Universidade Federal do Paraná (UFPR) Especialização em Engenharia Industrial 4.0

Introdução ao Weka

Data Mining with Open Source Machine Learning Software in Java

> David Menotti www.inf.ufpr.br/menotti/am-201

Hoje

- Weka
 - Introdução
 - Como instalar
 - Datasets
 - Usando algoritmos de:
 - Classificação
 - Clustering
 - Regressão

Introdução

- Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Ele contém ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização.
- Encontrada apenas nas ilhas da Nova Zelândia, a Weka é uma ave que não voa e tem uma natureza inquisitiva. O nome é pronunciado <u>assim</u>, e o pássaro soa <u>assim</u>.

Introdução

- Weka é um software de código aberto emitido sob a GNU General Public License.
- Sim, é possível aplicar a Weka para processar big data e realizar aprendizado profundo (deep learning)!

Introdução

 No <u>site</u>, existem vários cursos on-line gratuitos que ensinam aprendizado de máquina e mineração de dados usando o Weka. Confira no <u>site</u> os cursos para detalhes sobre quando e como se inscrever. Os vídeos dos cursos estão disponíveis no Youtube.



Free online courses on data mining with machine learning techniques in Weka

To help you explore the Weka software and learn about machine learning techniques for data mining and how to apply them, we have put together a series of three online courses that come with videos and plenty of exercises! They are hosted on the **FutureLearn** platform and are free of charge, but you can upgrade to receive an official FutureLearn Certificate of Achievement to use when applying for jobs or courses.

Data Mining with Weka

Everybody talks about Data Mining and Big Data nowadays. Weka is a powerful, yet easy to use tool for machine learning and data mining. **Data Mining with Weka** introduces you to practical data mining.

Manual

- Weka Manual
 - (v3-6-8) 03/05/2012
 - <u>http://www.nilc.icmc.usp.br/elc-ebralc2012/minicursos/WekaManual-3-6-8.pdf</u>
 - (v3-7-8) 21/01/2013
 - <u>http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf</u>

Como Instalar

• Weka website (latest version 3.8/3.9)

<u>https://www.cs.waikato.ac.nz/ml/weka/</u>

https://www.cs.waikato.ac.nz/ml/weka/downloading.html



Downloading and installing Weka

There are two versions of Weka: Weka 3.8 is the latest stable version, and Weka 3.9 is the development version. For the bleeding edge, it is also possible to download nightly snapshots.

Stable versions receive only bug fixes, while the development version receives new features. Weka 3.8 and 3.9 feature a package management system that makes it easy for the Weka community to add new functionality to Weka. The package management system requires an internet connection in order to download and install packages.

Como Instalar



Basta baixar / executar Weka-3-8-2jre-x64.exe ou Basta baixar / executar Weka-3-8-2-x64.exe

Como Instalar

https://www.cs.waikato.ac.nz/ml/weka/downloading.html

• Mac OS X

Click **here** to download a disk image for OS X that contains a Mac application including Oracle's Java 1.8 JVM (weka-3-8-2-oracle-jvm.dmg; 124.2 MB)

Other platforms (Linux, etc.)

Click **here** to download a zip archive containing Weka (wek 8-2.zip; 51.2 MB)

unzip the zip file. This will create a new directory called weka-3-8-2. To run Weka, change into that directory

java -jar weka.jar

Note that Java needs to be installed on your system for this to work. Also note, that using -jar will override your current CLASSPATH variable and only use the weka.jar.



CLI vs GUI

Início



CLI vs GUI

😣 🖨 🗊 SimpleCLI

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of the window. Use the up and down arrows to move through previous commands. Command completion for classnames and files is initiated with <Tab>. In order to distinguish between files and classnames, file names must be either absolute or start with './' or '~/' (the latter is a shortcut for the home directory). <Alt+BackSpace> is used for deleting the text in the commandline in chunks.

> help

```
Command must be one of:

java <classname> <args> [ > file]

kill

capabilities <classname> <args>

cls

history

exit

help <command>
```

😣 🖨 🗊 🛛 Weka Explorer	
Preprocess Classify Cluster As	sociate Select attributes Visualize
Ope Ope Ope Ge Filter	en Undo Edit Sav
Choose None	Apply Stop
Current relation	Selected attribute
Relation: Attributes: 5 Instances: Sum of weights: 150	Name: sepallen Type: Missi Distinct: Unique:
Attributes	Statistic Value
All I P	Minimum 4.3 Maximum 7.9 Mean 5.843 StdDay 0.828
No. Name	
1 sepallength 2 sepalwidth 3 petallength 4 petalwidth 5 class	Class: class (Nom) Visualize All
Remove	
Status	
ок	Log 🛷 X

Atributos

- Nominal: um de uma lista predefinida de valores – e.g. vermelho, azul, amarelo
- Numérico: Um número real ou inteiro
- String: delimitada por "aspas duplas"
- Data
- Relational

Arquivos ARFF

- A representação das instâncias
- Consiste em:
 - Um cabeçalho (header): Descreve os tipos de atributos e seus valores
 - Seção de dados: lista de dados separada por vírgula

Exemplo de Arquivo ARFF

% This is a toy example, the UCI weather dataset % Any relation to real weather is purely coincidental

@relation weather.symbolic



Nome do Dataset

Comment

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data

sunny, hot, high, FALSE, no sunny, hot, high, TRUE, no overcast, hot, high, FALSE, yes rainy, mild, high, FALSE, yes rainy, cool, normal, FALSE, yes rainy, cool, normal, TRUE, no overcast, cool, normal, TRUE, yes sunny, mild, high, FALSE, no sunny, cool, normal, FALSE, yes rainy, mild, normal, FALSE, yes sunny, mild, normal, TRUE, yes overcast, mild, high, TRUE, yes overcast, hot, normal, FALSE, yes rainy, mild, high, TRUE, no Atributos

Classe / Meta

Dados/Valores

ARFF



% This is a toy example, the UCI weather dataset % Any relation to real weather is purely coincidental

@relation weather.symbolic



Comment

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data

sunny, hot, high, FALSE, no sunny, hot, high, TRUE, no overcast, hot, high, FALSE, yes rainy, mild, high, FALSE, yes rainy, cool, normal, FALSE, yes rainy, cool, normal, TRUE, no overcast, cool, normal, TRUE, yes sunny, mild, high, FALSE, no sunny, cool, normal, FALSE, yes rainy, mild, normal, FALSE, yes sunny, mild, normal, TRUE, yes overcast, mild, high, TRUE, yes overcast, hot, normal, FALSE, yes rainy, mild, high, TRUE, yes Atributos

Classe / Meta

Dados/Valores

Abrindo um Dataset

Preprocess Clus	ter Associate	Select attributes V			\$ 7
Open file	pen URL	Open DB	Generate	Undo Edit	Save
Filter					
None					Apply
nt relation			Selected a	attribute	
Relation: None	😣 🗊 🛛 Open				Type: None
Instances: None	Look In:	09-weka	V		Unique: None
Attributes					
	uci-2007)111		Invoke options dialog	
	weather.	iominai.am		Note:	
				Some file formats offer additional	
				options which can be customized when invoking the options dialog.	
	File <u>N</u> ame:	weather.nominal.arf	f		
	Files of Type:	Arff data files (*.arff)		
					Visual
				Open Cancel]
	_		_		
	Remove				

Visualizando

🛇 🖨 💷 Weka Explorer					
Preprocess Classify Cluster Associate Select att	ributes Vis	sualize			
Open file Open UR Open DB Gene	erate	Undo	Edit	Save	
Choose None				Apply Stop	
Current relation	Selected	attribute			
Relation: weather.symbolic Attributes: 5 Instances: 14 Sum of weights: 14	Name Missing	: outlook : 0 (0%) D	istinct: 3	Type: Nominal Unique: 0 (0%)	
Attributes	No.	Label	Count	Weight	
All None Invert Pattern No. Name	Class: play	sunny overcast rainy ((Nom)	5 4 5	5.0 4.0 5.0 Visualize All	
Remove Status OK				Log ×0	

Visualizando



Excel => CSV

	A	В	С	D	E
1	outlook	temperature	humidity	windy	play
2	sunny	hot	high	FALSE	no
3	sunny	hot	high	TRUE	no
4	overcast	hot	high	FALSE	yes
5	rainy	mild	high	FALSE	yes
6	rainy	cool	normal	FALSE	yes
7	rainy	cool	normal	TRUE	no
8	overcast	cool	normal	TRUE	yes
9	sunny	mild	high	FALSE	no
10	sunny	cool	normal	FALSE	yes
11	rainy	mild	normal	FALSE	yes
12	sunny	mild	normal	TRUE	yes
13	overcast	mild	high	TRUE	yes
14	overcast	hot	normal	FALSE	yes
15	rainy	mild	high	TRUE	no
16	a state of the sta	and a second second	Second Contraction		

weather.csv

outlook,temperature,humidity,windy,play sunny,hot,high,FALSE,no sunny,hot,high,TRUE,no overcast,hot,high,FALSE,yes rainy,mild,high,FALSE,yes rainy,cool,normal,FALSE,yes rainy,cool,normal,TRUE,no overcast,cool,normal,TRUE,yes sunny,mild,high,FALSE,no sunny,cool,normal,FALSE,yes rainy,mild,normal,FALSE,yes sunny,mild,normal,TRUE,yes overcast,mild,high,TRUE,yes overcast,hot,normal,FALSE,yes rainy,mild,high,TRUE,yes

Excel => CSV



Software > Datasets

https://www.cs.waikato.ac.nz/ml/weka/datasets.html



Collections of Datasets

Some example datasets are included in the Weka distribution.

Available separately:

- A jarfile containing 37 classification problems, originally obtained from the UCI repository (datasets-UCI.jar, 1,190,961 Bytes).
- A jarfile containing 37 regression problems, obtained from various sources (datasets-numeric.jar, 169,344 Bytes).
- A jarfile containing 6 agricultural datasets obtained from agricultural researchers in New Zealand (agridatasets.jar, 31,200 Bytes).
- A jarfile containing 30 regression datasets collected by Luis Torgo (regression-datasets.jar, 10,090,266 Bytes).
- A gzip'ed tar containing UCI and UCI KDD datasets (uci-20070111.tar.gz, 17,952,832 Bytes)
- A gzip'ed tar containing StatLib datasets (statlib-20050214.tar.gz, 12,785,582 Bytes)
- A gzip'ed tar containing ordinal, real-world datasets donated by Dr. Arie Ben David (Holon Inst. of Technology/Israel) (datasets-arie_ben_david.tar.gz, 11,348 Bytes)
- A zip file containing 19 multi-class (1-of-n) text datasets donated by George Forman/Hewlett-Packard Labs (19MclassTextWc.zip, 14,084,828 Bytes)
- A bzip'ed tar file containing the Reuters21578 dataset split into separate files according to the ModApte split (reuters21578-ModApte.tar.bz2, 81,745,032 Bytes)
- A zip file containing 41 drug design datasets formed using the Adriana.Code software www.molecularnetworks.com/software/adrianacode - donated by Dr. M. Fatih Amasyali (Yildiz Technical Unversity) (Drugdatasets.zip, 11,376,153 Bytes)
- A zip file containing 80 artificial datasets generated from the Friedman function donated by Dr. M. Fatih Amasyali (Yildiz Technical Unversity) (Friedman-datasets.zip, 5,802,204 Bytes)

After expanding into a directory using your jar utility (or an archive program that handles tar-archives/zip files in case of the gzip'ed tars/zip files), these datasets may be used with Weka.

Other datasets in ARFF format:

- · Protein data sets, maintained by Shuiwang Ji, CS Department, Louisiana State University/USA
- Kent Ridge Biomedical Data Set Repository, maintained by Jinyan Li and Huiqing Liu, Institute for Infocomm Research, Singapore
- · Repository for Epitope Datasets (RED), maintained by Yasser El-Manzalawy, lowa State University.

WEKA Datasets

- Alguns datasets em formato ARFF <u>http://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html</u>
- <u>contact-lens.arff</u>
- <u>cpu.arff</u>
- <u>cpu.with-vendor.arff</u>
- <u>diabetes.arff</u>
- <u>glass.arff</u>
- <u>ionospehre.arff</u>
- <u>iris.arff</u>
- <u>labor.arff</u>

- <u>ReutersCorn-train.arff</u>
- <u>ReutersCorn-test.arff</u>
- <u>ReutersGrain-train.arff</u>
- <u>ReutersGrain-test.arff</u>
- <u>segment-challenge.arff</u>
- <u>segment-test.arff</u>
- soybean.arff
- <u>supermarket.arff</u>
- <u>vote.arff</u>
- <u>weather.arff</u>
- weather.nominal.arff

Classificação

- Como gerar:
 - uma árvore de decisão J48
 - um k-NN
 - Naive Bayes classifier
 - MLP
 - SVM
 - PCA

Classify > Choose

Classifier	Cluster	Associate	Select attributes	visualize	
Choose 148-0	12				
		Classi	*		
Use training set		Classi	ier output		
Supplied test se	t St				
Cross-validation	Folds 10				
O Percentage split	t % 66				
More op	tions				
-					
(Nom) play					
Start	Stop				
Result list (right-clic	ck for options)				
() · · · · · · · · · · · · · · · · · ·					

Classify > tree > J48

Preprocess Class	sify Cluster	Associate	Select attributes	Visualize	
Classifier Choose 148 -C	12				
		Classif			
Use training set		Classi			
Supplied test se	t s				
Cross-validation	Folds 10				
Percentage split	% 66				
More on	tions				
(Nom) play		-			
Start	Stop				
Result list (right-clic	k for options)				
E Lu					

Classify > tree > J48

Chassa U/R C 0.25 M 2	
148 -C 0.25 -M 2	
Test options	Classifier output
Use training set	Time taken to build model: 0.01 seconds
○ Supplied test set Set	
O Cross-validation Folds 10	<pre> Evaluation on training set ===</pre>
O Percentage split % 66	
	Correctly Classified Instances 14 100 %
More options	Kappa statistic 1
	Mean absolute error 0
(Nom) play	Relative absolute error 0 %
Start Stop	Root relative squared error 0 %
Besult list (right-click for options)	Total Number of Instances 14
04:38:07 - trees.148	=== Detailed Accuracy By Class ===
	TR Data ER Data Presizion Decell E Massura P(
	weighted Avg. I U I I I
	=== Confusion Matrix ===
	a h co- classified as
	9 0 a = yes
	0 5 b = no

Classifier output

🛇 🖨 🚯 Weka Explorer		
Classifier	Associate Select attributes Visualize	
Chaose 149 C 0.25 M 2	4	
140 -0 0.23 44 2		
Test options	Classifier output	Local
Use training set	=== Run information ===	
Supplied test set Set	Scheme:weka classifiers trees 148 -C 0.25 -M 2	
Cross-validation Folds 10	Relation: weather.symbolic	
Percentage split % 66	Instances: 14	
	outlook	
More options	temperature	
	humidity	
(Nom) play	play	
Start Stop	Test mode:evaluate on training data	
Result list (right-click for options)	=== Classifier model (full training set) ===	
04:38:07 - trees.J48	.148 pruned tree	
	autlack cuppy	
	humidity = high: no (3.0)	
	humidity = normal: yes (2.0)	Ányoro / Pogras Coradas
	outlook = overcast: yes (4.0)	AIVOIE / Regias Gelauas
	windy = TRUE: no (2.0)	
	windy = FALSE: yes (3.0)	
	Number of Leaves : 5	
	Size of the tree . 9	•
Status		
ок	Log	×0

Visualize



Visualize



Código em Java

import java.awt.BorderLayout; import java.io.BufferedReader; import java.io.FileReader;

import weka.classifiers.*; import weka.classifiers.trees.J48; import weka.core.Instances; import weka.gui.treevisualizer.PlaceNode2; import weka.gui.treevisualizer.TreeVisualizer;

```
public class WekaJ48 {
public static void main(String args[]) throws Exception {
    // train classifier
    J48 cls = new J48();
    Instances data = new Instances(new BufferedReader(new FileReader("D:\\sample.arff")));
    data.setClassIndex(data.numAttributes() - 1);
    cls.buildClassifier(data);
```

Código em Java

```
// display classifier
 final javax.swing.JFrame if =
  new javax.swing.JFrame("Weka Classifier Tree Visualizer: J48");
 jf.setSize(500,400);
 jf.getContentPane().setLayout(new BorderLayout());
 TreeVisualizer tv = new TreeVisualizer(null,
   cls.graph(),
   new PlaceNode2());
 jf.getContentPane().add(tv, BorderLayout.CENTER);
 jf.addWindowListener(new java.awt.event.WindowAdapter() {
  public void windowClosing(java.awt.event.WindowEvent e) {
   if.dispose();
 });
 jf.setVisible(true);
 tv.fitToScreen();
}
```

}

Classify > Lazy > k-NN (IBk)

😣 🖨 🗇 🛛 Weka Explorer					
Preprocess Classify Cluster As	sociate Select attributes	s Visualize			
Classifier					
Choose IBk -K 1 -W 0 -A "weka.core.n	eighboursearch.LinearNNSear	rch -A \"weka.co	pre.EuclideanDist	ance -R first-	last\""
Test options	Classifier output				
Use training set	Time taken to build mode?	· O seconds			
○ Supplied test set Set		. 0 0000100			
O Cross-validation Folds 10	=== Evaluation on trainir	ng set ===			
O Percentage split % 66	Summary				
	Correctly Classified Inst	tances	14	100	%
More options	Kappa statistic	is callees	1	0	
(Nom) play	Mean absolute error Root mean squared error		0.0625		
	Relative absolute error		13.4615 %		
Start Stop	Root relative squared err	ror	13.0347 %		
Result list (right-click for options)			14		
04:38:07 - trees.J48	=== Detailed Accuracy By	Class ===			
04:51:55 - bayes.NaiveBayes	TP Rate	FP Rate Pre	cision Recall	F-Measure	R
05:09:31 - Iazy.IBK	1	0	1 1	1	
	1	0	1 1	1	=
	weighted Avg. I	U	1 1	1	
	=== Confusion Matrix ===				
	a h a classified as				
	901a = ves				
	05 b = no				
					-
Status					
ΟΚ				LUG A	

Classify > function > SVM

- Deve-se instalar o LibSVM
 - LIBSVM A Library for Support Vector Machines
 - <u>https://www.csie.ntu.edu.tw/~cjlin/libsvm/</u>

Classify > function > SMO

Choose SMO - C 1.0 - L 0.001 - P 1.0E-12 - N 0 - V - 1 - W 1 - K "weka.classifiers.functions.supportVector.PolyKernel - C 250 Test options Classifier output © Use training set Set O Supplied test set Set Cross-validation Folds 10 Percentage split % 66 More options Correctly Classified Instances 14 100 More options Correctly Classified Instances 14 100 More options Other absolute error 0.0245 Root mean squared error 0.0354 Result list (right-click for options) Foot mean squared error 7.3845 % Total Number of Instances 14 O4:38:07 - trees.J48 O4:51:55 - bayes.NaiveBayes Distlice Arg. 1 0 1 1 O5:11:06 - functions.MultilayerPerceptron 05:14:48 - functions.SMO 0 1 1 1 weighted Avg. 0 1 1 1 1 1 1 O5:10:10 Functions.SMO I arg.like as 9 0 1 1 1 1 O5:11:06 - functions.SMO I arg.like as 0 1 <	Classifier		10	
Test options Classifier output • Use training set Supplied test set Supplied test set Set Cross-validation Folds O Percentage split % 66 More options Correctly Classified Instances 14 More options More options 0.0245 (Nom) play Correctly Classified Instances 0 (Nom) play Mean absolute error 0.0245 Result list (right-click for options) Start Stappe statistic O4:38:07 - trees.J48 O4:38:07 - trees.J48 Total Number of Instances 14 e== Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure 1 O5:10:6.6 - functions.MultilayerPerceptron O5:11:6.6 - functions.SMO 1 1 1	Choose SMO -C 1.0 -L 0.001 -P 1.0E-12	-N 0 -V -1 -W 1 -K "weka.classifiers.fu	nctions.supportVector.	PolyKernel -C 2500
<pre>● Use training set Supplied test set Cross-validation Percentage split % 66 More options Nore options Start Stop Result list (right-click for options) 04:38:07 - trees.J48 04:51:55 - bayes.NaiveBayes 05:09:31 - lazy.IBk 05:11:06 - functions.SMO Start Stop Correctly Classified Instances 14 100 Orrectly Classified Instances 0 0 Nore options Correctly Classified Instances 14 100 Orrectly Classified Instances 0 0 Nore options Correctly Classified Instances 14 100 Incorrectly Classified Instances 0 0 Relative absolute error 0.0354 Root relative squared error 7.3845 % Total Number of Instances 14 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure 1 0 1 1 1 Weighted Avg. 1 0 1 1 1 === Confusion Matrix === a b < classified as 9 0 a = yes 0 5 b = no </pre>	Test options	Classifier output		
 Supplied test set Set Cross-validation Folds 10 Percentage split % 66 More options More options Correctly Classified Instances 14 100 Incorrectly Classified Instances 0 0 Incorrectly Classified Instances 0 0 Kapa statistic 1 Mean absolute error 0.0334 Relative absolute error 7.3845 % Total Number of Instances 14 TP Rate FP Rate Precision Recall F-Measure 1 0 Signed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure 1 0 I 0 1 1 1 Weighted Avg. 1 0 1 1 1 Weighted Avg. 1 0 1 1 1 The confusion Matrix === a b < classified as 9 0 a = yes 0 5 b = no 	Use training set	Time taken to build model: 0.0	8 seconds	4
<pre>Cross-validation Folds 10 Percentage split % 66 Correctly Classified Instances 14 100 More options (Nom) play ▼ Start Stop Start Stop 04:38:07 - trees.J48 04:51:55 - bayes.NaiveBayes 05:09:31 - lazy.IBk 05:11:06 - functions.SMO</pre> Crecetly Classified Instances 14 100 Incorrectly Classified Instances 0 0 Kappa statistic 1 Mean absolute error 0.0354 Relative absolute error 7.3845 % Total Number of Instances 14 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure 1 0 1 1 1 1 Weighted Avg. 1 0 1 1 1 === Confusion Matrix === a b < classified as 9 0 a = yes 0 5 b = no	Supplied test set Set			
<pre> Percentage split % 66 More options Correctly Classified Instances 14 100 Incorrectly Classified Instances 0 0 Incorrectly Classified Instances 0 Incorrectly Class 0 Incorrectly Class</pre>	O Cross-validation Folds 10	=== Evaluation on training set		
More optionsIncorrectly Classified Instances00(Nom) playIncorrectly Classified Instances00(Nom) playNot mean squared error0.0245(Nom absolute error0.0354Relative absolute error5.2713 %Result list (right-click for options)Relative absolute error7.3845 %04:38:07 - trees.J48TP Rate FP RatePrecision Recall F-Measure05:09:31 - lazy.IBk1105:11:06 - functions.MultilayerPerceptron1005:14:48 - functions.SMO11Weighted Avg.011=== Confusion Matrix ===a b < classified as	O Percentage split % 66	Correctly Classified Instances	14	100
Wolf mean squared error 0.0534 Start Stop Result list (right-click for options) Root relative squared error 04:38:07 - trees.J48 14 04:51:55 - bayes.NaiveBayes 14 05:09:31 - lazy.IBk TP Rate FP Rate Precision Recall F-Measure 1 0 1 1 05:11:06 - functions.MultilayerPerceptron 1 0 1 1 Weighted Avg. 0 1 1 1 === Confusion Matrix === a b < classified as 9 0 a = yes 0 5 b = no	More options	Incorrectly Classified Instand Kappa statistic Mean absolute error	es 0 1 0.0245	0
StartStopResult list (right-click for options)Root relative squared error7.3845 %04:38:07 - trees.J48Od:51:55 - bayes.NaiveBayesDetailed Accuracy By Class ===05:09:31 - lazy.IBkTP RateFP RatePrecisionRecall F-Measure1011105:11:06 - functions.MultilayerPerceptron101105:14:48 - functions.SMOI111a b< classified as9 0 a = yes0 5 b = no		Relative absolute error	5.2713 %	
Result list (right-click for options) Image: State option in the state option in	Start Stop	Root relative squared error	7.3845 % 14	
	Result list (right-click for options) 04:38:07 - trees.J48 04:51:55 - bayes.NaiveBayes 05:09:31 - lazy.IBk 05:11:06 - functions.MultilayerPerceptron 05:14:48 - functions.SMO	=== Detailed Accuracy By Class TP Rate FP Ra 1 0 1 0 Weighted Avg. 1 0 === Confusion Matrix === a b < classified as 9 0 a = yes 0 5 b = no	=== te Precision Rec 1 1 1 1 1 1	all F-Measure 1 1 1

Classify > function > MLP

Classifier Choose MultilayerPerceptron L 0.3 ·M 0.2 ·N 500 ·V 0 ·S 0 ·E 20 ·H a Test options © Use training set © Supplied test set Set © Cross-validation Folds 10 Percentage split % 66 (Nom) play © Start Stop Gesult list (right-click for options) O4:38:07 - trees.J48 O4:51:55 - bayes. NaiveBayes O5:09:31 - lazy.IBk O5:11:06 - functions. MultilayerPerceptron © S a = yes 05 b = n0 Classified output Time taken to build model: 0.08 seconds === Evaluation on training set === === Summary === Correctly Classified Instances 14 100 Incorrectly Classified Instances 0 0 Root reality exquered error 0.0235 Root reality exquered error 7.3845 % Total Number of Instances 14 === Detailed Accuracy By Class === 0 1 0 1 1 1 Weighted Avg. 1 0 1 1 1 === Confusion Matrix === a b < classified as 9 0 a = yes 0 5 b = n0	Preprocess Classify Cluster Associa	ate Select attributes Visualize
Choose MultilayerPerceptron -1.0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a Test options Classifier output • Use training set Set • Supplied test set Set • Cross-validation Folds 10 • Percentage split % 66 • More options Orrectly Classified Instances 14 • More options 0 and bsolute error 0.0245 • Nor options Noge statistic 1 • Main absolute error 0.0354 • Rot mean squared error 0.0354 • Rot relative squared error 7.3845 % • Total Number of Instances 14 • == Detailed Accuracy By Class === TP Rate Precision Recall F-Measure 1 0 1 1 1 05:11:06 - functions.MultilayerPerceptron Weighted Avg. 0 1 1 • < classified as	Classifier	
Test options Classifier output • Use training set Supplied test set Supplied test set Set Cross-validation Folds Percentage split % 66 More options % 66 More options Correctly Classified Instances 14 More options More options 0 More options Mean absolute error 0.0245 Robult error 0.0354 Result list (right-click for options) Detailed Accuracy By Class === Of:38:07 - trees.J48 Ot = 1 1 O4:38:07 - trees.J48 Ot = 1 1 O5:09:31 - lazy.IBk The Rate FP Rate Precision Recall F-Measure Do: 11:06 - functions.MultilayerPerceptron Detailed Accuracy By Class 1 weighted Avg. 1 1	Choose MultilayerPerceptron -L 0.3 -M	0.2 -N 500 -V 0 -S 0 -E 20 -H a
 Use training set Supplied test set Cross-validation Folds 10 Percentage split % 66 More options More options Correctly Classified Instances 14 100 Correctly Classified Instances 0 Correctly Classified Instances 0 Correctly Classified Instances 14 100 More options Kappa statistic 1 Mean absolute error 0.0245 Root mean squared error 7.3845 % Total Number of Instances 14 The taken to build model: 0.08 seconds Start Stop Correctly Classified Instances 14 Mean absolute error 7.3845 % Total Number of Instances 14 Start Stop Correctly Class Perceptron Weighted Avg. 1 O 1 I <l< th=""><th>Test options</th><th>Classifier output</th></l<>	Test options	Classifier output
Result list (right-click for options) 14 04:38:07 - trees.J48 12 04:51:55 - bayes.NaiveBayes 12 05:09:31 - lazy.IBk 1 05:11:06 - functions.MultilayerPerceptron 1 Weighted Avg. 1 0 1 1 0 1 0 1 1 1	 Use training set Supplied test set Cross-validation Percentage split More options (Nom) play	Time taken to build model: 0.08 seconds === Evaluation on training set === === Summary === Correctly Classified Instances 14 100 Incorrectly Classified Instances 0 0 Kappa statistic 1 Mean absolute error 0.0245 Root mean squared error 0.0354 Relative absolute error 5.2713 % Root relative squared error 7.3845 % Tatal Number of Instance
	Result list (right-click for options) 04:38:07 - trees.J48 04:51:55 - bayes.NaiveBayes 05:09:31 - lazy.IBk 05:11:06 - functions.MultilayerPerceptron	<pre>=== Detailed Accuracy By Class ===</pre>
	Status OK	Log x0

Select Attributes > PCA

😣 🕒 💿 Weka Explorer	
Preprocess Classify Cluster As	ociate Select attributes Visualize
Attribute Evaluator	
Choose PrincipalComponents -R 0.	95 -A 5
Search Method	
Choose Ranker -T -1.797693134862	3157E308 -N -1
Attribute Selection Mode	Attribute selection output
Use full training set Cross-validation Folds 10 Seed 1 (Nom) class	Correlation matrix 1 -0.11 0.87 0.82 -0.11 1 -0.42 -0.36 0.87 -0.42 1 0.96 0.82 -0.36 0.96 1
Start Stop Result list (right-click for options)	eigenvalue proportion cumulative 2.91082 0.7277 0.7277 0.581petallength+0.5 0.92122 0.23031 0.95801 -0.926sepalwidth-0.3
05:32:45 - Ranker + PrincipalComponer	Eigenvectors V1 V2 0.5224 -0.3723 sepallength -0.2634 -0.9256 sepalwidth 0.5813 -0.0211 petallength 0.5656 -0.0654 petalwidth Ranked attributes: 0.2723 1 0.581petallength+0.566petalwidth+0.522sepallength-0.263seg 0.042 2 -0.926sepalwidth-0.372sepallength-0.065petalwidth-0.021pe
■ III III III III III IIII IIII IIII I	Selected attributes: 1,2 : 2
Status OK	Log 💉 X

Visualizar



Clustering & Regressão

- Como gerar:
 - Um kMeans
 - Uma regressão linear

Cluster > SimpleKmeans

😣 🖨 🗊 🛛 Weka Explorer							
Preprocess Classify Cluster	Associate	Select attributes	Visualize				
Clusterer							
Choose SimpleKMeans -N 3 -A "w	veka.core.Eu	clideanDistance -R firs	t-last" -l 500 -9	S 10			
Cluster mode	Cl	usterer output					
Use training set	Nu W	umber of iterations: ithin cluster sum of	3 squared erro	ors: 7.817456892309	574		_
⊖ Supplied test set Set	M	issing values global	ly replaced v	with mean/mode			
O Percentage split %	66 C	luster centroids:					
Classes to clusters evaluation				Cluster#			
(Nom) class 💌	A	ttribute	Full Data (150)	a 0) (50)	1 (50)	2 (50)	
Store clusters for visualization	S	epallength	5,843	3 5.936	5.006	6.588	
	Se	epalwidth	3.054	4 2.77	3.418	2.974	
Ignore attributes		etalwidth	1,198	7 1.326	0.244	2.026	_
Start Stop	c	lass	Iris-setosa	a Iris-versicolor	Iris-setosa	Iris-virginica	
Result list (right-click for options)							
05:40:13 - SimpleKMeans							
	T:	ime taken to build m	o <mark>del (full t</mark> i	raining data) : 0.0	l seconds		
	=	== Model and evaluat	ion on train:	ing set ===			=
	C	lustered Instances					
	0	50 (22%)					
	1	50 (33%)					
	2	50 (33%)					
	•			11			
Status OK						Log	x 0

Cluster > SimpleKmeans



Classify > LinearRegression

Classifier	
Choose LinearRegression -S 0 -R	1.0E-8
Test options	Classifier output
Use training set	Attributes: 3 nassenger numbers
○ Supplied test set Set	Date
O Cross-validation Folds 10	NewData Test mode:evaluate on training data
O Percentage split % 66	
More options	=== Classifier model (full training set) ===
More options	
(Num) passenger numbers	Linear Regression Model
	passenger_numbers =
Start Stop	2.657 * NewData +
Result list (right-click for options)	90.3866
06:05:12 - functions.LinearRegression	Time taken to build model: 0.01 seconds
	=== Evaluation on training set === === Summary ===
	Mean absolute error 34.4219
	Root mean squared error 45.757
	Relative absolute error 34.2701 %
	Total Number of Instances 144

Classify > LinearRegression

X: NewData (Num) Colour: passenger_numbers (Num)				-	Y: predictedpassenger_numbers		
					Select Instance		
Reset	Clear	Open	Save		Jitter 🖵		
470.3167 280.368-	and the second se	and the second sec	A CONTRACT	R	x 2000/0000 2000/0000 x 2000/0000 x		
90.419	12	71.5	03		142.9938	-	



References

- Weka 3: Data Mining Software in Java
 - <u>https://www.cs.waikato.ac.nz/ml/weka/</u>
- Weka Datasets
 - <u>http://storm.cis.fordham.edu/~gweiss/data-mining/da</u> <u>tasets.html</u>