Aalto University
School of Business

Weka

Lecturer: Associ. Prof. Yong Liu (yong.liu@aalto.fi);

Outline

1. What is Weka?
2. Merits of Weka
3. Comparison of Weka with machine learning in Python/R
4. Weka tutorial design

What is Weka?

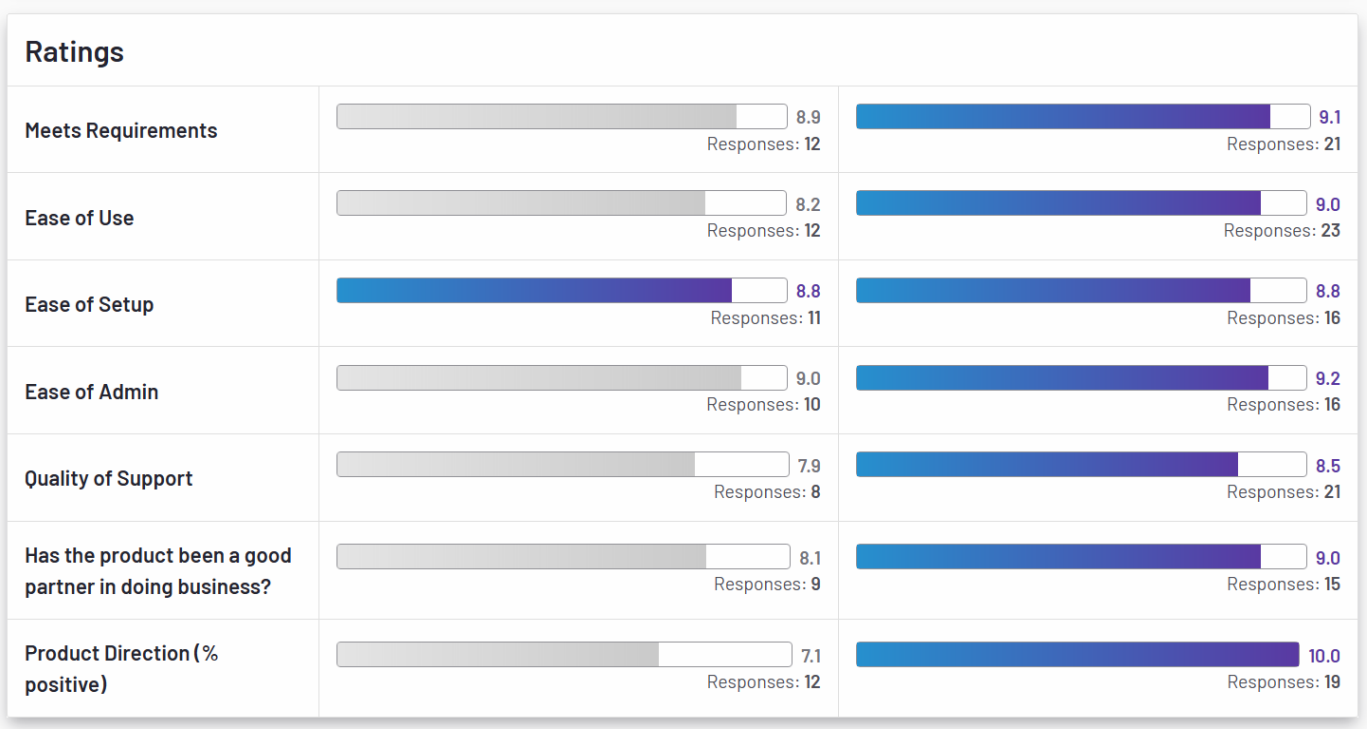
- **It is a collection of machine-learning algorithms for data mining tasks.**
- **It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.**
- **It is open-source software issued under the GNU General Public License, developed at the University of Waikato, New Zealand**

The **weka** is a flightless bird in New Zealand.

Merits of Weka

- **It offers tens of different machine-learning algorithms for use.**
- **More machine-learning algorithms can be downloaded through its package management function.**
- **It does not require a background in programming!!**

Weka vs. machine learning in Python/R



https://www.g2.com/compare/weka-vs-machine-learning-in-python

Weka vs. machine learning in Python/R

- **Weka has a flat learning curve.**
- **You can learn how to do and actually do machine learning in a couple of days with Weka.**

- **Python/R has a steep learning curve.**

coursera.org
https://www.coursera.org › ... › Data ▾ Käännä tämä sivu

How Long Does it Take to Learn Python? (+ Tips for Learning)

13.7.2022 — In general, it takes **around two to six months** to learn the fundamentals of Python. But you can learn enough to write your first short ...

- **Debugging takes lots of time and causes stress!**

This Is How Long It Takes To Learn Machine Learning

Updated: 02/18/21 • 9 min read

Analytics

Learning machine learning is much like learning any new skill. It is dependent on factors such as your existing knowledge of machine learning topics, computer literacy, attitude, and the time you have to devote to learning it. With that in mind, here are some guidelines.

To learn machine learning, expect it to take 6 months if you dedicate 40 hours a week to just learning it. Expect to study 2 years at 10 hours a week or through practical experience weaved into your job to have a solid foundation with machine learning.

Link to the post: https://coffeebreakdata.com/how-long-to-learn-machine-learning/

Weka vs. machine learning in Python/R

Theory of Machine Learning ?

Fundamental of Python/R → Learning of Specific Packages of Machine Learning → Implementation of Machine Learning Package

2-6 months | **A few hours or days** | **One hour (if lucky) ++**

What if you need to compare the performance of, for instance, five different machine learning algorithms on the dataset?

What will we learn?

- **Implementation of machine learning via Weka**
- **A very brief introduction to machine learning**
  - It is highly recommended to learn the theory of machine learning from the data science course

Summary

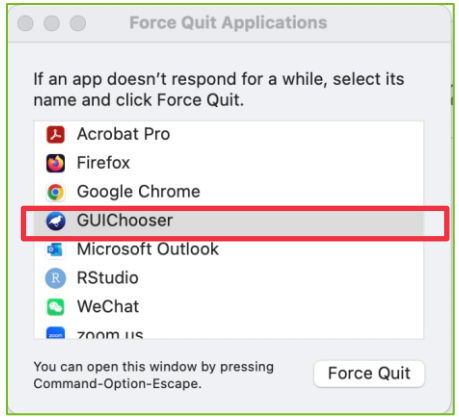
1. What Weka is.
2. Merits of Weka
3. Differences between Weka vs. programming-based machine learning tools, like Python and R.
4. Tutorial design.

Installation of Weka

1. **Download Weka: https://waikato.github.io/weka-wiki/downloading_weka/**

2. **The link to the downloading webpage is provided in MyCourse right below this tutorial video.**

3. **A possible bug of Weka for Mac users.**

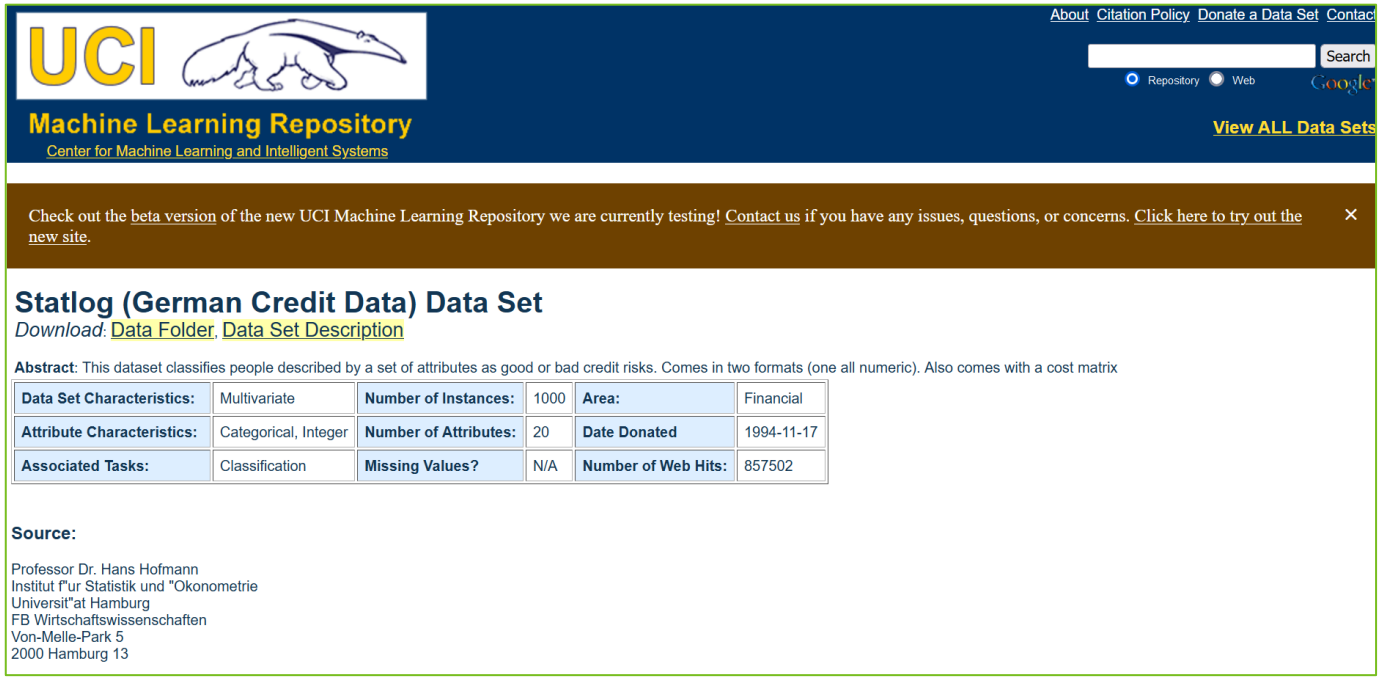


Importing data to Weka

1. Importing a CSV data file to Weka.
2. Importing data via url
3. Converting a CSV data file to an Arff data file.
4. Arff data file

Dataset to be imported

This dataset classifies people described by a set of attributes as good or bad credit risks.



UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

Repository Web

View ALL Data Sets

Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site.

Statlog (German Credit Data) Data Set

Download: Data Folder, Data Set Description

Abstract: This dataset classifies people described by a set of attributes as good or bad credit risks. Comes in two formats (one all numeric). Also comes with a cost matrix

Data Set Characteristics:	Multivariate	**Number of Instances:**	1000	**Area:**	Financial
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20	**Date Donated**	1994-11-17
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	857502

Source:

Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

https://github.com/selva86/datasets/blob/master/GermanCredit.csv

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

selva86 / datasets Public

Watch 25 Fork 1.7k Star 487

Code Issues Pull requests 3 Actions Projects Wiki Security Insights

master datasets / GermanCredit.csv Go to file ...

selva86 adding binary datasets Latest commit e112d95 on Dec 3, 2015 History

1 contributor

1001 lines (1001 sloc) 245 KB Raw Blame

Search this file...

	status	duration	credit_history	purpose	amount	savings	employment_duration	installment_rate	personal_status_
2	... < 100 DM	6	critical account/other credits existing	domestic appliances	1169	unknown/no savings account	... >= 7 years	4	male : single
3	0 <= ... < 200 DM	48	existing credits paid back duly till now	domestic appliances	5951	... < 100 DM	1 <= ... < 4 years	2	female : divorce
4	no checking account	12	critical account/other credits existing	retraining	2096	... < 100 DM	4 <= ... < 7 years	2	male : single
5	... < 100 DM	42	existing credits paid back duly till now	radio/television	7882	... < 100 DM	4 <= ... < 7 years	2	male : single
6	... < 100 DM	24	delay in paying off in the past	car (new)	4870	... < 100 DM	1 <= ... < 4 years	3	male : single
7	no checking account	36	existing credits paid back duly till now	retraining	9055	unknown/no savings account	1 <= ... < 4 years	2	male : single
8	no checking account	24	existing credits paid back duly till now	radio/television	2835	500 <= ... < 1000 DM	... >= 7 years	3	male : single
9	0 <= ... < 200 DM	36	existing credits paid back duly till now	car (used)	6948	... < 100 DM	1 <= ... < 4 years	2	male : single
10	no checking account	12	existing credits paid back duly till now	domestic appliances	3059	... >= 1000 DM	4 <= ... < 7 years	2	male : divorced/

What is Arff?

ARFF stands for Attribute-Relation File Format. It is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

https://datahub.io/blog/attribute-relation-file-format-arff

Outline of Tutorial video

1. **Understand data structure for analysis in Weka**
2. **How to specify the dependent variable (class) of dataset**
3. **How to convert a numeric variable to a nominal variable.**
   - We learn how to use the filter function
4. **Data visualization**

Data structure in Weka

1. **The last attribute in the attribute list represents the dependent variable.**
2. **All other variables will be automatically selected/included in machine learning as independent variables.**
3. **Often, we need to specify the last attribute (dependent variable) as a nominal variable**

Summary

1. Understand data structure for analysis in Weka
2. Specifying (dependent variable) class of the variable
3. Data visualization

Outline of Tutorial video

1. **It is good to keep the instances with missing values in Weka**
2. **Dealing with missing values**
   - Mean-replacement (filter -> ReplaceMissingValue)
   - Drop the instances with missing values(filter -> ReplaceMissingValue)

Outline of Tutorial video

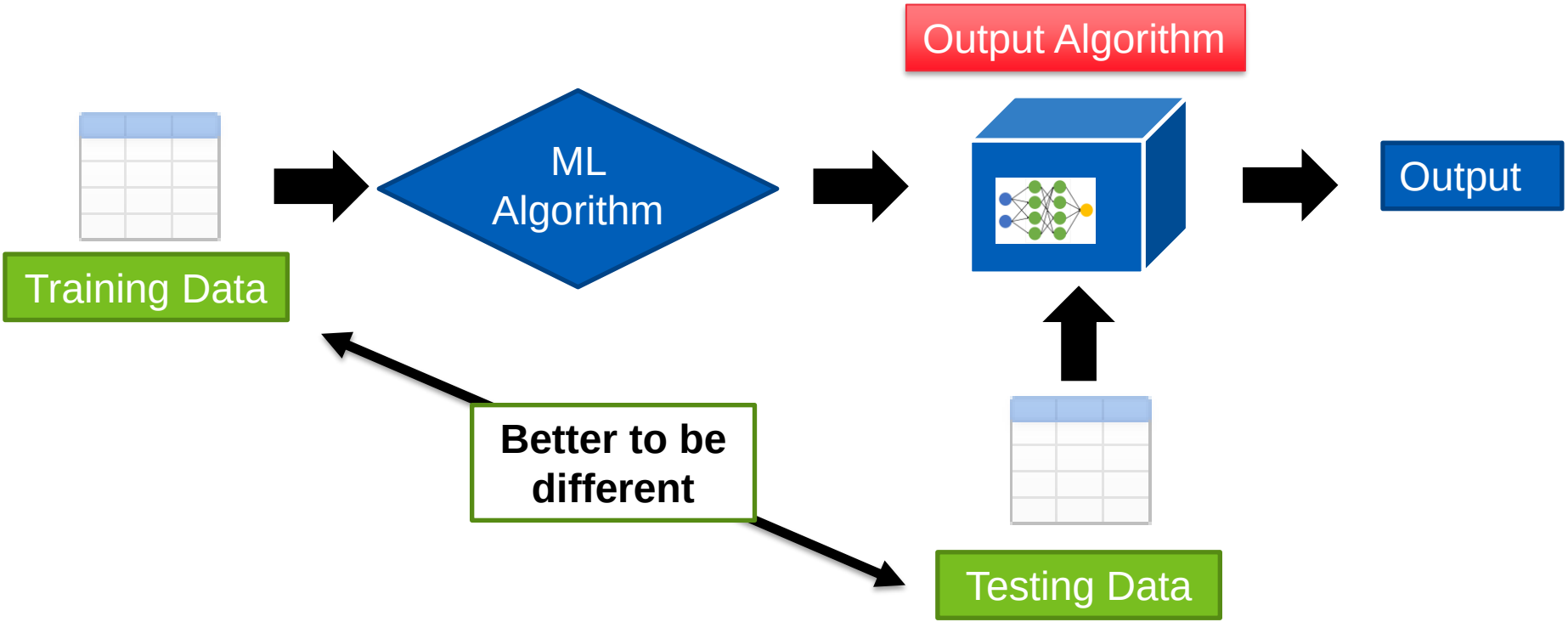
1. Basics of machine learning
2. Basics of decision tree

What is Machine Learning?

Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.

https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/

Basics of machine learning



Machine learning as an experimental science

Posted by u/apple_tau 5 months ago

154

[D] Why is ML research so experimental?

Discussion

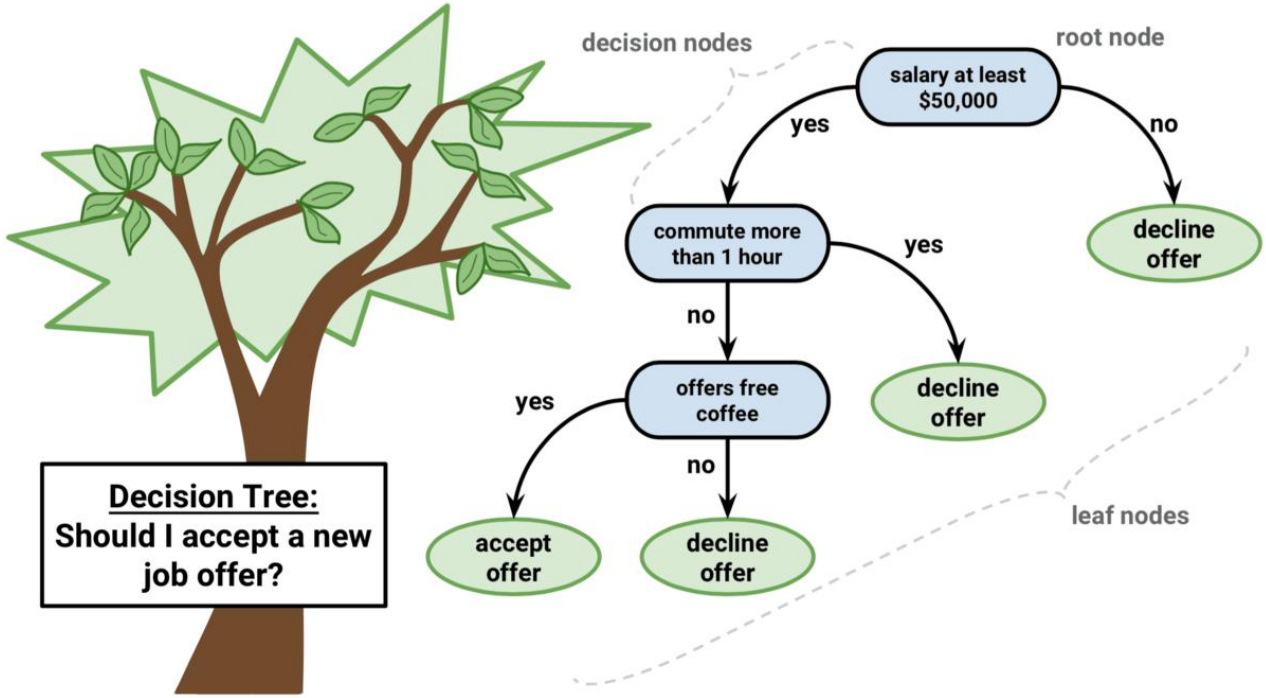
I'm still a bit of an ML noob, so this might be my inexperience talking, but why is so much research in ML experimental? My understanding is that areas such as physics have a strong experimental branch because they study already existing systems, but this doesn't seem to be the case with ML. I mean, we study mathematical objects, so it seems to me that we should be trying to understand them as such.

Like, if someone wants to propose a shortest path algorithm, they report its time complexity, not that it took 1min on average to run it, right?

131 Comments Share Save Hide Report 86% Upvoted

https://www.reddit.com/r/MachineLearning/comments/wgbmsr/d_why_is_ml_research_so_experimental/

Decision tree as a machine learning algorithm

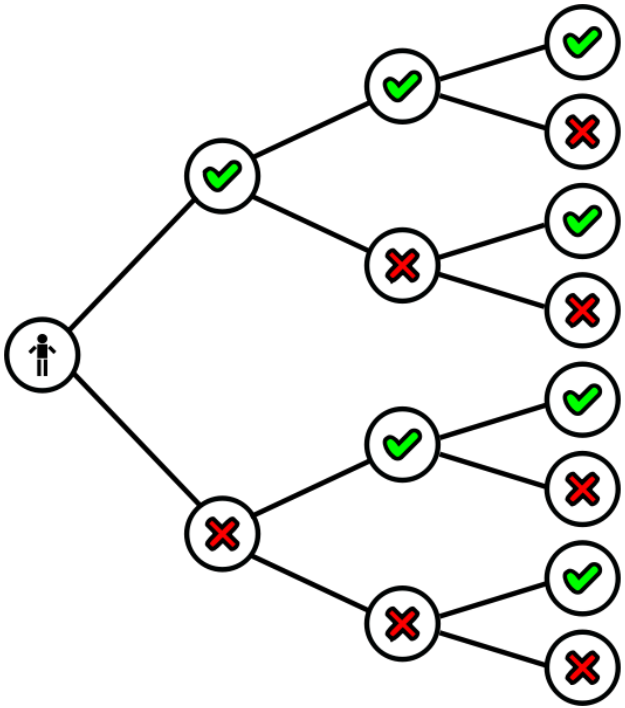


- **Nodes**
- **Leaf nodes**

https://regenerativetoday.com/simple-explanation-on-how-decision-tree-algorithm-makes-decisions/

Pruning

- **Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances.**
- **Pruning reduces the complexity of the final classifier and hence improves predictive accuracy by the reduction of overfitting.**



https://freesvg.org/mcol-event-tree

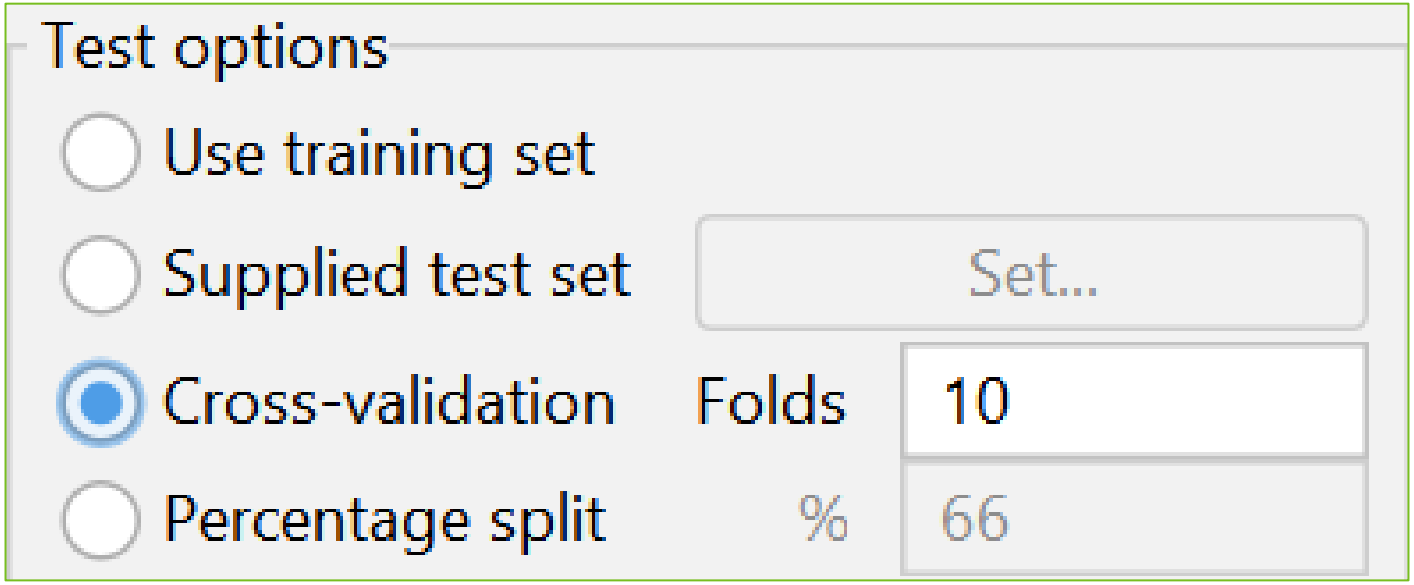
Summary

1. Basics of machine learning
2. Basics of decision tree

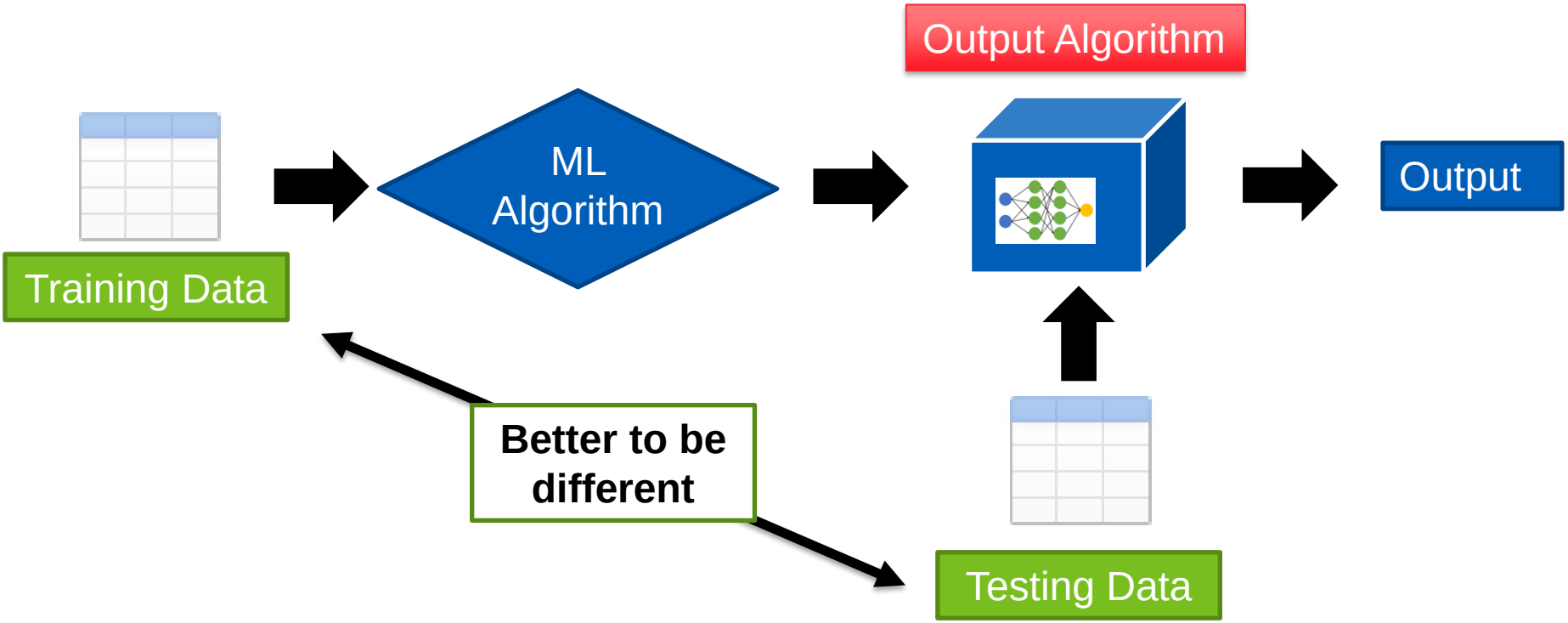
Outline of Tutorial video

1. Configuration of decision tree
2. Visualization of decision tree

Explaining Test Options in Weka

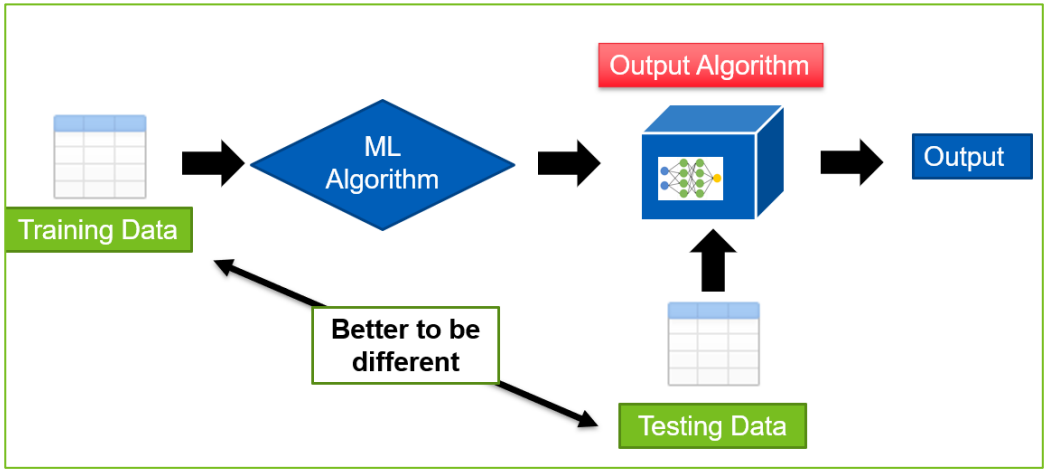


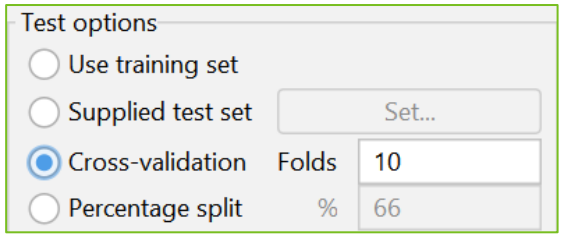
Basics of machine learning



Use training set: Classifies your model based on the dataset which you originally trained your model with.

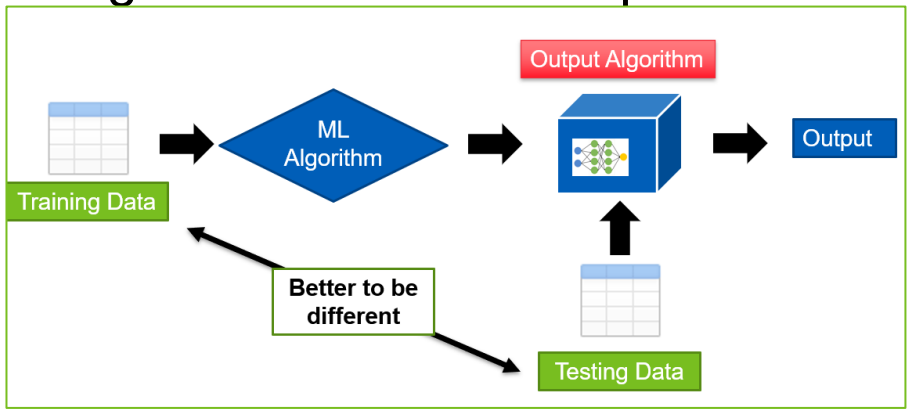
Percentage split: Divide your dataset into train and test according to the number you enter. By default, the percentage value is 66%, it means 66% of your dataset will be used as training set and the other 34% will be your test set.

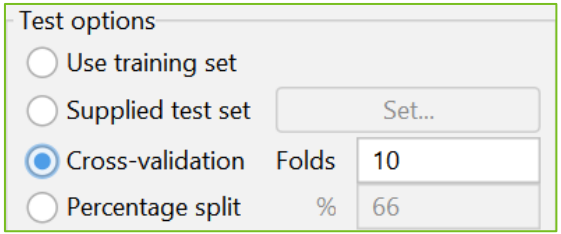




https://code.likeagirl.io/beginning-to-weka-step-by-step-93f6564d9f2

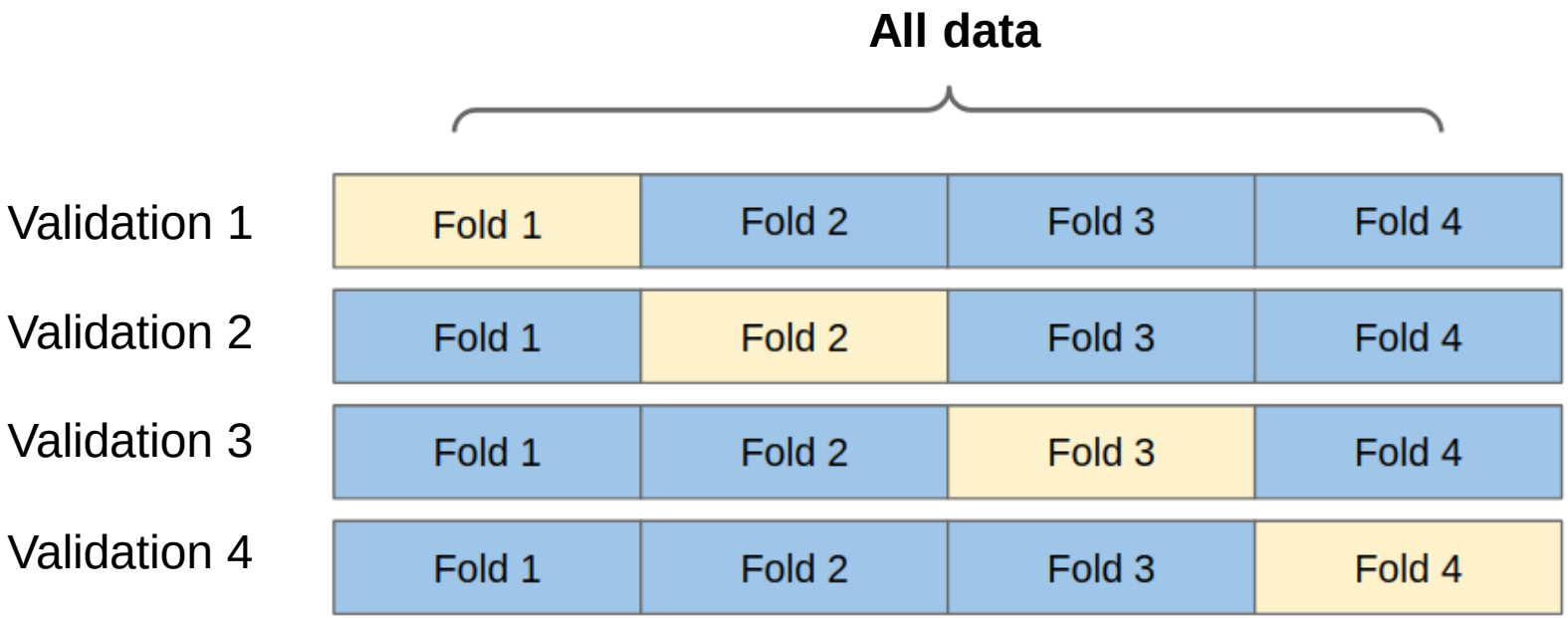
Cross-validation: The cross-validation option is a widely used one, especially if you have a limited amount of datasets. The number you enter in the *Fold* section are used to divide your dataset into Fold numbers (let's say it is **ten**). The original dataset is randomly partitioned into ten subsets. After that, Weka uses set **one** for testing and **nine** sets for training for the first training, then uses set **two** for testing and the other **nine** sets for training, and repeats that **ten** times in total by incrementing the set number each time. In the end, the average success rate is reported to the user.





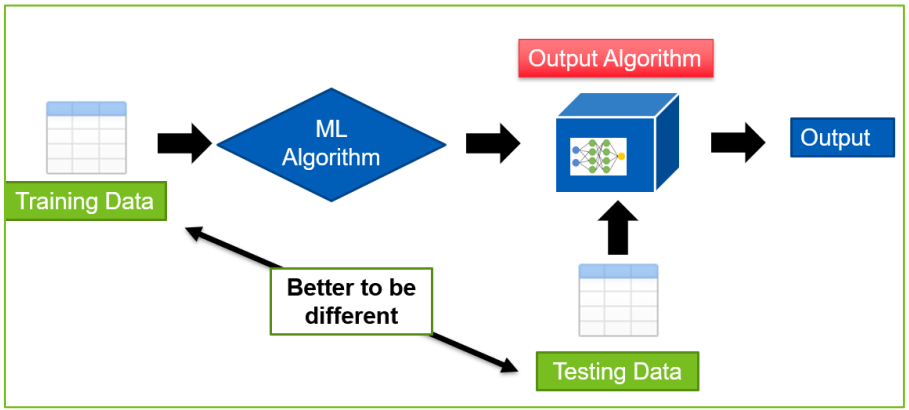
https://code.likeagirl.io/beginning-to-weka-step-by-step-93f6564d9f2

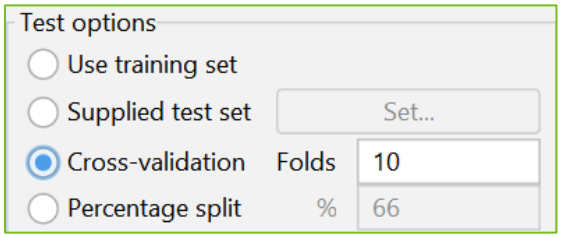
An example 4-Fold cross validation



https://commons.wikimedia.org/wiki/File:K-Fold_Cross-Validation.png

Supplied test set: Controls how your model is classified based on the dataset you supply from externally. Select a dataset file by clicking the Set button.





https://code.likeagirl.io/beginning-to-weka-step-by-step-93f6564d9f2

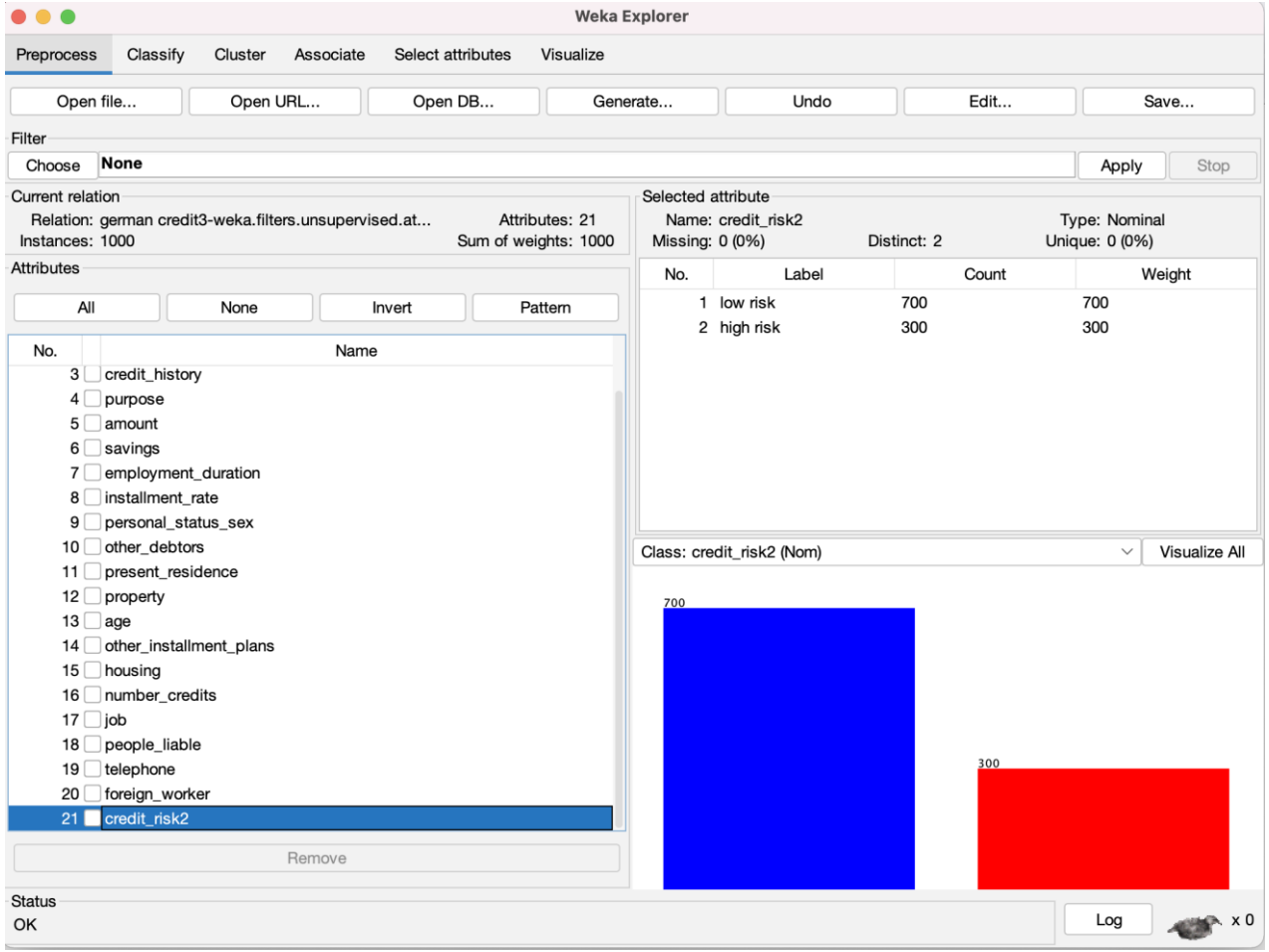
ZeroR Classifier

- **ZeroR is the simplest classification method that relies on the target and ignores all predictors.**
- **ZeroR classifier simply predicts the majority category (class).**
- **ZeroR is useful for determining a baseline performance as a benchmark for other classification methods.**

https://www.saedsayad.com/zeror.htm

ZeroR classifier as a benchmark

Benchmark= 700/1000 = 0.7



Outline

1. Briefly explain the cost-sensitive classification
2. How to generate cost-sensitive output in Weka
3. How to conduct cost-sensitive classification in Weka

Cost-sensitive classification (CSS)

- **Cost-sensitive classification: an approach to force machine learning algorithms to consider the costs caused by different kinds of errors. The costs of different kinds of errors are not assumed to be equal, and the objective of CSS is to minimize the expected costs.**
- **Basis: wrongly classifying an instance to a category incurs different costs.**

COST	Actually, give a loan	Actually, not give a loan
Should give a loan	0	1
Should not give a loan	5	0

https://mlr-org.github.io/mlr-tutorial/release/html/cost_sensitive_classif/index.html

Result 1

a	b	<-- classified as
600	50	a = Good
100	250	b = Bad

Correctly Classified Instances
N = 850
Ratio = 85 %

Cost = 50 X 1 + 100 X 5
= 550

Result 2

a	b	<-- classified as
600	120	a = Good
50	230	b = Bad

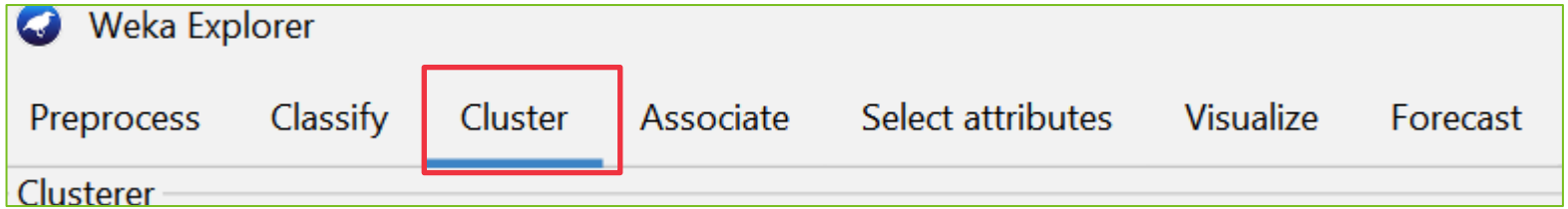
Correctly Classified Instances
N = 830
Ratio = 83 %

Cost = 120 X 1 + 50 X 5
= 370

A need for cost-sensitive classifier!

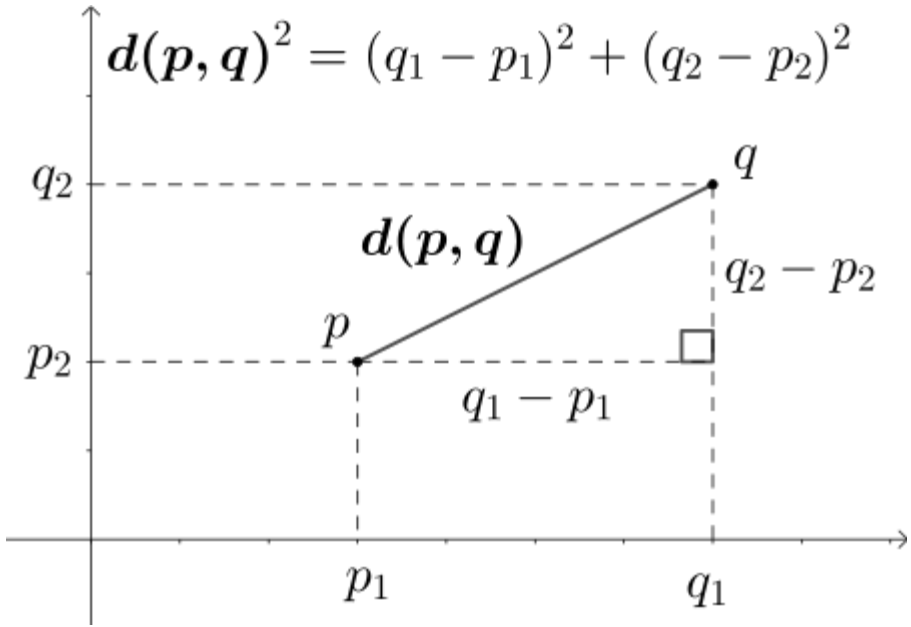
Outline

- **Euclidean Distance**
- **Cluster analysis**
- **Customer segmentation**
- **How to implement cluster analysis in Weka**



Euclidean Distance

ID	p	q	p2
1	3	5	4
2	4	4	5
3	2	5	4



https://www.cuemath.com/euclidean-distance-formula/

Higher dimensions

In three dimensions, the distance is

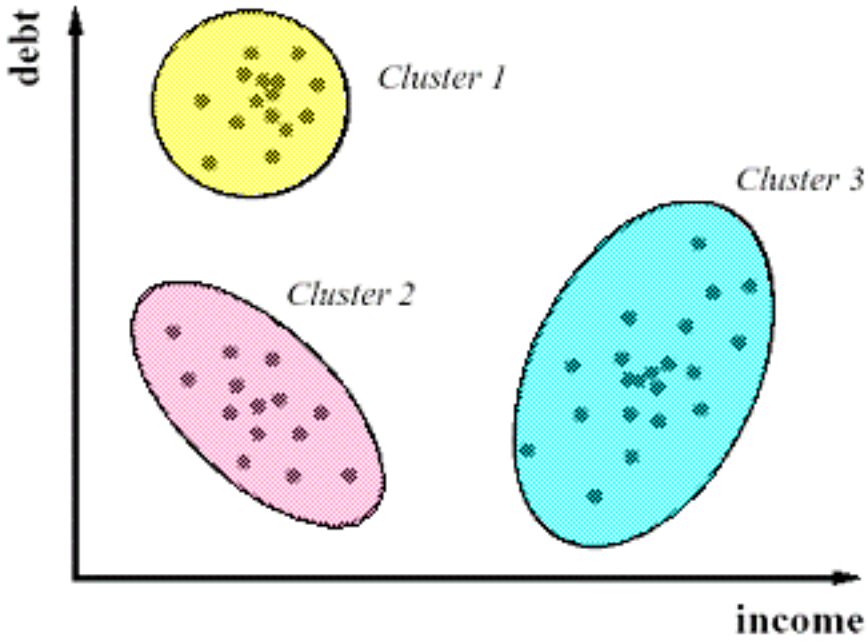
$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}.$$

In *N* dimensions, the distance is

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

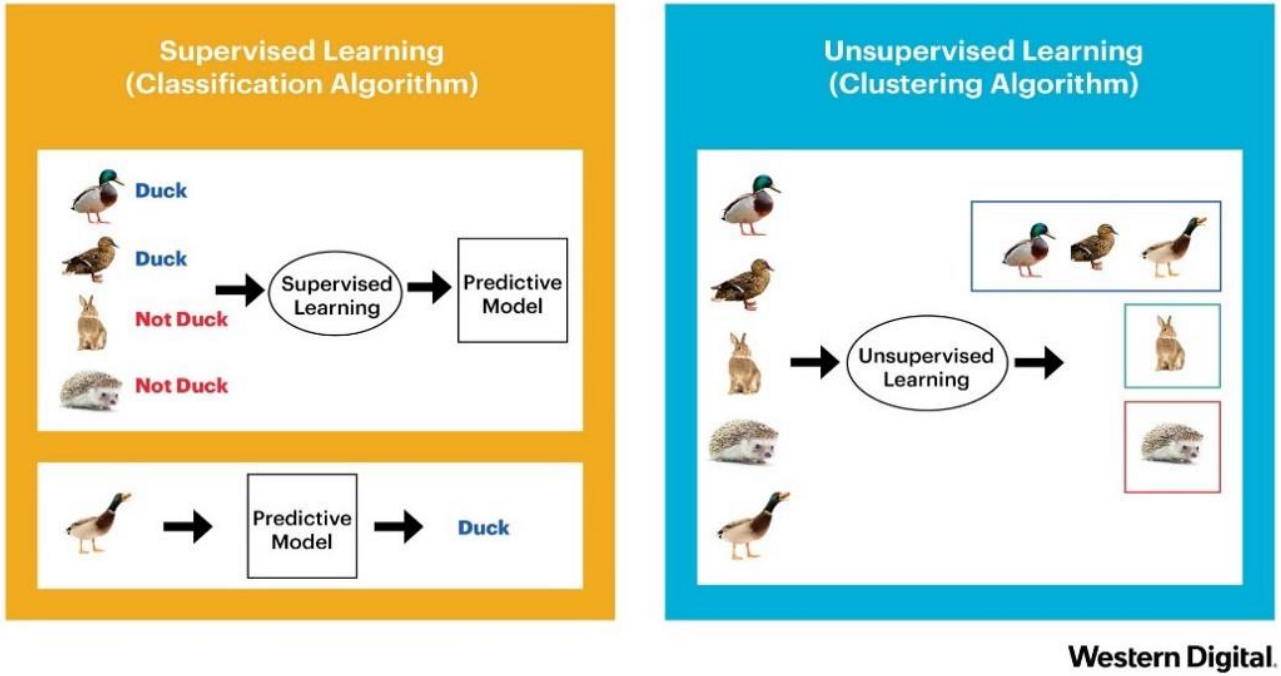
https://en.wikipedia.org/wiki/Euclidean_distance

Cluster



https://www.analyticsvidhya.com/blog/2013/11/getting-clustering-right/

Supervised vs Unsupervised Learning



https://medium.com/hengky-sanjaya-blog/supervised-vs-unsupervised-learning-aae0eb8c4878

Customer segmentation

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

https://www.optimove.com/resources/learning-center/customer-segmentation
https://www.qualitygurus.com/types-of-customers-and-customer-segmentation/

Outline

- **Imbalanced data in machine learning and its outcome**
- **SMOTE (Synthetic Minority Oversampling Technique) method**
- **How to install a package in Weka**

https://www.saedsayad.com/zeror.htm

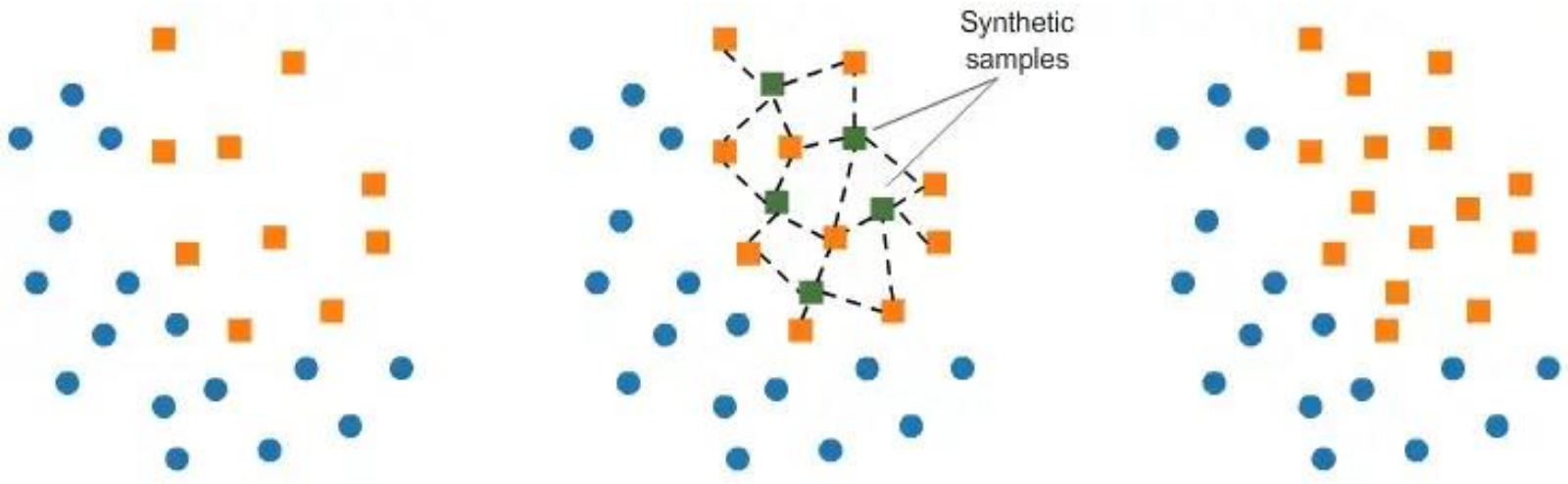
Imbalanced data

Imbalanced data is data in which observed frequencies are very different across the different possible values of a categorical variable. Basically, there are many observations of some type and very few of another type.

Using imbalanced data, we may have a model that appears very accurate for predicting the category that is over-presented but is useless for predicting the category that is under-presented.

https://towardsdatascience.com/smote-fdce2f605729

SMOTE
(Synthetic Minority Oversampling Technique)



SMOTE finds out ‘k’ nearest neighbors of a data point in the minority class. After the nearest data points have been identified, SMOTE then creates some synthetic data points on the lines joining the primary point and the neighbors so that these data points share the similar features/characteristics of the other minority data points.

Outline

1. Briefly explain multiple linear regression.
2. How to implement multiple linear regression in Weka.
3. Export regression result as a new variable.
4. Show the predicted values in the result.
5. Export the predicted values in a new data file.

Multiple Linear Regression

Multiple linear regression formula

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- ... = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

https://www.scribbr.com/statistics/multiple-linear-regression/

Outline

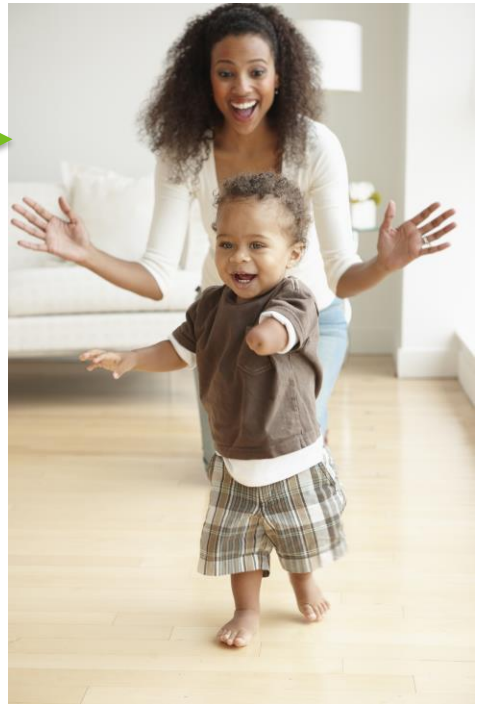
1. Briefly explain neural network algorithm.
2. How to implement a neural network in Weka.

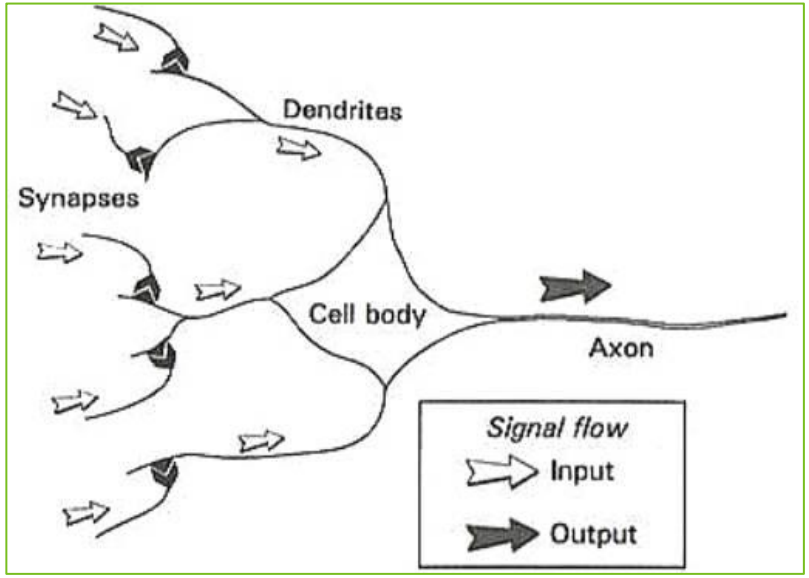
What is a neural network?

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.

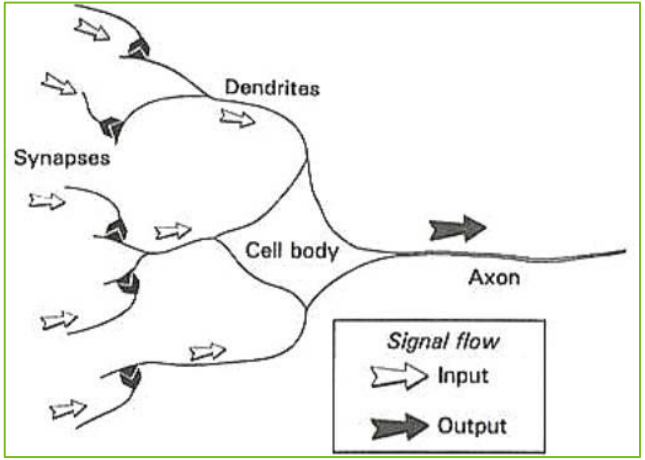
https://www.investopedia.com/terms/n/neuralnetwork.asp
https://en.wikipedia.org/wiki/Artificial_neural_network

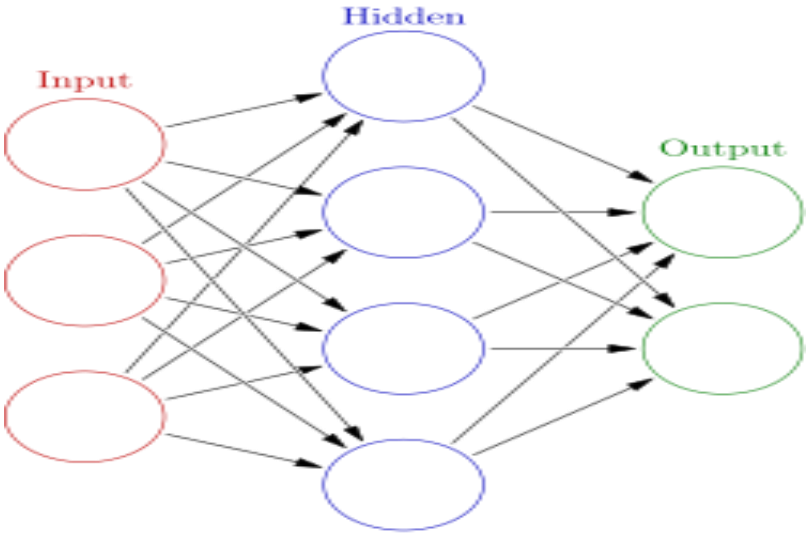
How do we learn to control our body?

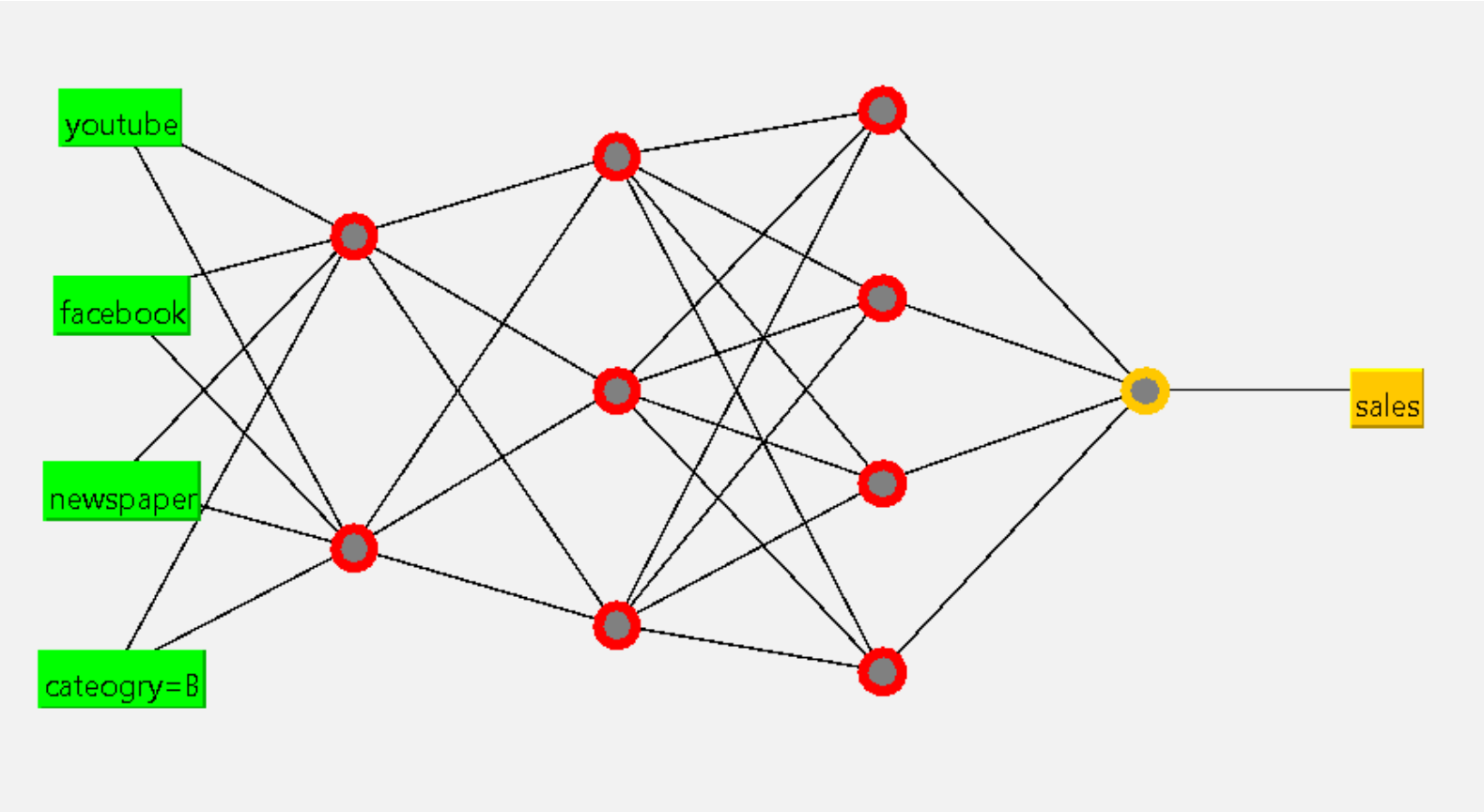




How does a neural network work?






youtube
facebook
newspaper
cateogry=B
sales

Outline

1. Briefly explain random forest.
2. How to implement random forest in Weka.

Why use random forest? A reflection of the decision tree algorithm

1. **Decision tree is very sensitive to small variations in the training data, which makes the model very unstable.**
2. **Very likely to overfit the training data**

- "**Variance** is an error resulting from sensitivity to small fluctuations in the dataset used for training. High variance will cause an algorithm to model irrelevant data, or noise, in the dataset instead of the intended outputs, called signal. This problem is called **overfitting**. An overfitted model will perform well in training, but won't be able to distinguish the noise from the signal in an actual test."

https://towardsdatascience.com/decision-trees-6-important-things-to-always-remember-85636858da51
https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/

Random Forest

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a "forest."

Steps of Random Forest Algorithm:
Step 1: Bootstrapping
Step 2: Feature selection
Step 3: Construction of trees
Step 4: Voting and aggregation

https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/
https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm

ID	x1	x2	x3	class
1	3	3	5	1
2	4	6	3	0
3	0	6	0	1
4	1	5	6	1
5	3	2	9	0

Sample 1
ID: 1, 2, 4, 3, 3

Sample 2
ID: 3, 2, 1, 5, 5

Sample 3
ID: 4, 2, 1, 3, 5

Sample 4
ID: 1, 2, 1, 3, 5

Step 1

Sample 1
Features: x1, x2

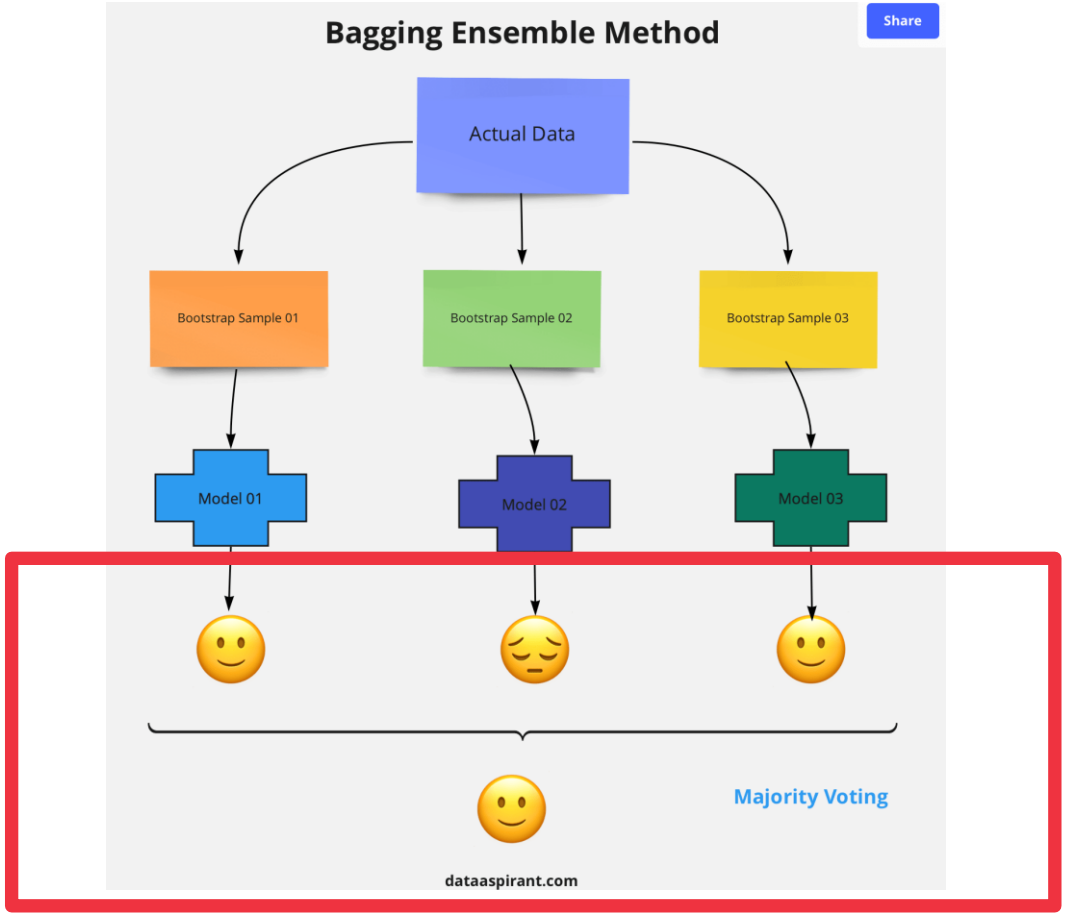
Sample 2
Features: x3, x2

Sample 3
Features: x1, x3

Sample 4
Features: x1, x2

Step 2

Step 4: Voting and aggregation



Outline

- **Briefly explain text mining in Weka**
- **How to implement text mining in Weka.**

https://www.saedsayad.com/zeror.htm

Example: text to attributes

Reviews	Rating
This is an **excellent** hotel!	5
Excellent hotel, highly recommended.	4
Many hotels nearby, and this one is not good!	3

Parameters to be considered:
1. **Stop words**
2. **Upper/lower case**
3. **Minimum frequency of words**
4. **Stemmer**

…

Convert to be…

	This	is	an	excellent	hotel	recommend	…
This is an excellent hotel!	1	1	1	1	1	0	…
Excellent hotel, highly recommended.	0	0	0	1	1	1	…

Outline

- **Briefly explain association analysis**
- **How to implement association analysis in Weka**

Association Analysis

- **Association analysis (AA)** discovers the probability of the co-occurrence of items in a collection.
- **Association rules**: the relationships between co-occurring items.

Market-basket analysis

Valuable for direct marketing, sales promotions, and for discovering business trends. Market-basket analysis can also be used effectively for store layout, catalog design, and cross-sell.

Example: An association model might find that a user who bought products A and B is 70% likely to buy product C in the same session.

https://docs.oracle.com/cd/E18283_01/datamine.112/e16808/market_basket.htm

Market Basket Example

Example II

- Where should detergents be placed in the Store to maximize their sales?
- Are window cleaning products purchased when detergents and orange juice are bought together?
- Is soda typically purchased with bananas? Does the brand of soda make a difference?
- How are the demographics of the neighborhood affecting what customers are buying?

Image source: deepclimate.org

https://www.analyticsvidhya.com/blog/2014/08/effective-cross-selling-market-basket-analysis/

Association rules

$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

An example of Association Rules

1. Assume there are 100 customers.
2. 10 of them bought milk, 8 bought butter and 6 bought both of them.
3. bought milk => bought butter.
4. support = P(Milk & Butter) = 6/100 = 0.06.
5. confidence = support/P(Butter) = 0.06/0.08 = 0.75.
6. lift = confidence/P(Milk) = 0.75/0.10 = 7.5.

Please note the rule A $\Rightarrow$ D differs from the rule D $\Rightarrow$ A

https://ghumare64.medium.com/market-basket-analysis-using-association-rule-mining-2f87e4064ee7
https://www.saedsayad.com/association_rules.htm

Please pay attention to the data format requirement

Order	Product
1	Product 1
1	Product 2
1	Product 3
2	Product 2
2	Product 3
3	Product 2
3	Product 3
3	Product 4

Unacceptable

Product 1	Product 2	Product 3	…	Product n
1	1		..	1
	1		..	
	1	1	..	
		1	..	1
	1	1	..	1
			..	1

Acceptable

ZeroR Classifier

- **ZeroR is the simplest classification method that relies on the target and ignores all predictors.**
- **ZeroR classifier simply predicts the majority category (class).**
- **ZeroR is useful for determining a baseline performance as a benchmark for other classification methods.**

https://www.saedsayad.com/zeror.htm

ZeroR classifier as a benchmark

Benchmark= 700/1000 = 0.7