U3 MACHINE LEARNING ALGORITHMS

U3.E1 SUPERVISED MODELS

Machine Learning Engineer

January 2021, Version 1

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

LEARNING OBJECTIVES

The student is able to

MLE.U3.E1.PC1	Define and understand supervised learning.
MLE.U3.E1.PC2	Understand the differences between regression and classification.
MLE.U3.E1.PC3	Know some of the most commonly used algorithms in supervised learning.
MLE.U3.E1.PC4	Categorize each algorithm within the regression and classification.
MLE.U3.E1.PC5	Know the domain applications of supervised models.

Supervised Learning method trains a function (or algorithm) to compute output variables based on a given data in which both input and output variables are known.

The goal of a learning process is to find a function that minimizes the risk of a prediction error that is expressed as a difference between the actual and the computed output values when tested on a dataset. In such cases, the learning process may be controlled by a predetermined acceptable error threshold.

DRIVES
Development and Research on Innovative Vocational Education Skills

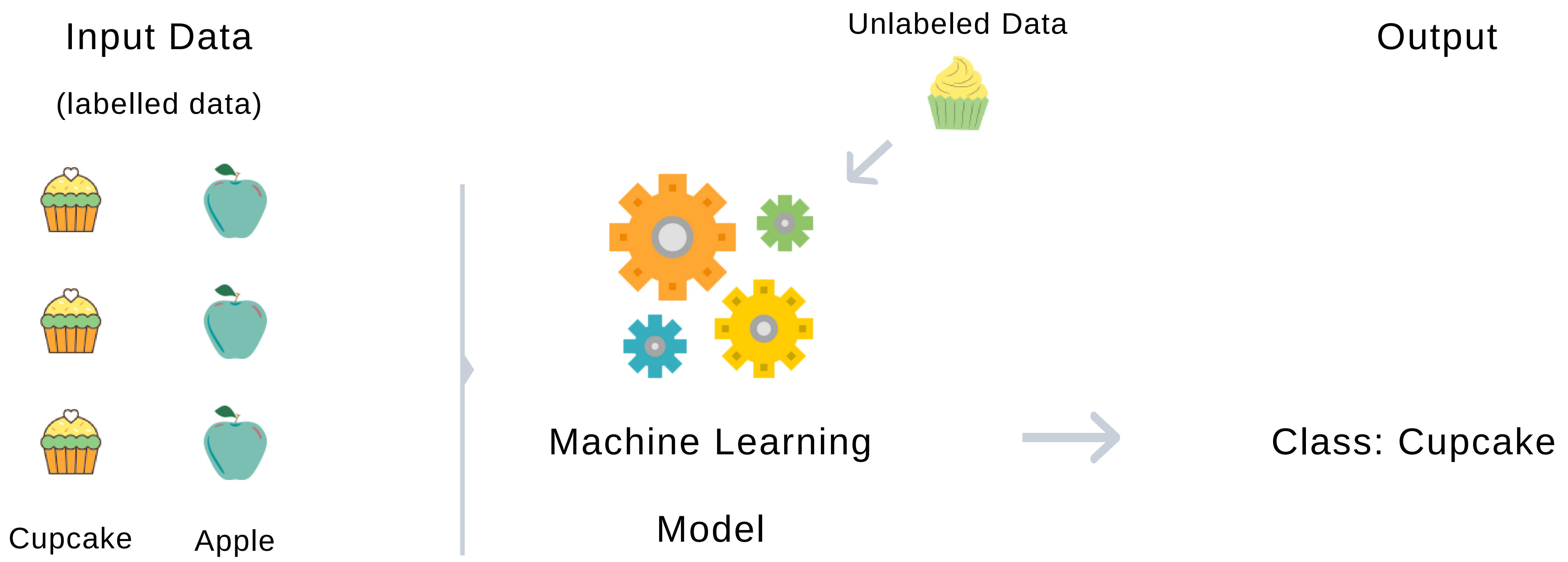
As the name suggests, here the learning occurs under supervision.

The training dataset is labelled

You already know how the output should be

There is a relationship between the input and the output

SUPERVISED LEARNING



ANALOGY WITH DRIVING LESSONS

The supervised learning process can be seen as a collection of guidelines provided by a driving instructor to explain what should be done (output variables) in different situations (input variables). These guidelines are adapted by a student driver and turned into a driver behavior.

The predetermined thresholds can be seen as the standards to pass the driving exam. In this case, the student driver knows the standard way to drive (i.e., actual output) and steps to achieve it (i.e., actual inputs) before he/she starts the driving lessons.

For the student driver, it becomes an iterative process to achieve acceptable performance. In each iteration, the student makes mistakes that are corrected by the driving instructor (i.e., training the model). This iterative process ends when the student gets the driving license (i.e., the model achieves a satisfactory performance).

IMPORTANT NOTIONS

Data: a set of labeled data records/instances/cases/examples <x_i, y>

Training Set

The portion of the dataset from which the ML algorithms discover or learn relationships between the features and the target attribute. The training set is labeled in supervised learning.

Validation Set

Another portion of the dataset to which the ML algorithms are applied in order to check how well they identify relationships between the known outcomes of the target variable and the other features of the dataset.

Test/Holdout Set

The subset of the dataset that provides a final estimate of the performance of the ML models after they have been trained and validated. Holdout sets should never be used to make decisions about which algorithms to use or to improve or tune algorithms.

The training set should not be used in testing and the test set should not be used in learning!

↓

An unseen test set provides an unbiased estimation of the model's performance.

DRIVES
Development and Research on Innovative Vocational Education Skills

IMPORTANT NOTIONS

Features: a measurable property of the object being analysed. In datasets, features appear as columns. Features are also sometimes referred to as “variables” or “attributes”.

# Glucose	# BloodPres...	# SkinThickn...	# Insulin	# BMI	# Age	# Outcome
148	72	35	0	33.6	50	1
85	66	29	0	26.6	31	0
183	64	0	0	23.3	32	1
89	66	23	94	28.1	21	0
137	40	35	168	43.1	33	1
116	74	0	0	25.6	30	0
78	50	32	88	31	26	1
115	0	0	0	35.3	29	0
197	70	45	543	30.5	53	1

https://www.kaggle.com/uciml/pima-indians-diabetes-database

DRIVES
Development and Research on Innovative Vocational Education Skills

IMPORTANT NOTIONS

Features:

The quality of the dataset's features has a huge influence on the quality of the insights that will be achieved when using the dataset for ML.

In addition, different business problems within the same industry do not necessarily require the same features, which is why it is important to have a strong understanding of the business objectives of the data science project.

The quality of dataset's features can be improved with processes such as **feature selection** and

IMPORTANT NOTIONS

Features:

Feature Selection

eliminates irrelevant or redundant columns from the dataset without sacrificing accuracy.

1. Reduce the chance of overfitting;

2. Boost the run speed of the algorithm by reducing the CPU, I/O, and RAM load the production system needs to construct and use the model;

3. Increase the interpretability of the model by identifying the most informative factors that drive the results of the model.

Feature Engineering

constructs additional variables to the dataset to improve the performance and accuracy of the ML model.

1. Provide a deeper understanding of the data;

2. Improve the predictive power of the ML model;

3. Deliver more valuable insights;

DRIVES
Development and Research on Innovative Vocational Education Skills

IMPORTANT NOTIONS

Experimentation cycle:

- Data gathering
- Data preparation
- Learn parameters on the training set
- Hyper-parameter tuning on the validation set
- Compute evaluation metrics on the test set and make predictions

DRIVES
Development and Research on Innovative Vocational Education Skills

Supervised learning can further be categorized as Regression and Classification problems.

In the case of a **classification** problem, the aim of the ML algorithm is to categorize or classify the inputs based on the training dataset. The training dataset in a classification problem includes a set of input:output pairs categorized in classes.

- Is a given patient with covid-19 or healthy?
- Is this an image of a dog, a cat, or a horse?

For a **regression** problem, the goal of the ML algorithm is to develop a relationship between outputs and inputs using a continuous function to help machines understand how outputs change for inputs. The relationship between output variables and input variables can be defined by various mathematical functions such as linear, nonlinear, and logistic.

- What is the company's revenue by the end of the year?
- What is the probability that tomorrow will rain?

While classification methods are used when the output is of categorical nature, the regression methods are used for continuous output.

input data → **CLASSIFICATION** → one or more discrete labels (gender, marital status, nationality)

input data → **REGRESSION** → a continuous quantity (weight, height, price, age, temperature)

DRIVES
Development and Research on Innovative Vocational Education Skills

CLASSIFICATION	REGRESSION
The discovery of models is done into predefined classes.	The discovery of models is done into values.
Involves prediction of discrete values.	Involves prediction of continuous values.
The nature of the predicted data is unordered.	The nature of the predicted data is ordered.
The method of calculation is measuring accuracy.	The method of calculation is measurement of root mean square error.
The output variable takes class labels.	The output variable takes continuous labels.

DRIVES
Development and Research on Innovative Vocational Education Skills

CLASSIFICATION

- Decision Trees
- Naïve Bayes
- Support Vector Machines
- Random Forest
- Logistic Regression
- Ensemble Methods
- K-Nearest Neighbors
- Kernel SVM

REGRESSION

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

- Decision Trees are classic learning models and one of the most widely used for inductive inference.

- They are closely associated to the fundamental notion of "divide and conquer" in computer science.

- A decision tree is a flowchart-like tree structure where an internal node represents an attribute, a branch represents a decision rule, and each leaf node represents the outcome.

- Decision trees can be used to solve regression and classification problems.

- The aim of a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data, i.e., the training data.

TYPES OF DECISION TREE ALGORITHMS

Decision tree types are based on the type of target variable we have. It can be of two types:

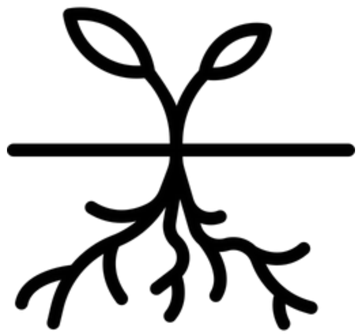
CATEGORICAL VARIABLE DECISION TREE

Decision trees that have a categorical target variable are therefore called a Categorical Variable Decision Tree.

CONTINUOUS VARIABLE DECISION TREE

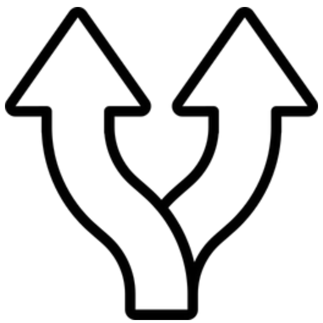
Decision trees that have a continuous target variable are called a Continuous Variable Decision Tree.

TERMINOLOGY



ROOT NODE

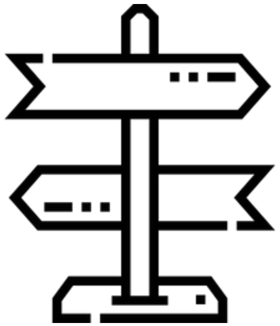
The topmost node in the decision tree is known as the root node, which represents the whole population or sample and is divided into two or more sets.



DIVISION

It is the process of dividing a node into two or more subnodes.

TERMINOLOGY



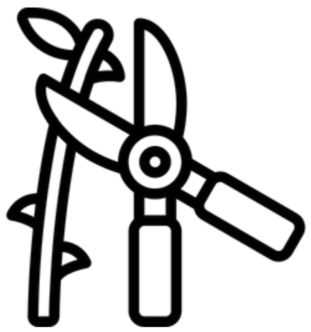
DECISION NODE

When a subnode is divided into other subnodes, it is called a decision node.

TERMINAL NODE

Nodes that do not split are called Sheet or Terminal nodes.

TERMINOLOGY



PRUNING

The process of removing subnodes from a decision node is called pruning, which is the opposite process of dividing.

BRANCH / SUBTREE

A subsection of the tree is called a branch or subtree.

PARENT AND CHILD NODE

A node divided into subnodes is called the parent node of these subnodes, while the subnodes are the children of the parent node.

DRIVES
Development and Research on Innovative Vocational Education Skills

ASSUMPTIONS

01

In the beginning, the whole training set is considered the root.

02

Resource values are more convenient to be categorical. If the values are continuous, they will be discredited before the model is built.

03

Records are distributed recursively based on the attribute values.

04

The order for placing attributes as root or node within the tree is done using one statistical approach.

DRIVES
Development and Research on Innovative Vocational Education Skills

→ For a given training set, there are many trees that code it without any error.

→ It is computationally infeasible to consider all the trees. Finding the simplest and smallest tree is a NP-complete problem (Hyafil & Rivest 1976; Quinlan 1986).

→ Hence, it is necessary to resort to a greedy heuristic:
- Start from an empty decision tree;
- Split on next best attribute;
- Recurse.

How to determine which attribute is the best?

↓

UNCERTAINTY MEASURES

UNCERTAINTY MEASURES

- ENTROPY
- INFORMATION GAIN
- GINI INDEX
- GAIN RATIO
- CHI SQUARE
- VARIANCE REDUCTION

RANDOM FOREST ALGORITHM

- It consists of a large number of individual decision trees that function as a whole.
- A large number of relatively unrelated models (trees), functioning as a committee, will perform better than any of the individual constituent models.
- The fundamental concept behind the random forest is a simple but powerful concept - the wisdom of the crowds.

DRIVES
Development and Research on Innovative Vocational Education Skills

RANDOM FOREST ALGORITHM

The prerequisites for the random forest to perform successfully are:

1. There must be some real sign in our characteristics so that the models built using these characteristics do better than random guessing.

2. The forecasts (and therefore errors) made by individual trees need to have low correlations with each other.

- Supervised learning can naturally be studied from a probabilistic point of view.

- Naive Bayes classifiers are a collection of classification algorithms from the family of "probabilistic classifiers" based on the Bayes's theorem. The main feature of these algorithms, and also the reason for receiving "naive" in the name, is that they completely ignore the correlation between variables (features).

- These algorithms predict the probability of a given record belonging to each class. The class with the highest probability is considered to be the class.

Given a test example *d* with observed attribute values a_1 through a_k, the classification is basically done by computing the following *posteriori* probability. The prediction is the class c_j such that Eq. 1 is maximal.

Where:

- A_1 through A_k are the attributes with discrete values;
- C is the target class;
- $Pr(C=c_j)$ is the class *prior* probability;
- $P(A_1=a_1,...,A_k=a_k)$ is the same for every class;

$$\Pr(C = c_j \mid A_1 = a_1, \ldots, A_{|A|} = a_{|A|})$$
$$= \frac{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = c_j)\Pr(C = c_j)}{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|})}$$
$$= \frac{\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = c_j)\Pr(C = c_j)}{\sum_{r=1}^{|C|} \Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = c_r)\Pr(C = c_r)} \quad (1)$$

- Hence, the focus is $P(A_1=a_1,...,A_k=a_k \mid C=c_j)$, which can be written as $Pr(A_1=a_1|A_2=a_2,...,A_k=a_k, C=c_j)$* $Pr(A_2=a_2,...,A_k=a_k \mid C=c_j)$. Recursively, the second factor above can be written in the same way, and so on.

- All attributes are conditionally independent given the class $C = c_j$. Formally, $Pr(A_1=a_1 \mid A_2=a_2, ..., A_{|A|}=a_{|A|}, C=c_j) = Pr(A_1=a_1 \mid C=c_j)$ and so on for A_2 through $A_{|A|}$. i.e.,

$$\Pr(A_1 = a_1, \ldots, A_{|A|} = a_{|A|} \mid C = c_i) = \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_j)$$

$$\downarrow$$

$$\Pr(C = c_j \mid A_1 = a_1, \ldots, A_{|A|} = a_{|A|})$$

$$= \frac{\Pr(C = c_j) \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_j)}{\sum_{r=1}^{|C|} \Pr(C = c_r) \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_r)}$$

DRIVES
Development and Research on Innovative Vocational Education Skills

How to estimate P(A_i = a_i| C=c_j)?

It is only necessary to calculate the numerator because the denominator is the same for each class. Thus, given a test example, it is necessary to calculate the following to determine the most probable class for the test instance.

$$c = \arg\max_{c_j} \Pr(c_j) \prod_{i=1}^{|A|} \Pr(A_i = a_i \mid C = c_j)$$

R	**U**	**C**
v	g	h
o	n	h
v	z	h
o	z	m
v	n	m
o	g	m

DRIVES
Development and Research on Innovative Vocational Education Skills

How to estimate $P(A_i = a_i | C=c_j)$?

1. compute all probabilities

Pr(C=h) = 1/2 | Pr(C=m) = 1/2

Pr(R=v|C=h) = 2/3 | Pr(R=v|C=m) = 1/3

Pr(R=o|C=h) = 1/3 | Pr(R=o|C=m) = 2/3

Pr(U=g|C=h) = 2/3 | Pr(U=g|C=m) = 1/3

Pr(U=z|C=h) = 1/3 | Pr(U=z|C=m) = 1/3

Pr(U=n|C=h) = 0 | Pr(U=n|C=m) = 1/3

R	U	C
v	g	h
o	g	h
v	z	h
o	z	m
v	n	m
o	g	m

2. For a new instance R=o U=n C=?

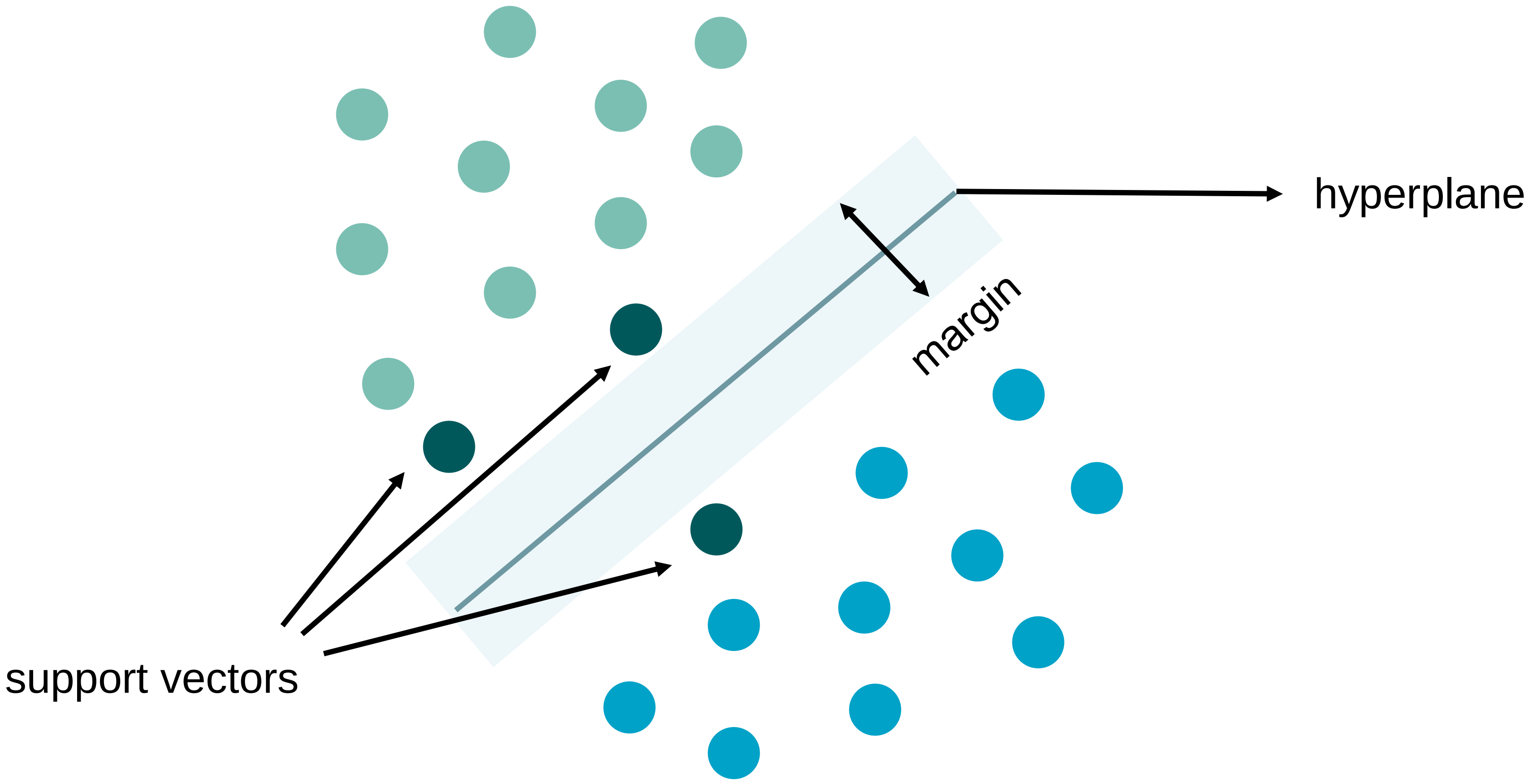
- For C=h: 1/2*1/3*0 = 0
- For C=m: 1/2*2/3*1/3 = 2/18 = 1/9

C = m is more probable → m is the final class!

DRIVES
Development and Research on Innovative Vocational Education Skills

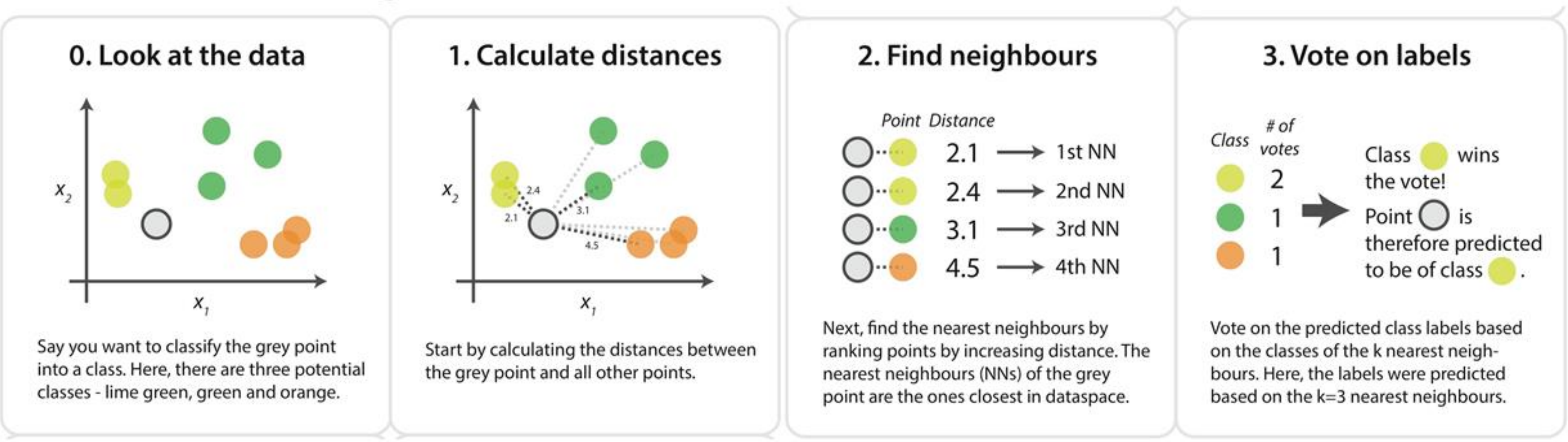
- Naïve Bayesian learning assumes that all attributes are categorical, so numeric attributes need to be discretized.
- The missing values are ignored in the Naïve Bayes classifier.
- Naïve Bayes is efficient and easy to implement.
- The main drawback is the assumption of class conditional independence. Thus, the loss of accuracy occurs when the assumption is seriously violated, i.e., in the case of highly correlated datasets.

SVMs have the ability to solve classification and regression problems, acquiring the ability to generalize in the training stage. The algorithm creates a line or a hyperplane that separates the data into classes in order to find an optimal boundary between the possible outputs.



K NEAREST NEIGHBOR (K-NN) ALGORITHMS

The k-NN algorithm assumes similar things are close together. It is therefore a classifier where learning is based on "how similar" the data is (one vector) to another.



- Linear Regression is a method for modeling a target value based on independent predictors.
- Regression techniques differ mainly based on the number of independent variables and the type of relationship between independent and dependent variables.
- It is mainly used to predict and discover the cause and effect relationship between the variables.
- Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

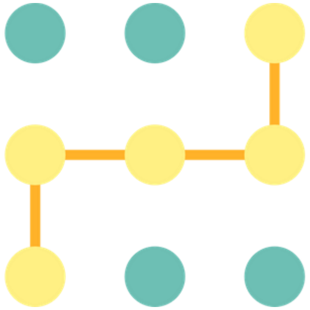
SUPERVISED LEARNING APPLICATIONS

Bioinformatics

Cheminformatics

Database

Marketing



Pattern

Recognition

Object

Recognition

Information Retrieval

Spam Detection

Speech Recognition

Optical Character Recognition

Handwriting Recognition

ADVANTAGES OF SUPERVISED LEARNING

DRIVES
Development and Research on Innovative Vocational Education Skills

01 The input data is very well known and is labeled.

02 One can determine the number of classes he/she wants to have.

03 The answers in the analysis and output of the algorithm are likely to be known due to the fact that all the classes used are known.

04 It allows the algorithm training to distinguish between different classes where one can set an ideal decision boundary.

05 The results produced by the supervised method are more accurate and reliable.

DISADVANTAGES OF SUPERVISED LEARNING

DRIVES
Development and Research on Innovative Vocational Education Skills

01 Supervised learning can be a complex method compared to unsupervised learning. The main reason is that it is required to understand and label inputs in supervised learning.

02 It doesn't take place in real time, while unsupervised learning is in real time. Supervised machine learning uses offline analysis.

03 A lot of computing time is needed for training.

04 When faced with dynamic, big and growing data, there is no assurance of the labels that will pre-define the rules, which can be a real challenge.

- In Supervised Learning the learning occurs under supervision, i.e., the training data is labelled and both input and output variables are known.
- Supervised learning can further be categorized as Regression and Classification.
- In Classification, the discovery of models is done into predefined classes, while in Regression, the discovery of models is done into values.
- Classification methods are used when the output is of categorical nature.
- Regression methods are used for continuous output.
- Decision Trees, Naïve Bayes, Support Vector Machines, Random Forest and k-Nearest Neighbor are popular ML algorithms used in Classification tasks.
- Simple Linear Regression, Multiple Linear Regression, and Polynomial Regression are popular techniques used in Regression tasks.

REFERENCES

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

Phyu, T. N. (2009, March). Survey of classification techniques in data mining. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, No. 5).

Kesavaraj, G., & Sukumaran, S. (2013, July). A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICCCNT.2013.6726842.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168). https://doi.org/10.1145/1143844.1143865.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning* (pp. 9-41). Springer, New York, NY.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia* (pp. 21-49). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2.

Liu, B. (2011). Supervised learning. In *Web data mining* (pp. 63-132). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19460-3_3.

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: supervised methods. https://doi.org/10.1038/nmeth.4551.

REFERENCES

Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., & Dera, D. (2017). Machine learning in transportation data analytics. In *Data analytics for intelligent transportation systems* (pp. 283-307). Elsevier. https://doi.org/10.1016/B978-0-12-809715-1.00012-2.

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). Ieee.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.

Muhammad, I., & Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*, *5*(3).

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. *Behavior Therapy*, *51*(5), 675-687. https://doi.org/10.1016/j.beth.2020.05.002.

REFERENCE TO AUTHORS

Diana Ferreira

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

iD 0000-0003-2326-2153

Regina Sousa

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

iD 0000-0002-2988-196X

José Machado

- Associate Professor with Habilitation at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

iD 0000-0003-4121-6169

António Abelha

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0001-6457-0756

Victor Alves

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0003-1819-7051

REFERENCE TO AUTHORS

This Training Material has been certified according to the rules of **ECQA – European Certification and Qualification Association.**

The Training Material was developed within the international job role committee **“Machine Learning Engineer”:**

UMINHO – University of Minho (https://www.uminho.pt/PT)

The development of the training material was partly funded by the EU under Blueprint Project DRIVES.

Thank you for your attention

DRIVES project is project under **[The Blueprint for Sectoral Cooperation on Skills in Automotive Sector]**, as part of New Skills Agenda.

The aim of the Blueprint is **to support an overall sectoral strategy and to develop concrete actions to address short and medium term skills needs.**

Follow DRIVES project at:

More information at:

www.project-drives.eu

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.