supervised Learning. ANN

How Multi-Layer Artificial Neural Network works?

In the first step , for every neurons of hidden layers, the same process in the perceptron is applied:

- The weighted sum(z) is calculated.
- It is transmitted to related hidden neuron, then the activation function present in the neuron (ReLU or SELU) is applied.
- In the next step, the outputs of hidden layers are transmitted to output layer. As said before, the number of neurons depends on the problem in here:
  - Regression: consists of 1 neuron ,
  - Binary Classification: consists of 1 neuron,
  - Multi-label Classification: consists of 1 neuron per label,
  - Multi-class Classification: consists of 1 neuron per class in the output layer.

[0] ReLU (rectified linear unit) activation function operates by outputting the input directly if it is positive; otherwise, it outputs zero

[0] Scaled Exponential Linear Units, or SELUs, are activation functions that induce self-normalizing properties

Supervised Learning. ANN

How Multi-Layer Artificial Neural Network works?
The activation functions in neurons of output layer also depends on the task:

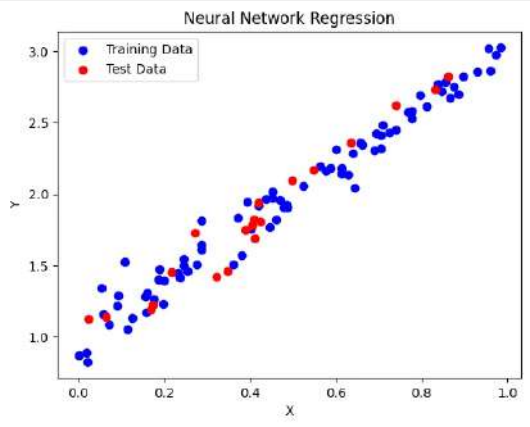
- Regression: None or ReLU/Softplus (if positive outputs) or Logistic/tanh (if bounded outputs),
- Binary Classification: Logistic(sigmoid) function,
- Multi-label Classification: Logistic(sigmoid) function,
- Multi-class Classification: Softmax function.

The main goal is to enable ANN to learn the most accurate weight values (so achieving most accurate result) with correct hidden layer and neuron numbers.

[0]Softplus $f(x) = log(1 + e^x)$, Sigmoid $f(x) = \frac{1}{1+e^{-x}}$
[0]Tanh $f(x) = \frac{2}{1+e^{-2x}} - 1$, Softmax $f(x_i) = \frac{e^{x_i}}{\sum e^{x_j}}$

Supervised Learning. Regression. ANN

In case the task is to predict the continuous variable, we solve a regression problem.



Supervised Learning. Regression. ANN

Steps to build a simple ANN model:

- Data Preparation: collect and preprocess your dataset, divide the data into training and testing sets
- Model Architecture:
  - Define the architecture of the neural network. This typically includes input, hidden, and output layers.
  - The input layer has nodes corresponding to your input features.
  - The hidden layers contain one or more layers with nodes (neurons) that apply non-linear transformations to the data.
  - The output layer has a single neuron, which provides the regression prediction.

Supervised Learning. Regression. ANN

Steps to build a simple ANN model:

- Select an appropriate loss function for regression tasks. Mean Squared Error (MSE) is a common choice:

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \widehat{y}_i)^2$$

- Training:
  - Feed the training data through the neural network and calculate the loss.
  - Use backpropagation and optimization algorithms (e.g., gradient descent) to update the network's weights to minimize the loss.
  - Continue this process for multiple epochs until the model converges or until a stopping criterion is met.

Supervised Learning. Regression. ANN

Steps to build a simple ANN model:

- Hyperparameter Tuning:
  - Experiment with hyperparameters like the number of layers, number of neurons in each layer, learning rate, batch size, and activation functions to optimize model performance.
  - Cross-validation can help assess how well your model generalizes to new data.
- Evaluation: use the test data to evaluate the model's performance by calculating metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or R-squared (coefficient of determination).

Supervised Learning. Artificial NN vs Deep NN

Criteria	ANN	DNN
Architecture	Typically consist of an input layer, one or a few hidden layers, and an output layer	Can have tens or even hundreds of hidden layers, allowing them to model more complex patterns.
Complexity and Representation	Generally suited for simpler tasks where the relationships in the data are not too complex.	Capable of learning high-level features and representations from raw data, making them ideal for image recognition, natural language processing,etc.
Computational Requirements	Training is generally faster and can be done on less powerful hardware	Often require specialized hardware, such as GPUs or TPUs, for efficient training.

Supervised Learning. Artificial NN vs Deep NN

Criteria	ANN	DNN
Layer Composition	The layers are typically simple, with each layer fully connected to the next. Layers used are usually restricted to basic fully connected (dense) layers	DNNs can contain a variety of layers, not just fully connected ones, e.g. Convolutional Layers (for images), Recurrent Layers (for sequence data like time series or text), Pooling Layers for downsampling data, Normalization Layers for stabilizing and speeding up training, etc.
Activation Functions and Regularization	Often uses simple activation functions like sigmoid or tanh, especially in older models. Simple regularization techniques such as L2	Employs more advanced activation functions like ReLU and its variants, which help to avoid issues like vanishing gradients. Uses sophisticated regularization techniques like dropout, batch normalization, and more, to combat overfitting and ensure stable training

Supervised Learning. Artificial NN vs Deep NN

Criteria	ANN	DNN
Feature Learning	Often rely on predefined features or simpler feature sets, as they do not have enough layers to automatically extract complex features from raw data	The multiple layers allow the network to learn complex features directly from the raw data, without needing manual feature engineering. This ability is crucial in tasks like image and speech recognition, where features are abstract and complex
Capacity	Limited capacity for learning due to fewer layers and parameters	High capacity for learning because of the large number of layers and parameters. Suitable for complex tasks and high-dimensional datasets, where intricate patterns and structures need to be learned

Supervised Learning. Testing

The purpose of evaluation is threefold:

- to determine which model is the most suitable for a task
- to estimate how the model will perform
- to convince users that the model will meet their needs

The most important part of the design of an evaluation experiment for a predictive model is ensuring that the data used to evaluate the model is not the same as the data used to train the model

Supervised Learning. Testing

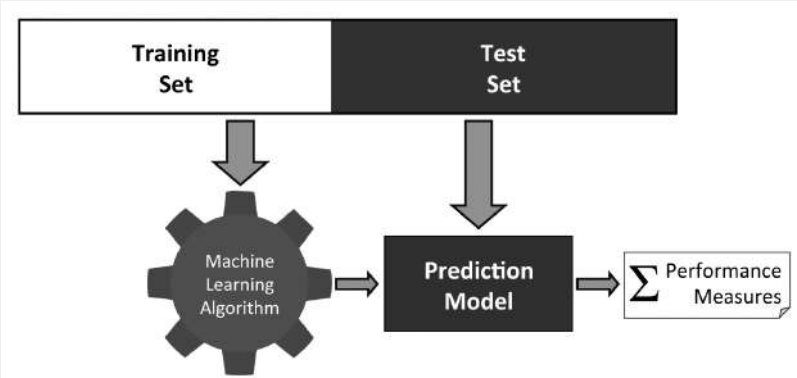


Figure: The process of building and evaluating a model using a hold-out test set

Supervised Learning. Testing

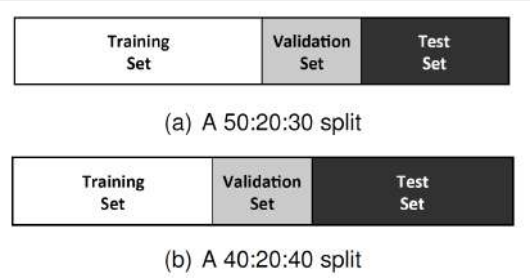


Figure: Hold-out sampling can divide the full data into training, validation, and test sets

Supervised Learning. Testing

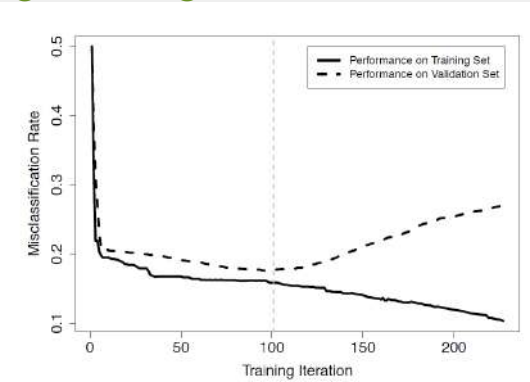


Figure: Using a validation set to avoid overfitting in iterative machine learning algorithms: it shows how the misclassification rate on a train set of changes as the training continues. At some point, overfitting will begin to occur, and the ability of the model to generalize well to new data (validation/test set) will diminish (here, overfitting has begun to occur at iteration = 100)

Supervised Learning. Testing

Using hold-out sampling requires that we have enough data available to make suitably large training, validation, test sets. Second, performance measured using hold-out sampling can be misleading if we happen to make a lucky split of the data that places the difficult instances into the training set and the easy ones into the test set. To address these two issues is k-fold cross validation is typically used.

Figure: The division of data during the k-fold cross validation process. Black rectangles indicate test data, and white spaces indicate training data

Supervised Learning. Testing

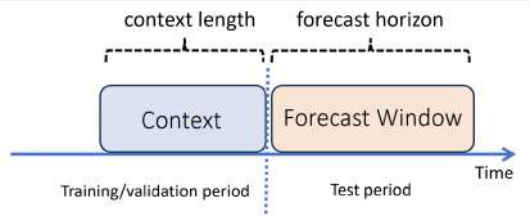


Figure: The out-of-time sampling process

Supervised Learning. Testing

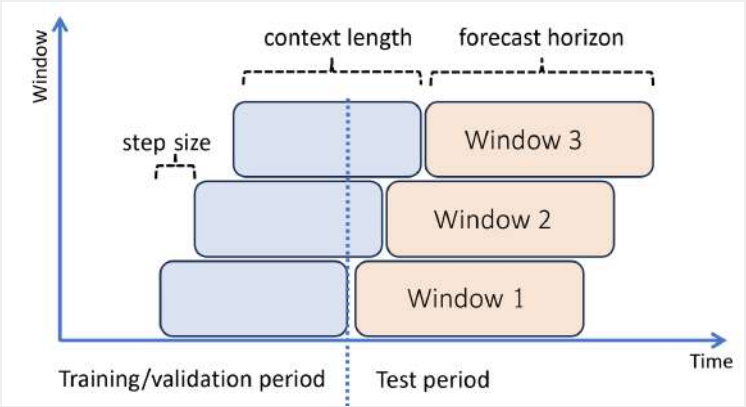


Figure: The out-of-time sampling process

Supervised Learning. Testing. Performance Measures. Categorical labels

In case of binary classification: the words "positive" and "negative" refer to the target and non-target classes.

A confusion matrix is an evaluative tool for displaying different prediction errors: a table that compares a model's predicted values with the actual values. The performance of a binary classifier can be represented by a 2×2 confusion matrix.

Table: The structure of a confusion matrix: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)

		Prediction	
		positive	**negative**
Target	**positive**	TP	FN
	negative	FP	TN

Supervised Learning. Testing. Performance Measures. Categorical labels

Figure: Correct predictions

Supervised Learning. Testing. Performance Measures. Categorical labels

There are two types of errors the model can make: the first type of error is called a false positive. The second is called a false negative.

Figure: Incorrect predictions

Supervised Learning. Testing. Performance Measures. Categorical labels

Technically, FP shows the number of incorrectly predicted positive cases; FN shows the number of incorrectly predicted negative cases.

Figure: The distinction between false positives (FP) and false negatives (FN) is important because the consequences of errors are different. You might consider one error less or more harmful than the other

Supervised Learning. Testing. Performance Measures. Categorical labels

$$\text{accuracy} = \frac{\text{number correct predictions}}{\text{total predictions}}$$

$$\text{misclassification rate} = \frac{\text{number incorrect predictions}}{\text{total predictions}}$$

Accuracy can be misleading if there is an imbalance in the number of examples of each class. For instance, if 95% of cases are not fraud, a classifier assigning all cases to the "not fraud" category could achieve 95% accuracy. Accuracy applies when there is a well-defined sense of right and wrong.

Supervised Learning. Testing. Performance Measures. Categorical labels

$$\text{TPR} = \frac{TP}{(TP + FN)}$$
$$\text{TNR} = \frac{TN}{(TN + FP)}$$
$$\text{FPR} = \frac{FP}{(TN + FP)}$$
$$\text{FNR} = \frac{FN}{(TP + FN)}$$

All these measures can have values in the range $[0, 1]$. TPR and TNR $\rightarrow$ 1, while FPR and FNR $\rightarrow$ 0. Also, FNR $= 1-$ TPR, and FPR $= 1-$ TNR.

Supervised Learning. Testing. Performance Measures. Categorical labels

$$\text{precision} = \frac{TP}{(TP + FP)}$$

$$\text{recall} = \frac{TP}{(TP + FN)} = \text{TPR} = \text{sensitivity}$$

$$\text{specificity} = \text{TNR} = \frac{TN}{(TN + FP)}$$

$$\text{F}_1\text{-measure} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$$

Recall, Precision, Specificity, F_1 range from 0 to 1 (where 1 is best).
Recall shows the share of true positive predictions made by the model out of all positive samples in the dataset, i.e. how many instances of the target class the model can find.

Supervised Learning. Testing. Performance Measures. Categorical labels

- The AUROC (Area Under the Receiver Operating Characteristic) score measures how well a classification model can distinguish between different classes.
- The ROC curve shows the performance of a classification model by plotting the rate of true positives against false positives as thresholds in a model are changed.
- AUROC scores range from zero to one, where a score of 50% indicates random-chance performance and 100% indicates perfect performance.
- We can interpret the AUROC as the probability that a positive example scores higher than a negative example.

Supervised Learning. Testing. Performance Measures. Categorical labels

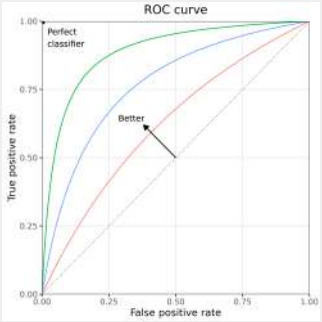


Figure: AUROC increases as it moves in the northwest direction, with more true positives and fewer false positives. It is useful for comparing the performance of different classifiers

.

Supervised Learning. Testing. Performance Measures. Categorical labels

The predictions shown in the confusion matrix and performance measures are based on a prediction score threshold of 0.5 (by default).

- This threshold can be changed, however, which leads to different predictions and a different confusion matrix and performance measures.
- Therefore, it makes sense to explore different thresholds.
- The ROC curve is drawn by plotting a point for every feasible threshold value and joining them, and thus could be used for optimizing the threshold value.

Supervised Learning. Testing. Performance Measures. Categorical labels

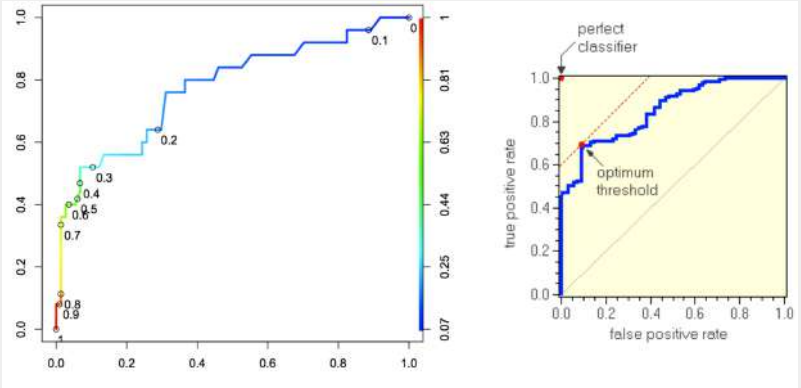


Figure: ROC curve with thresholds

Supervised Learning. Testing. Performance Measures. Continuous labels

- For continuous labels, the model testing process is the same as for categorical labels. We have a test set containing instances for which we know the actual label values, and we have a set of predictions made by a model.
- Recall that we have continuous labels when we have a regression problem.
- In continuous case, it is not possible to summarise by table, as we have many different actual labels in the test set and predicted labels, respectively.
- Suppose that y denotes the actual labels and $\hat{y}$ denotes the predicted labels.

Supervised Learning. Testing. Performance Measures. Continuous labels

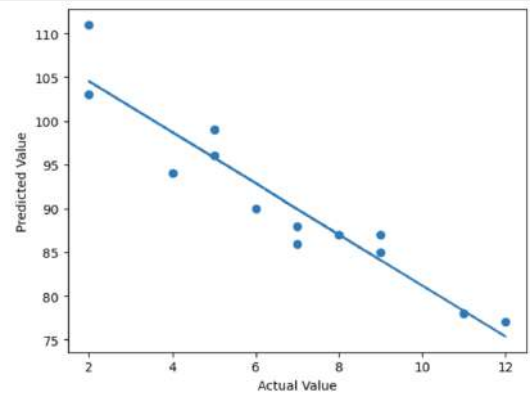


Figure: Actual labels vs. Predicted labels

Supervised Learning. Testing. Performance Measures. Continuous labels

The most popular performance metrics of regression model:

- MSE, RMSE, MAE, MAPE, R-squared, and Adjusted R-squared

Each metric provides a different lens to evaluate the performance of a regression model. Choosing the right metric depends on the specific context and objectives of our analysis. Understanding these metrics intuitively helps in selecting the most appropriate model and communicating its performance effectively.

Supervised Learning. Testing. Performance Measures. Continuous labels

Mean Squared Error (MSE) performance measure captures the average difference between the expected label values in the test set and the values predicted by the model:

$$\text{MSE} = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2$$

Accordingly, we have Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}$$

Supervised Learning. Testing. Performance Measures. Continuous labels

- RMSE is in the same units as the target variable being predicted, while MSE is in squared units. This makes RMSE more interpretable.
- Similarly, RMSE is scale-dependent, meaning it is related to the scale of the data. When comparing model performance across datasets with different scales, RMSE can provide a more intuitive sense of the error magnitude relative to the scale of the data.
- While both MSE and RMSE are sensitive to large errors due to the squaring of the residuals, MSE tends to be more sensitive. This is because the squaring of errors before averaging, followed by taking the square root, magnifies the impact of larger errors more than smaller ones.

Supervised Learning. Testing. Performance Measures. Continuous labels

Mean Absolute Error (MAE) is the average absolute difference between the predicted and actual values.

$$\text{MAE} = \frac{1}{k}\sum_{i=1}^{k} |y_i - \hat{y}_i|$$

- MAE is less sensitive to outliers compared to MSE and RMSE. MAE, by taking the absolute value of errors, treats all deviations from the true values equally, providing a more robust error metric in such cases.
- MAE is intuitively easier to understand since it's simply the average error in the same units as the data. However, it is scale-dependent measure and cannot be used to compare different models.

Supervised Learning. Testing. Performance Measures. Continuous labels

Mean Absolute Percentage Error (MAPE) expresses the error as a percentage of the actual value.

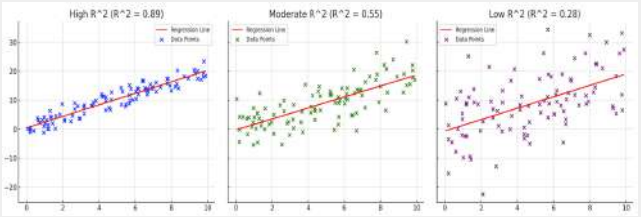
$$\text{MAPE} = \frac{100\%}{k} \sum_{i=1}^{k} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- MAPE expresses the error as a percentage, providing a relative measure of error. This is particularly useful in scenarios where it's important to understand the size of the error in proportion to the actual value.
- Unlike MAE or MSE/RMSE, MAPE offers a scale-independent view of the error. This makes it especially valuable for comparing the performance of models across datasets with different scales or units.
- When the actual values zero or close to zero, the division within the MAPE formula yields an undefined or inadequate results.

Supervised Learning. Testing. Performance Measures. Continuous labels

R-Squared shows the proportion of variance explained by the model.

$$\text{R-Squared} = 1 - \frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{k}(y_i - \bar{y})^2}$$



A high R-Squared (max=1) means that model very closely predict the actual values.

Supervised Learning. Testing. Performance Measures. Continuous labels

- Unlike MAE, MSE, or RMSE, R-Squared is not affected by the scale of the data. This allows for easier comparison between models on different scales and makes it a useful tool in model selection.
- R-Squared values range from 0 to 1, with 0 indicating that the model explains none of the variability of the response data around its mean and 1 indicating that it explains all the variability.
- R-squared tends to increase as more predictors are added to the model. Thus, Adjusted R-Squared modifies the R-Squared formula to address this issue, and it increases only if the new predictor improves the model. This makes Adjusted R-Squared a more reliable metric, particularly when comparing models with a different number of predictors.

Supervised Learning. Testing. Performance Measures. Continuous labels

Graphically, in case we have enough test data, it makes sense to plot the distribution of the residuals $= y_i - \hat{y}_i$.

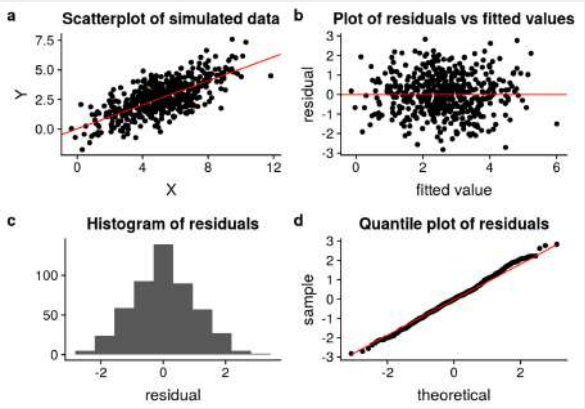


Table of Contents

Unsupervised Learning

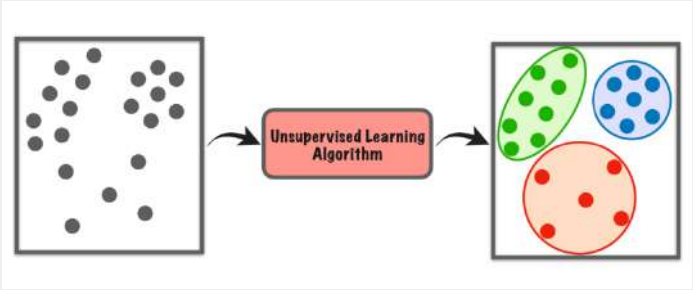
In the previous topic, we learned supervised machine learning in which models are trained using labeled data under the supervision of training data.

- Unsupervised learning is learning from unlabeled data.
- Instead of matching its inputs to the correct labels, the model must identify patterns within the data to help it understand the underlying relationships between the variables.
- As no labels are provided, a model is left to its own devices to discover valuable patterns in the data.

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data.

Unsupervised Learning

The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.



Unsupervised Learning

Challenges of Unsupervised Learning:

- Evaluation: Assessing the performance of unsupervised learning algorithms is difficult without predefined labels or categories.
- Interpretability: Understanding the decision-making process of unsupervised learning models is often challenging.
- Overfitting: Unsupervised learning algorithms can overfit to the specific dataset used for training, limiting their ability to generalize to new data.
- Computational complexity: Some unsupervised learning algorithms, particularly those dealing with high-dimensional data or large datasets, can be computationally expensive.

Unsupervised Learning

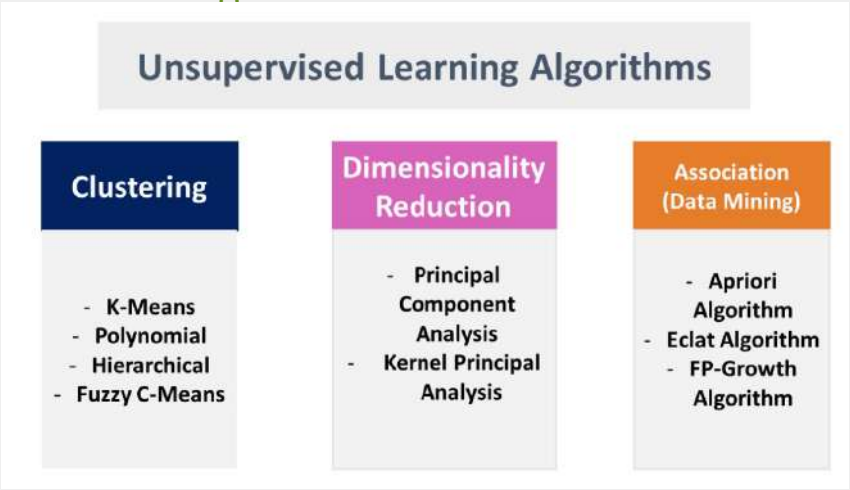


Figure: Types of Unsupervised Learning

Unsupervised Learning. Clustering

- Clustering in unsupervised machine learning is the process of grouping unlabeled data into clusters based on their similarities.
- The goal of clustering is to identify patterns and relationships in the data without any prior knowledge of the data's meaning.
- Broadly this technique is applied to group data based on different patterns, such as similarities or differences, our machine model finds. These algorithms are used to process raw, unclassified data objects into groups. For example, to group clients based on the input parameters provided by our data.

Unsupervised Learning. Clustering

Some common clustering algorithms:

- K-means Clustering: Partitioning Data into K Clusters
- Hierarchical Clustering: Building a Hierarchical Structure of Clusters
- Density-Based Clustering (DBSCAN): Identifying Clusters Based on Density
- Fuzzy Clustering: Assign a membership degree between 0 and 1 for each data point for each cluster (a data point to belong to more than one cluster with different degrees of membership)
- Spectral Clustering: Utilizing Spectral Graph Theory for Clustering

Unsupervised Learning. Association Rule Learning

- Association rule learning is also known as association rule mining is a common technique used to discover associations in unsupervised machine learning.
- This technique is a rule-based ML technique that finds out some very useful relations between parameters of a large data set.
- This method is very popular in market basket analysis to better understand the relationship between different products, e.g. if a customer buys milk, then he may also buy bread, eggs, or butter.
- Another example from a dataset of medical records of patients. Rule: High Blood Pressure, Obesity $\rightarrow$ Heart Disease. Explanation: This rule may indicate that patients with high blood pressure and obesity have a higher likelihood of developing heart disease. This can be used to identify at-risk patients and apply preventative measures.

Unsupervised Learning. Association Rule Learning

- Apriori Algorithm: A Classic Method for Rule Induction
- FP-Growth Algorithm: An Efficient Alternative to Apriori
- Eclat Algorithm: Exploiting Closed Itemsets for Efficient Rule Mining
- Efficient Tree-based Algorithms: Handling Large Datasets with Scalability

Unsupervised Learning. Dimensionality Reduction

- Dimensionality reduction is the process of reducing the number of features in a dataset while preserving as much information as possible.
- This technique is useful for improving the performance of LM algorithms and for data visualization.

Unsupervised Learning. Dimensionality Reduction

Examples of dimensionality reduction algorithms:

- Principal Component Analysis (PCA): Linear Transformation for Reduced Dimensions
- Linear Discriminant Analysis (LDA): Dimensionality Reduction for Discrimination
- Non-negative Matrix Factorization (NMF): Decomposing Data into Non-negative Components
- Locally Linear Embedding (LLE): Preserving Local Geometry in Reduced Dimensions
- Isomap: Capturing Global Relationships in Reduced Dimensions

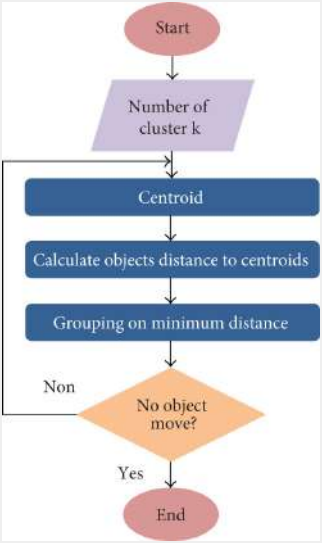
Unsupervised Learning. K-means clustering

The K-Means algorithm clusters data by trying to separate samples in n groups/clusters of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares

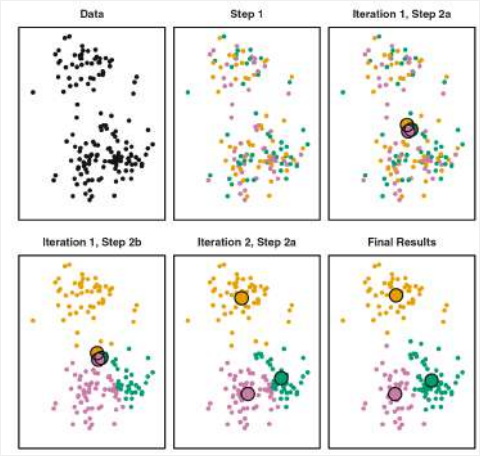
$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2),$$

where $x_i \in X$ - data points, n - sample size, K - number of clusters, C - clusters, each described by the centroid/mean μ_j of data points in the cluster.

Unsupervised Learning. K-means clustering



Unsupervised Learning. K-means clustering

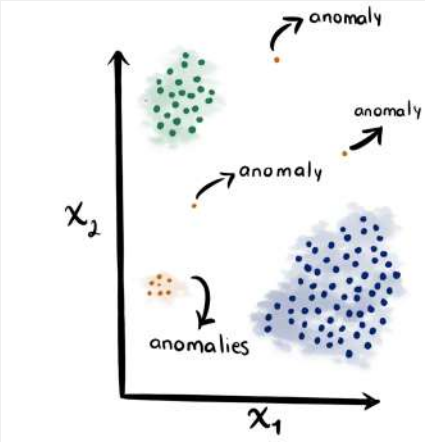


Unsupervised Learning. K-means clustering

Tasks to be solved:

- Try different similarity measures that best fit your data
- Replicate the choice of initial clusters few times
- Find the best number K of clusters
- Validate the quality of obtained clusters, e.g. using silhouette coefficient, which measures the similarity of a data point within its cluster (cohesion) compared to other clusters (separation)
- Consider feature importance

Unsupervised Learning. Anomaly Detection

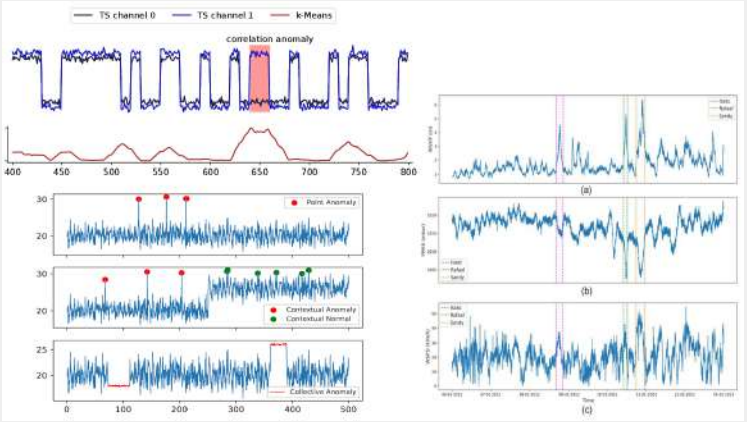


Unsupervised Learning. K-means clustering for Anomaly Detection

Different strategies available:

- Set Anomaly Threshold: Define a threshold for anomaly detection. Data points with distances above this threshold are considered anomalies. The threshold can be set based on statistical methods or domain knowledge.
- Focus on small clusters: The small clusters with less than a threshold (1% of total number of data points) or isolation data points not belong to any cluster

Unsupervised Learning. Anomaly Detection in Time Series



Techniques like isolation forests, clustering-based approaches, and autoencoders have proven effective in unsupervised anomaly detection for time series data.

References

- Julia Simpson (2024) Introduction to Artificial Intelligence (AI) Technology https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/2024-wttc-introduction-to-ai.pdf
- Fredrik Filipsson (2024) Artificial Intelligence: An Introduction to AI Fundamentals https://redresscompliance.com/artificial-intelligence-an-introduction-to-ai-fundamentals/
- Dan Hendrycks (2024) Introduction to AI Safety, Ethics and Society https://www.aisafetybook.com/textbook
- MarkovML (2024) Exploratory Data Analysis in Predictive Modeling: Techniques & Strategies https://www.markovml.com/blog/exploratory-data-analysis
- GeeksforGeeks https://www.geeksforgeeks.org/
- Alexander Jung (2022) Machine Learning: The Basics https://alexjungaalto.github.io/MLBasicsBook.pdf
- John D. Kelleher, Brian Mac Namee and Aoife D'Arcy (2015) Fundamentals of Machine Learning for Predictive Data Analytics https://mitpress.mit.edu/9780262029445/fundamentals-of-machine-learning-for-predictive-data-analytics/

WELLBEING@ARTIFICIAL INTELLIGENCE
2023-1-PT01-KA220-SCH-00155442

AI Concepts and Definitions

Kristina Sutiene and Liepa Bikulciene

Kaunas University of Technology

Darica, Türkiye
Sep 16-20, 2024

Co-funded by the European Union