Laboratório Classificação com o WEKA Explorer

Para esse laboratório considere os seguintes classificadores:

- C4.5 (J4.8)
- KNN
- Naïve Bayes

Você pode realizar o tutorial básico a partir da página abaixo, ou as atividades descritas no restante desta página são avançadas.

Considere as bases de treinamento e teste de dígitos manuscrítos*

- digTrain1k.arff, digTrain2k.arff digTrain3k.arff digTrain4k.arff e digTrain5k.arff
- digTest5k.arff
- 1. Compare o desempenho desses classificadores em função da disponibilidade de base de treinamento. Alimente os classificadores com blocos de 1000 exemplos e plote num gráfico o desempenho na base de testes e analise em qual ponto o tamanho da base de treinamento deixa de ser relevante.
- 2. Qual é o classificador que tem o melhor desempenho com poucos dados (1000 exemplos)?
- 3. Qual é o classificador que tem melhor desempenho com todos os dados?
- 4. Qual é o classificador mais rápido para classificar os 5k exemplos de teste.
- 5. O que você pode dizer a respeito das matrizes de confusão. Os erros são os mesmos para todos os classificadores quando todos eles utilizam toda a base de teste?

*disponível em: <u>www.inf.ufpr.br/menotti/am-211/data.zip</u>

Nas páginas abaixo, encontram-se tutoriais explicando como usar classificadores no Weka.

Preparando os dados para classificação

Inicie uma sessão do Weka ou execute em linha de comando: java –jar weka.jar.
 Quando a GUI Chooser surgir, selecione o Explorer a partir das quatro opções do lado direito.

| | | Weka Explorer | | | | |
|--|---------------|---|---------------------------------|---|----------------|----------------------------|
| | | Preprocess Classify Outer Ass | ocure Select attributes Vousion | | | |
| | | Open Ne Open UP | it Open DB G | enerate | ER. | 1 1ave |
| | | Choose None | | | | (Apple) |
| | | Current relation Relation: None Instances: None | Altributes: None | Selected attribute Name: None Missing: None | Distinct: None | Type: None Unique: None |
| 🐼 Weka GUI Chooser | | Attributes All tame | Diviet. Patien. | | | |
| Program Visualization Tools Help | | | | | | |
| | Applications | | | | | |
| WEKA | Explorer | | | | | Visualce All |
| of Waikato | Experimenter | | | | | |
| Waikato Environment for Knowledge Analysis Version 3.6.11 | KnowledgeFlow | | | | | |
| (c) 1999 - 2014 | Cruck CIT | | Nettores | | | |
| Hamilton, New Zealand | Simple CLI | Status Welcome to the Welca Explorer | | | | Log 🛷 |

2. Estamos no **Preprocess** agora. Clique no botão **Open** para abrir a caixa de diálogo padrão através da qual você pode selecionar um arquivo. Escolha o arquivo **customer_lab2.csv**.

1. Elegendo o atributo meta ou classe

 Para realizar a classificação com Weka, o último atributo no conjunto de dados é considerado como classe / meta e deve ser nominal. Como o último atributo do dataset customer_lab2.csv é do tipo numérico (1/0), devemos convertê-lo para o tipo nominal próximo passo.

| 2 | w | leka | Explorer | - 🗆 × |
|--|---|------|--|---------------------------------|
| Preprocess | Classify Cluster Associate Select attributes Visualize | | | |
| Open fl | e Open URL Open DB | Gen | erate Undo Edit | Save |
| | | | | |
| direct | latered and the second s | | | Save the w |
| Choose | Pone | | | ADDA |
| Current relation: Relation: Instances: | ion customer_labThree 999 Attributes: 19 | (| Selected attribute Name: response_01 Missing: 0 (0%) Distinct: 2 | Type: Numeric Unique: 0 (0%) |
| Attributes | | | Statistic Value | / |
| | | | Minimum | |
| AI | None Invert Pattern | | Maximum 1 | |
| | 1 mars | - | Mean 0.418 | |
| NO , | Name | | StdDev 0.494 | |
| 7 | jobcat | ^ | | |
| 8 | union | | | |
| 9 | empcat | - 11 | | |
| 10 | card2tenurecat | -12 | Characterization of Altern) | 10 united Al |
| 11 | equip | - 11 | Class: response_01 (vum) | VISUAI/22 AV |
| 12 | internet | -11 | | |
| 14 | calid | - 11 | dB1 | |
| 15 | celweit | - 11 | | |
| 16 | forward | | | 418 |
| 17 | confer | | | |
| 18 | _ebi | | | |
| 19 | response_01 | ¥ | | |
| < | | _ | | |
| _ | Remove | | | 0 |
| | | | 0 0.5 | |
| Status | | | | |
| OK | | | | L00 400° |

 O filtro de atributo não supervisionado NumericToNominal é escolhido para executar esta conversão. Como gostaríamos de converter apenas o último atributo, altere o attributeIndices para last.

| 🖸 Weka E | xplorer | - • × |
|---|--|-----------------------------------|
| Preprocess Classify Cluster Associate Select attributes Visualize | | |
| Open file Open URL Open DB Gener | ate Undo | Edit Save |
| Current relation Relation: customer_labThree-weka.filters.unsupervised.attribute.Nu Instances: 999 Attributes: 19 | Selected attribute Name: response_01 Missing: 0 (0%) Distinct: 3 | Type: Nominal 2 Unique: 0 (0%) |
| Attributes | No. Label | Count |
| All None Invert Pattern | 1 0 2 1 | 581 418 |
| No. Name 7 jobcat 8 union 9 empcat 10 CardZtenurecat 11 equip 12 wireless 13 Internet 14 caliid 15 callwait 16 forward 17 confer 18 ebil 19 response_01 | Class: response_01 (Nom) | Visualize All |
| Remove | | |
| Status OK | | Log 💉 x 0 |

| ٢ | weka.gui.GenericObjectEditor | × | | | | | |
|---------------------------|--|--------|--|--|--|--|--|
| weka.filters.uns About | supervised.attribute.NumericToNominal | _ | | | | | |
| A filter for tu | A filter for turning numeric attributes into nominal ones. More Capabilities | | | | | | |
| attributeIndice | es last | | | | | | |
| invertSelectio | g False | × × | | | | | |
| Open | Save OK Cancel | | | | | | |

3. Depois de aplicar o filtro, o último atributo torna-se **nominal** e é considerado como o rótulo de classe para o dataset - agora o dataset é visualizado em duas cores.

| 9 | Weka | Explorer | | | | - 🗆 × |
|--|---|-------------------------------|---|---------|----------------------|-----------------|
| Preprocess | Classify Cluster Associate Select attributes Visualize | | | | | |
| Open fil | e Open URL Open DB Gen | erate | Undo | Edit | | Save |
| Filter | | | | | | |
| Choose | NumericToNominal -R last | | | | | Apply |
| Current relat Relation: Instances: | ion customer_labThree-weka.filters.unsupervised.attribute.Nu 999 Attributes: 31 | Selected Name: Missing: | attribute : response_01 : 0 (0%) Dist | inct: 2 | Type: N Unique: 0 | lominal (0%) |
| Attributes | | No. | Label | C | ount | |
| Al | None Invert Pattern | | 1 0 2 1 | 58 | 81 18 | |
| No. 19 20 21 22 | Name cord cardtenurecat card2 card2tenurecat | | | | | |
| 23 | equip | Class: resp | ponse_01 (Nom) | | × | Visualize All |
| 24 | wireless Internet | | | | | |
| 25 | callid | 591 | | | | |
| 27 | callwait | | | | | |
| 28 | forward | | | 418 | | |
| 29 | _ conter | | | | | |
| 30 | response_01 | | | | | |
| | Remove | | | | | |
| Status OK | | | | | Log | . x |

4. Se o atributo classe não for o último atributo, você poderá definí-lo na janela de edição (Edit).

| l | 6 | | | 1 | Viewer | | | | × | l | Weka_LabTh | iree [C |
|----|--------------|-----------|----------|--------------|------------|------------|-----------|----------------|----|-----|-----------------------------|---------|
| Re | elation: cus | tomer_lab | Three-we | ka.filters.u | unsupervis | ed.attribu | te.Numeri | cToNominal-Rla | st | | er | |
| | wireless | internet | callid | callwait | forward | confer | ebill | response_0 | 1 | | | |
| c | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Nominal | 0 | Set | mean | |
| 0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0 | | | | |
| 0 | 1.0 | 4.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0 | s | iet | all values to | Edit. |
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0 | - | _ | | |
| 0 | 1.0 | 3.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0 | 5 | et | missing values to | |
| | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | | R | Rep | lace values with | |
| H | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | | 0 | | | | |
| H | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0 | R | Ren | ame attribute | |
| | 0.0 | 3.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0 | A | \tt | ibute as class | |
| H | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | | | | | |
| ľ | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0 | Ľ | Jei | ete attribute | |
| 6 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | × | C |)el | ete attributes | |
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0 | s | or | t data (ascending) | |
| 0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0 | - | - | , (co.co | |
| 0 | 1.0 | 3.0 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0 | C | Opt | imal column width (current) | |
| 0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0 | | | imal column width (all) | |
| 0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0 | | p | imai column width (all) | |
| 0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0 | | | | |
| 0 | 0.0 | 3.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0 | | | | |
| 0 | 0.0 | 3.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0 | | | response_01 (Nom) | |
| 0 | 1.0 | 4.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0 | | | | |
| 0 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0 | | | | |
| 0 | 1.0 | 2.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0 | ~ | | | |
| | c | | | | | | | | > | | | |
| | | | | | | | | | | 1 | | 418 |
| | | | | | | Und | 0 | Cano | el | | | |
| - | | | | | | | | | | - | | |

5. Você também deve converter os tipos de outros atributos. Os atributos region, townsize, agecat, edcat, jobcat, empcat, card2tenurecat e internet são todos valores nominais, no entanto, eles são tratados como tipo numérico pelo Weka. E os atributos gender, union, equip, wireless, callid, callwait, forward, confer, e ebill são todos de valores binários, e também são tratados como tipos numéricos pelo Weka. O filtro NumericToNominal deve ser aplicado para convertê-los. Você também pode Normalize o atributo educat para [0, 1], já que as categorias de educação são rankings.

| C Web | ka | Explo | er | | | | - 🗆 🗙 |
|---|-----|--------------------------|----------------------------|-----------------------|-------------|------------------|-------------------|
| Preprocess Classify Cluster Associate Select attributes Visualize | | | | | | | |
| Open file Open URL Open DB G | ene | erate | | Undo | Ed | it | Save |
| Filter | | | | | | | |
| Choose Normalize -5 1.0 -T 0.0 | | | | | | | Apply |
| Current relation Relation: customer_labThree-weka.filters.unsupervised.attribute.Nu Instances: 999 Attributes: 19 | | Selec Na Miss | ed att me: re ing: 0 | ibute gion (0%) | Distinct: 5 | Type: Unique: | Nominal 0 (0%) |
| Attributes | | No. | | Label | | Count | |
| All None Invert Pattern | | | 1 | 1 2 | | 195 203 | |
| No. Name | | | 3 | 3 | | 217 | |
| 1 🗹 custid | ^ | | 5 | 5 | | 208 | |
| 2 region 3 townsize 4 gender | | | | | | | |
| 5 agecat | | Class: response_01 (Nom) | | | | | Visualize All |
| 7 jobcat 8 union | | | | 202 | 217 | | 208 |
| 9 empcat 10 card2tenurecat | | 195 | | 203 | | 176 | |
| 11 equip | | | | | | - | |
| 13 internet | ~ | | | | | | |
| Remove | | | | | | | |
| Status OK | | | | | | Log | ×0 |

- Seleção de Atributos (Select attributes) Como nem todos os atributos são relevantes para o trabalho de classificação, você deve executar a seleção de atributos antes de treinar o classificador.
- 1. Você pode remover atributos irrelevantes à mão. Por exemplo, o primeiro atributo **custId** deve ser removido. Selecione-o e clique no botão **Remove** para removê-lo.

| 🖸 🔤 We | eka Explorer 🛛 🗕 🗖 💌 |
|--|--|
| Preprocess Classify Cluster Associate Select attributes Visualize | Generate Undo Edit Gene |
| operioka operioka operioa e | Generate Ondo Editari Save |
| Choose NumericToNominal -R. last | Apply |
| Current relation Relation: custome_labThree-weka.filters.unsupervised.attribute.Nu Instances: 999 Attributes: 31 | Selected attribute Type: Nominal Missing: 0 (0%) Distinct: 999 Unique: 999 (100%) |
| Attributes All None Invert Pattern | No. Label Count 1 4459-VLPQUH-30L 1 / / / / / / / / / / / / / / / / / / |
| No. Name 1. Cuetid | 2 21402A2447030 2 3 2228 KOLOPU-FY3 1 4 2866-TTOTKL-TA7 1 5 7217-UECHSF-PCR 1 |
| 2 region 3 townsize 4 gender | 6 4166-WEDDXN-SRK 1 7 1114-UELXX-QT7 1 |
| S Bgecat 0 defat 7 jobcat 8 Union 9 empcat 10 retire | Class: response_01 (Nom) v Visualize All |
| 11 Incet 12 jobsat 13 reside | Too many values to display. |
| Status | |
| OK | Log 💉 X |

 Você também pode executar a seleção automática de atributos. Introduzimos dois métodos de avaliação de atributos de forma individual - InfoGainAttributeEval e ChiSquaredAttributeEval. O método de seleção de atributo padrão de Weka é CfsSubsetEval, que avalia subconjuntos de atributos.

| 0 | | | | Weka Explorer | - | □ × |
|--|---|--|-------------------|---------------|-----|-----|
| Preprocess Attribute E | Classify Cl | uster Associate | Select attributes | Visualize | | |
| abev € 18 € 5 9 0 0 9 0 9 0 9 0 9 0 9 0 9 0 9 0 9 0 9 | tributeSelecti Gristionetti Chisquared Cassifietzi Consistency CostSensiti FilteredAttri InfocainAtt InfocainAtt LatentSeme OnerAttrib SyMatribut Symetrica Symetrica Symetrica | on Vol AttributeEval JobetEval VSubsetEval veSubsetEval ibuteEval setEval setEval stributeEval mitcAnalysis uteEval monents buteEval MucerAttributeE bsetEval | val | i output | | |
| | Filter | Remove fi | ilter Close | | | |
| Status OK | | | | | Log | ×0 |

3. Para usar o avaliador InfoGainAttributeEval, um método de busca Ranker é selecionado para ordenar todos os atributos usando o resultado da avaliação. Usamos todo dataset como conjunto de treinamento. Os resultados mostram que os primeiros 8 atributos são bons.

| 0 | | Weka Explorer | - 🗆 × |
|---|-------------------------------------|-------------------------------|-------|
| Preprocess | Classify Cluster Associat | 2 Select attributes Visualize | |
| Attribute Eva | aluator | | |
| Choose | InfoGainAttributeEva | | |
| Search Meth | bod | | |
| Choose | Ranker -T -1.79769313 | 8623157E308 -N -1 | |
| Attribute Sel | lection Mode | Attribute selection output | |
| Use full Cross-v | training set validation Folds 10 | | |
| | Seed 1 | | |
| (Nom) respo | onse_01 | × _ | |
| Start | Stop | | |
| Result list (rig | ight-click for options) | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| Status | | | |

| ٥ | Weka Explorer - | □ × |
|---|--|------|
| Preprocess Classify Cluster Associate | Select attributes Visualize | |
| Attribute Evaluator | | |
| Choose CfsSubsetEval | | |
| Search Method | | |
| Choose BestFirst -D 1 -N 5 | | |
| Attribute Selection Mode | Attribute selection output | |
| Ouse full training set Cross-validation Folds Seed Seed | Kanked attributes: 0.03779 5 edcat 0.03567 12 internet 0.03259 10 equip 0.01464 16 confer | ^ |
| (Nom) response_01 v | 0.01363 14 calwait 0.01241 17 ebil 0.01241 5 forward | |
| Result list (right-click for options) 11:10:19 - Ranker + InfoGainAttributeEx 11:10:41 - Ranker + InfoGainAttributeEx 11:11:19 - Ranker + InfoGainAttributeEx 11:11:18 - Ranker + InfoGainAttributeEx 11:11:19 - BestFirst + CfsSubsetEval | 0.0074 13 callid 0.0026 2 townsize 0.00254 9 card2tenurecat 0.00253 4 agecat 0.00253 11 wireless 0.00175 3 gender 0.00174 1 region 0.00174 1 region 0.00174 1 region 0.0012 7 union 0.0011 8 empcat Selected attributes: 5 12 10 16 14 17 15 13 2 9 4 11 3 1 6 7 8 • 17 | |
| < >> | < | > |
| Status OK | Log | 🐠 ×0 |

4. Execute a seleção de atributos uma segunda vez mas agora usando o avaliador CfsSubsetEval com o método de busca BestFirst. Compare os resultados dos dois métodos de seleção de atributos.

| ٥ | Weka Explorer – 🗖 💌 | ¢ |
|--|--|-----|
| Preprocess Classify Cluster Associate Attribute Evaluator | Select attributes Visualize | |
| Choose CfsSubsetEval | | |
| Search Method | | - |
| Choose BestFirst -D 1 -N 5 | | |
| Attribute Selection Mode | Attribute selection output | |
| Use full training set Cross-validation Folds 10 Seed 1 | Search Method: Best first. Start set: no attributes Search direction: forward | |
| (Nom) response_01 v | Stale search after 5 node expansions Total number of subsets evaluated: 120 | |
| Start Stop | Merit of best subset found: 0.052 | |
| Result list (right-click for options) 11:10:19 - Ranker + InfoGainAttributeEv 11:10:41 - Ranker + InfoGainAttributeEv 11:10:59 - Ranker + InfoGainAttributeEv | Attribute Subset Evaluator (supervised, Class (nominal): 18 response_01): CFS Subset Evaluator Including locally predictive attributes | 1 |
| 11:11:18 - Ranker + InfoGainAttributEV 11:14:19 - BestFirst + CfsSubsetEval | Selected attributes: 5,10,12,16 : 4 edcat equip internet confer | ~ |
| < > | < >> | |
| Status OK | Log 🗸 🔧 | x 0 |

5. Se você decidir reduzir o conjunto de dados removendo atributos sem importância, escolha **Save reduced data...** clicando com o botão Direito em **Result list**. Salve o arquivo com o nome **customer.arff**.

| 0 | Weka Explorer | - | | × | | |
|--|---|---|--|---|--|--|
| Preprocess Classify Cluster Associate | Select attributes Visualize | | | | | |
| Attribute Evaluator | | | | | | |
| Choose InfoGainAttributeEval | | | | | | |
| Search Method | | | | | | |
| Choose Ranker -T -1.79769313486 | 23157E308 -N 8 | | | | | |
| Attribute Selection Mode | Attribute selection output | | | _ | | |
| Use full training set Cross-validation Folds 10 Seed 1 | Search Method: Attribute ranking. | | | ^ | | |
| (Nom) response_01 v | Attribute Evaluator (supervised, Class (nominal): 30 response_01): Information Gain Ranking Filter | | | | | |
| Start Stop | Ranked attributes: | | | | | |
| Result list (right-dick for options) | 0.03779 5 edcat | | | | | |
| 10:19:02 - Ranker + InfoGainAttributeEv | 0.03274 24 internet | | | | | |
| 10:19:49 - Ranker + InfoGainAttributo | 0.03259 22 equip | | | | | |
| view | vin main window | | | | | |
| View | v in separate window | | | | | |
| Sav | e result buffer vard | | | | | |
| Dele | te result buffer .id | | | | | |
| Visu | alize reduced data :: 5,24,22,28,26,29,27,25 : 8 | | | | | |
| Save reduced data | | | | | | |
| | | | | ~ | | |
| × > | < | | | > | | |
| Status | | | | | | |

3. Classificador Naïve Bayes: bayes / NaïveBayes

1. Abra o dataset salvo **customer.arff** e clique na guia **Classify** na parte superior da janela. Clique no botão Choose abaixo de *Classifier*. A lista drop-down de todos os classificadores são exibidos. Escolha **NaiveBayes** da pasta **bayes**.

| ٢ | Weka Explorer | - | | × |
|---|-------------------|---|----|-------|
| Preprocess Classify Cluster Associate Select at | ributes Visualize | | | |
| Classifier | | | | |
| Weka dassifiers dassifier dassifie | Close | | | |
| OK | Lo | g | C. | . × 0 |

2. Clique (com o botão Esquerdo) dentro da caixa Filter, e então a janela de propriedades é apresentada. Então, a janela de propriedades do NaiveBayes será aberta, se você não quiser usar a Distribuição Normal para dados numéricos, defina useKernelEstimator para true; Você também pode realizar discretização supervisionada em dados numéricos definindo useSupervisedDiscretization como true. Clique no botão OK para salvar todas as configurações.

| O 1 | weka.gui.GenericObjectEditor | | | | | | | |
|---|------------------------------|---|--|--|--|--|--|--|
| weka.classifiers.bayes.NaiveBayes About | | | | | | | | |
| Class for a Naive Bayes classifier using estimator classes. More Capabilities | | | | | | | | |
| deb | ig False | ~ | | | | | | |
| displayModelInOldForm | at False | ~ | | | | | | |
| useKernelEstimat | or False | ~ | | | | | | |
| useSupervisedDiscretizatio | False | ~ | | | | | | |
| Open | Save OK Cancel | I | | | | | | |

3. Para dividir o dataset em training set and testing set, escolha a Cross-validation de 10 vezes. Para usar conjuntos de training, validation and testing set, escolha Supplied test set, após realizar Cross-validation. Neste formato de Cross-validation+Supplied test set, todo dataset inicial é usado para training e validation, e o testing set é aquele escolhido.

| 0 | Weka Explorer | - | |
|--------------------------|---|-----|-----|
| Preprocess Classify | Y Cluster Associate Select attributes Visualize | | |
| Classifier | | | |
| Choose Naiv | reBayes | | |
| Test options | Classifier output | | |
| O Use training set | t | | |
| Supplied test se | et Set | | |
| Cross-validation | n Folds 10 | | |
| O Percentage spli | lit % 66 | | |
| | fore options | | |
| | | | |
| (Nom) play | ~ | | |
| | | | |
| Start | Stop | | |
| Result list (right-click | k for options) | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| Status | | | |
| OK | | Log | 100 |

4. Clique no botão **Start**, à esquerda da janela, e então o algoritmo começa a ser executado. A saída é apresentada na janela da direita.

| 3 | | Weka Ex | plorer | |
|---------------------------------------|--------------------------|---------|--------|-----------------------------|
| Preprocess Classify Cluster Associat | e Select attributes Visu | alize | | |
| Classifier | | | | |
| Choose NaiveBayes | | | | |
| Test options | Classifier output | | | |
| O Use training set | | Class | | |
| O Supplied test sat | Attribute | 0 | 1 | |
| Supplied test set | | (0.58) | (0.42) | |
| Cross-validation Folds 10 | | | | |
| O Percentage split % 66 | edcat | 0 4255 | 0 205 | |
| | atd day | 0.4355 | 0.305 | |
| More options | weight sum | 581 | 418 | |
| | precision | 0.25 | 0.25 | nonomators of normal |
| (Nom) response_01 | Y | 0.20 | | parameters of normal |
| | equip | | | distributions for numeric |
| Start Stop | 0 | 361.0 | 341.0 | |
| Result list (right-click for options) | 1 | 222.0 | 79.0 | |
| 11:23:37 - bayes.NaiveBayes | [total] | 583.0 | 420.0 | |
| | | | | |
| | internet | | | |
| | 0 | 269.0 | 283.0 | frequency counts of |
| | 1 | 100.0 | 52.0 | nominal values |
| | 2 | 71.0 | 32.0 | nominal values |
| | 3 | 72.0 | 34.0 | |
| | 4 | 74.0 | 22.0 | |
| | [tota1] | 586.0 | 423.0 | |
| | confer | | | NaiveBayes avoids zero |
| | 0 | 304.0 | 159.0 | |
| | 1 | 279.0 | 261.0 | frequencies by applying the |
| | [total] | 583.0 | 420.0 | Laplace correction |
| | | | | Lupiuce correction. |
| itatus | | | | |
| OK. | | | | Log |

| ٥ | Weka Explorer | - 🗆 × |
|---|--|---------------------------|
| Preprocess Classify Cluster Associate S | elect attributes Visualize | |
| Classifier | | |
| Choose NaiveBayes | | |
| Test options | Classifier output | |
| O Use training set | Time taken to build model: 0.03 seconds | |
| O Supplied test set Set | === Stratified cross-validation === | Accuracy |
| Cross-validation Folds 10 | === Summary === | |
| O Percentage split % 66 | Correctly Classified Instances 616 | 61.6617 % |
| More options | Incorrectly Classified Instances 383 | 38.3383 % |
| | Kappa statistic 0.22 | 235 |
| (Nom) response 01 | Mean absolute error 0.42 | 267 |
| (tony) coponec_or | Root mean squared error 0.40 | 331 |
| Start Stop | Relative absolute error 87.60 Root relative squared error 97.92 | 567 % 272 % |
| Result list (right-click for options) | Total Number of Instances 999 | |
| 11:23:37 - bayes.NaiveBayes | | |
| | === Detailed Accuracy By Class === | |
| | TP Rate FP Rate Precision | Recall F-Measure ROC Area |
| | 0.633 0.407 0.684 | 0.633 0.658 0.663 |
| | 0.593 0.367 0.538 | 0.593 0.564 0.663 |
| | Weighted Avg. 0.617 0.39 0.623 | 0.617 0.619 0.663 |
| | === Confusion Matrix === | |
| | a b < classified as | |
| | 368 213 a = 0 | |
| | 170 248 b = 1 | |
| | | ~ |
| | | |
| Status | | Log x0 |
| UN | | |

K-Nearest-Neighbor: lazy/IBK

1. Gostaríamos de realizar a classificação K-Nearest-Neighbor no mesmo dataset.

Para isso você deverá escolher: classifiers => lazy => IBK.

Você pode experimentar valores diferentes de K e ver qual valor dá um resultado melhor. Compare os resultados com o classificador Naïve Bayes.

| 0 | | | | | Weka Ex | plorer | | | | - 🗆 | × |
|-------------------------|-----------|---------|------------|-------------------|-----------------|---------------|------------------|----------------|-----------|-------|------|
| Preprocess | Classify | Cluster | Associate | Select attributes | Visualize | | | | | | |
| Classifier | | | | | | | | | | | |
| 🚺 weka | | | | | hearNNSearch -A | \"weka.core.E | uclideanDistance | -R first-last\ | | | |
| 🔄 🗌 🛅 🔒 da | assifiers | | | | | | | | | | |
| i 🖶 🗍 | bayes | | | | | | | | | | |
| Ē. ₽. | function | s | | | | | | | | | ^ |
| P | lazy | | | | o build mode | 1: 0 seco | nds | | | | |
| | • IB1 | | | | | | | | | | |
| | | - | | | ed cross-val | idation = | | | | | |
| | • 1 BR | | | | | | | | | | |
| | | | | | | | | | | | |
| 📕 🕴 🔒 | meta | | | | assified Ins | stances | 11 | | 78.5714 | 8 | |
| - ÷ | mi | | | | Classified I | instances | 3 | | 21.4286 | ş | |
| 📕 🌔 👜 🗐 | misc | | | | tic | | 0.55 | 32 | | | |
| E 🕂 🕂 🕂 | rules | | | | e error | | 0.25 | 37 | | | |
| ∎ [] ⊕•] | trees | | | | uared error | | 0.43 | 49 | | | |
| -F | | | | | olute error | | 53.28 | 57 % | | | |
| | | | | | e squared er | ror | 88.15 | 83 % | | | |
| 1 | | | | | of instance | :3 | 14 | | | | |
| | | | | | Decument Pr | | _ | | | | |
| | | | | | Accuracy by | CI455 | - | | | | |
| | | | | | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC A | rea |
| | | | | | 0.778 | 0.2 | 0.875 | 0.778 | 0.824 | 0.8 | 11 |
| | | | | | 0.8 | 0.222 | 0.667 | 0.8 | 0.727 | 0.8 | 11 |
| | | | | | . 0.786 | 0.208 | 0.801 | 0.786 | 0.789 | 0.8 | 11 |
| | Filter | ··· | Remove fil | ter Close | | | | | | | |
| | | | | === Couras | ron Matrix === | | | | | | |
| | | | | | | | | | | | |
| | | | | ab < | classified as | 1 | | | | | |
| | | | | 72 a= | yes | | | | | | |
| | | | | 14 b = | no | | | | | | |
| | | | | | | | | | | | ~ |
| | | | | < | | | | | | | > |
| Status | | | | | | | | | | _ | |
| OK | | | | | | | | | Log | .409 | A X0 |
| | | | | | | | | | | 100 | P |

| weka.gui.GenericObjectEditor | | | | | | |
|---------------------------------------|--------------------------------------|---------|--|--|--|--|
| weka.classifiers.lazy.IBk About | | | | | | |
| K-nearest neighbours classifier. More | | | | | | |
| Capabilities | | | | | | |
| KNN | 1 | | | | | |
| crossValidate | False | ~ | | | | |
| debug | False | ~ | | | | |
| distanceWeighting | No distance weighting | | | | | |
| meanSquared | False | ~ | | | | |
| nearestNeighbourSearchAlgorithm | Choose LinearNNSearch - A "weka.core | .Euclie | | | | |
| windowSize | 0 | | | | | |
| Open Save. | OK Cancel | | | | | |

| 0 | Weka Explorer | | | - 🗆 × | | |
|---------------------------------------|--|---------------------|------------------|----------|--|--|
| Preprocess Classify Cluster Associate | Select attributes Visualize | | | | | |
| Classifier | | | | | | |
| Choose IBk -K 3 -W 0 -A "weka.core | neighboursearch.LinearNNSearch -A \"weka.core.Eu | uclideanDistance -R | R first-last\"" | | | |
| Test options | Classifier output | | | | | |
| ○ Use training set | | | | ^ | | |
| O Supplied test set Set | Correctly Classified Instances | 596 | 59.6597 | 8 | | |
| | Incorrectly Classified Instances | 403 | 40.3403 | 8 | | |
| Cross-validation Folds | Kappa statistic | 0.1521 | L | | | |
| O Percentage split % 66 | Mean absolute error | 0.4414 | 1 | | | |
| More options | Root mean squared error | 0.4891 | L 5 e | | | |
| | Relative absolute error | 90.0020 | 1020 % | | | |
| (h) | Root relative squared error 99.14 | | | | | |
| (Nom) response_01 | Total Number of Instances | 555 | | | | |
| Start Stop | === Detailed Accuracy By Class === | | | | | |
| Result list (right-click for options) | TD Data FD Data | Precision | Decall E-Measure | POC Are: | | |
| 12:27:06 - bayes.NaiveBayes | 0.711 0.562 | 0.637 | 0.711 0.672 | 0.628 | | |
| 12:27:48 - bayes.NaiveBayes | 0.438 0.289 | 0.521 | 0.438 0.476 | 0.628 | | |
| 12:32:16 - lazy.IBk | Weighted Avg. 0.597 0.448 | 0.589 | 0.597 0.59 | 0.628 | | |
| | | | | | | |
| | === Confusion Matrix === | | | | | |
| | a b < classified as | | | | | |
| | | | | | | |
| | 235 183 b = 1 | | | | | |
| | | | | ~ | | |
| | < | | | > ` | | |
| | | | | - | | |
| Status OK | | | Log | x0 | | |
| | | | | | | |

| 0 | Weka Explorer | - 🗆 × |
|---------------------------------------|---|-----------------------------------|
| Preprocess Classify Cluster Associate | Select attributes Visualize | |
| Classifier | | |
| Choose IBk -K 20 -W 0 -A "weka.cor | e.neighboursearch.LinearNNSearch -A \"weka.core.Euclide | leanDistance -R first-last\"" |
| Test options | Classifier output | |
| O Use training set | === Stratified cross-validation === | ^ |
| O Supplied test set Set | === Summary === | |
| Cross-validation Folds 10 | Correctly Classified Instances | 607 60.7608 % |
| O Percentage split % 66 | Incorrectly Classified Instances | 392 39.2392 % |
| More options | Kappa statistic | 0.1599 |
| Hore options | Rean absolute error | 0.4423 |
| (Nom) response 01 | Relative absolute error | 90.8728 % |
| (tony response_or | Root relative squared error | 96.5204 % |
| Start Stop | Total Number of Instances | 999 |
| Result list (right-click for options) | Detailed Accuracy By Class | |
| 12:27:06 - bayes.NaiveBayes | Decalled Accuracy by class | |
| 12:27:48 - bayes.NaiveBayes | TP Rate FP Rate Pre | ecision Recall F-Measure ROC Area |
| 12:32:16 - Iazy.IBk | 0.766 0.612 | 0.635 0.766 0.694 0.654 |
| 12:41:02 - lazy.IBk | 0.388 0.234 | 0.544 0.388 0.453 0.654 |
| 12:41:19 - lazy.IBk | Weighted Avg. 0.608 0.454 | 0.597 0.608 0.593 0.654 |
| | === Confusion Matrix === | |
| | a b < classified as | |
| | 445 136 a = 0 | |
| | 256 162 b = 1 | v |
| | ٢ | > |
| Status OK | | Log x0 |

Árvores de Decisão: trees/J48 (Implementing C4.5)

1. Gostaríamos de construir um modelo de árvore de decisão no mesmo dataset de treinamento. Para isso você deverá escolher: classifiers => lazy => IBK.

Utilize todos os valores padrão dos parâmetros e depois gere diferentes árvores de decisão mudando estes parâmetros (confidenceFactor, minNumObj e numFolds).

| ٢ | | | | | Weka Exp | olorer | | | | | × |
|---|---|------------------|---|------------------|--|--|---|---------------------------------|-----------------------------------|------------------------------------|--------|
| Preprocess | Classify Cluster | Associate | Select attribute | s V | isualize | | | | | | |
| Classifier | | | | | | | | | | | |
| Classifier weka - da - da - da - da - da - da - da - d | assifiers bayes functions loary meta misc rules trees • ADTree • DecisionStump • FT • 148graft • 148graft | 2 | | * | o build mode ed cross-val === assified Ins Classified I tic e error uared error olute error olute error e squared er | 1: 0 secon idation == tances nstances ror s | nds -0.02: 0.52: 0.57: 109.37: 115.68: 14 | 44 08 08 11 % 59 % | 57.1429 42.8571 | 8 | |
| | UMT MSP MSP RandomFores RendomForee REFTree SimpleCart Filter | t Remove filt | er Close a b <- 7 2 a 4 1 b < | cc = y = n | Accuracy By IP Rate 0.778 0.2 0.571 n Matrix ==== classified as res | Class === FF Rate 0.8 0.222 0.594 | Precision 0.636 0.333 0.528 | Recall 0.778 0.2 0.571 | F-Measure 0.7 0.25 0.539 | ROC Are 0.333 0.333 0.333 | e ~ |
| Status OK | | | | | | | | | Log | ~ | . x C |

| 0 | weka.gui.GenericObjectEc | ditor × | | | | | | |
|-------------------------|-----------------------------|---------|--|--|--|--|--|--|
| weka.classifiers.trees. | 148 | | | | | | | |
| Class for generat | ng a pruned or unpruned C4. | More | | | | | | |
| | Capabilities | | | | | | | |
| binarySplits | False | ~ | | | | | | |
| confidenceFactor | 0.25 | | | | | | | |
| debug | False | ~ | | | | | | |
| minNumObj | 2 | | | | | | | |
| numFolds | 3 | | | | | | | |
| reducedErrorPruning | False | ~ | | | | | | |
| saveInstanceData | False | ~ | | | | | | |
| seed | 1 | | | | | | | |
| subtreeRaising | True | ~ | | | | | | |
| unpruned | False | ~ | | | | | | |
| useLaplace | False | ~ | | | | | | |
| Open | Save OK | Cancel | | | | | | |



| 0 | Weka Explorer – 🗖 🗙 | |
|---------------------------------------|---|----|
| Preprocess Classify Cluster Associate | Select attributes Visualize | |
| Classifier | | |
| Choose 348 -C 0.25 -M 2 | | |
| Test options | Classifier output | 5 |
| Use training set | A Stratified gross_validation | |
| O Supplied test set Set | === Summary === | |
| Cross-validation Folds 10 | Correctly Classified Instances 613 61 3614 \$ | |
| O Percentage split % 66 | Incorrectly Classified Instances 386 38.6386 % | |
| Marris and Kana | Kappa statistic 0.1693 | |
| More options | Mean absolute error 0.4571 | |
| | Root mean squared error 0.4826 | |
| (Nom) response_01 v | Relative absolute error 93.9202 % | |
| | Root relative squared error 97.832 % | |
| Start Stop | Total Number of Instances 999 | |
| Result list (right-click for options) | | |
| 12:27:06 - bayes.NaiveBayes | === Detailed Accuracy By Class === | |
| 12:27:48 - bayes.NaiveBayes | | 11 |
| 12:32:16 - lazy.IBk | TP Rate FP Rate Precision Recall F-Measure ROC Area | |
| 12:40:47 - Iazy.IBK | 0.781 0.62 0.637 0.781 0.702 0.619 | |
| 12:41:02 - Id2y.IDK | 0.38 0.219 0.556 0.38 0.452 0.619 | |
| 12:46:25 - trees. 348 | Weighted Avg. 0.614 0.452 0.603 0.614 0.597 0.619 | |
| | Confusion Matrix | |
| | a b < classified as | |
| | 454 127 a = 0 | |
| | 259 159 b = 1 | |
| | | |
| | v | |
| | < > | |
| Statue | | 2 |
| OK | Log x | 0 |

2. Para visualizar a árvore de decisão que construímos, clique com o botão **Direito** no item **trees.J48** da lista de resultados.

| | we | |
|--|---------------------------------------|--|
| Preprocess Classify Cluster Asso | ciate Select attributes Visualize | |
| Classifier | | |
| Choose 046 -C 0.25 -M 2 | | |
| Test options | Classifier output | |
| Use training set | Scheme:weka.classif | iers.trees.148 -C 0.25 -M 2 |
| O Suppled test set Set | Relation: custo | mer labThree-weka.filters.unsupervised.attribute.Numeric |
| Oroccuralidation Ende 10 | Instances: 999 | |
| | Attributes: 5 | |
| O Percentage spit % 66 | edcat | |
| More options | equir | |
| | incer | net. |
| (Nom) response_01 | v respo | nse 01 |
| | Test mode:10-fold o | ross-validation |
| Start Stop | | |
| Result list (right-dick for options) | Classifier mode | l (full training set) |
| 12:27:06 - bayes.NaiveBayes | | |
| 12:27:48 - bayes.NaiveBayes | J48 pruned tree | |
| 12:32:16 - Iazy.16K 12:40:47 - Jazy 18k | | |
| 12:41:02 - lozy.IBk | equip = 0 | |
| 12:41:19 - lazy.IBk | edcat <= 0.25 | |
| 12:46:25 - trees.348 | and a sector of a dama | (213.0/101.0) |
| VI | ew in main window | (254.0/98.0) |
| Vi | ew in separate window | 33.0/83.0) |
| Sa | ve result buffer | .0) |
| De | elete result buffer | |
| | - Local I | |
| Lo | aa model | 7 |
| Sa | ve model | > |
| Re | -evaluate model on current test se | t |
| OK Vi | sualize classifier errors | Log |
| Vi | sualize tree | |
| Options v | sualize margin curve | |
| Vi | sualize threshold curve | • |
| AGE 14 OF 14 625 WOR | ost/Benefit analysis | • |
| | sualize cost curve | |
| | | |
| | | |



3. Os modelos de classificação treinados podem ser salvos clicando com o botão **Direito** nos itens da lista de resultados.

| e Charife at a | | Wek | a explorer | | | | |
|---|---|--|-------------|----------------------|---------|-----------|----------|
| Preprocess Classify Clust | er Associate S | Select attributes Visualize | | | | | |
| Classifier | | | | | | | |
| Choose 348 -C 0.25 | -M 2 | | | | | | |
| Test options | | Classifier output | | | | | |
| O Use training set | | Time taken to build | model: 0 s | econds | | | |
| O Supplied test set | Set | Stratified eres | -validatio | n | | | |
| Cross-validation Fol | ds 10 | === Summary === | -varidatio | | | | |
| Percentage split | % 66 | | | | | | |
| | | Correctly Classified | i Instances | 607 | | 60.7608 | 8 |
| More options. | | Incorrectly Classifi | ed Instanc | es 392 | | 39.2392 | ş |
| | | Kappa statistic | | 0.1 | 599 | | |
| (Nom) response_01 | ~ | Root mean squared er | ror | 0.44 | 123 | | |
| Start | Stop | Relative absolute en | ror | 90.8 | 128 % | | |
| | o top | Root relative square | ed error | 96.5 | 204 % | | |
| Result list (right-click for op | ions) | Total Number of Inst | ances | 999 | | | |
| 12:32:16 - lazy.IBk 12:40:47 - lazy.IBk 12:41:02 - lazy.IBk | | TP Re | te FP Ra | te Precision | Recall | F-Measure | ROC Area |
| 12:41:19 - lazy.IBk | View in mai | n window | 38 0.2 | 12 U.635 34 0.544 | 0.388 | 0.694 | 0.654 |
| 12:46:25 - trees. J48 | View in sen | arate window | 0.4 | 54 0.597 | 0.608 | 0.593 | 0.654 |
| | Save result I | huffer | | | | | |
| | Delete esuit i | buffer | | | | | |
| | Delete resul | t butter | lifted as | | | | |
| | Load mode | I | Liled as | | | | |
| | Save model | | | | | | |
| | | model on current test set | | | | | > |
| | Re-evaluate | | | | | | |
| Status | Visualize cla | ssifier errors | - | | | | |
| Status OK | Visualize cla | ssifier errors e | | | | Log | |
| Status OK | Visualize cla Visualize tre Visualize ma | e ergin curve | | | | Log | ~ |
| Status OK Options 💌 | Ke-evaluate Visualize cla Visualize tre Visualize ma Visualize the | e ergin curve reshold curve | | Ensembl | e: Meta | Log | ~ |
| Status OK Options 💌 | Ke-evaluate Visualize cla Visualize tre Visualize ma Visualize the Cost/Benefi | e argin curve reshold curve t analysis | | Ensembl | e: Meta | Log | 1 |
| Status OK Options V AGE 14 OF 15 639 W(| Ke-evaluate Visualize cla Visualize tre Visualize ma Visualize the Cost/Benefi Visualize co | e ergin curve reshold curve t analysis t curve | | Ensembl | e: Meta | Log | ~ |



Ensemble (Metaleaning) classifier.meta.Voting

1. Você pode combinar vários classificadores para executar um método conjunto. Para isso você deverá escolher: classifiers => meta => Vote.

| Preprocess Classify Cluster Associate Select | attributes Visualize | | |
|--|---|--------------------------------|---------------------------------|
| Classifier | | | |
| Choose Vote -S 1 -B "weka.classifiers.baye | s.NaiveBayes " -B "weka.classifiers.lazy.IB | k -K 20 -W 0 -A \"weka.core.ne | ighboursearch.LinearNNSearch -A |
| Test options Class | sifier output | | |
| ○ Use training set | | | ^ |
| Cuplied test est | Stratified cross-validation | === | |
| | = Summary === | | |
| Cross-validation Folds 10 | rrectly Classified Instances | 619 | 61 962 \$ |
| OPercentage split % 66 Inc | correctly Classified Instances | 380 | 38.038 % |
| More options Kar | opa statistic | 0.2089 | |
| More options | an absolute error | 0.442 | |
| Roc | ot mean squared error | 0.4756 | |
| (Nom) response_01 V Rel | lative absolute error | 90.8199 % | |
| Start Stop | ot relative squared error | 96.4051 % | |
| Tot | tal Number of Instances | 999 | |
| Result list (right-click for options) | | | |
| 12:27:06 - bayes.NaiveBayes === | = Detailed Accuracy By Class = | == | |
| 12:27:48 - Dayes.NaiveBayes | TD Date FD Date | Precision Pecall | E-Massure DOC Are: |
| 12:40:47 - Jazy.IBk | 0.704 0.498 | 0.663 0.704 | 0.683 0.657 |
| 12:41:02 - lazy.IBk | 0.502 0.296 | 0.55 0.502 | 0.525 0.657 |
| 12:41:19 - lazy.IBk Wei | ighted Avg. 0.62 0.413 | 0.616 0.62 | 0.617 0.657 |
| 12:46:25 - trees.J48 | | | |
| 13:38:26 - meta.Vote | = Confusion Matrix === | | |
| | | | |
| | a b < classified as | | |
| 40 | $09 \ 172 a = 0$ | | |
| 20 | 38 210 D = 1 | | |
| | | | |
| | | | Y |
| | | | |
| Status OK | | | Log 💉 🛛 |

| 0 | weka.gui.GenericObjectEditor | × | | |
|---------------------------------------|-------------------------------|---|--|--|
| weka.classifiers.me About | eta.Vote | | | |
| Class for combining classifiers. More | | | | |
| | Capabilities | | | |
| classifiers | 3 weka.classifiers.Classifier | | | |
| debug | False | • | | |
| seed | 1 | | | |
| Open | Save OK Cancel | | | |