GEAM
Grupo de Estudos de Aprendizado de Máquina

- Bolsista: Ítalo Lima Dantas- Bolsa de Iniciação Acadêmica (BIA)
- Orientador: Régis Pires Magalhães

Sumário

O Problema

- O exemplo a seguir é retirado do mundo da óptica oftálmica. O objetivo é determinar se um paciente é adequado ao uso de lentes de contato e para qual tipo de contato.
- O tipo de problema abordado é o de Classificação, no qual obteremos “Lenses” como Label, classificando-o em Hard, Soft ou None, o que indica se o paciente necessita do uso de lentes, e caso necessite, se serão lentes gelatinosas ou rígidas.
- Há 4 feactures: Age, Prescription, Astigmastic e Tear_rate, que nos orientarão para a resolução do problema. (Idade do paciente, prescrição oftálmica sobre distúrbios oculares, astigmatismo e taxa de produção de lágrimas.

Algoritmos

- 2 algoritmos serão utilizados: Árvore e Regressão Logística, no fim os dois modelos serão comparados.
- Regressão Logística é um recurso estatístico que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.
- Árvores de decisão são modelos estatísticos que utilizam um treinamento supervisionado para a classificação e previsão de dados. Em outras palavras, em sua construção é utilizado um conjunto de treinamento formado por entradas e saídas.

Datasets

Info

63 datasets
7 datasets cached

Filter

Title	Size	Instances	Variables	Target	Tags
• Car Evaluation	50.7 KB	1728	6	C categorical	synthetic
• Conferences	2.3 KB	42	5		
• Grades for English and Math	265 bytes	12	3		synthetic, educational
• Lenses	968 bytes	24	5	C categorical	medical
• Philadelphia Crime	90.5 KB	9666	4	C categorical	criminology, time, geo
• Sailing	455 bytes	20	4	C categorical	synthetic
• Titanic	44.1 KB	2201	4	C categorical	
Breast Cancer and Docetaxel ...	1.8 MB	24	9486	C categorical	biology
Smoking effect on B lymphoc...	1.8 MB	79	3000	C categorical	genomics
Bone marrow mononuclear c...	582.0 KB	96	1000	C categorical	genomics
HDI	65.1 KB	188	66	N numeric	economy, geo
Abalone	187.5 KB	4177	8	N numeric	biology
Adult	4.1 MB	32561	15	C categorical	economy
Attrition - Predict	838 bytes	3	18	C categorical	economy, synthetic, education
Attrition - Train	182.2 KB	1470	18	C categorical	economy, synthetic
Auto MPG	17.3 KB	398	9	N numeric	
Bank Marketing	466.1 KB	4119	20	C categorical	economy
Banking Crises	31.3 KB	211	73		time, economy
Bone Healing	11.6 KB	37	0	C categorical	image analytics, biology

Description

Lenses (1990), from UCI ML Repository

The following example is taken from the world of ophthalmic optics. The aim is to determine whether a patient is suitable for contact lens wear and for which type of contacts.

References

Cendrowska, J., PRISM: An algorithm for inducing modular rules, International Journal of Man-Machine Studies, 1987, 27, 349-370.

Send Data

O Processo

- Fazer upload do Data Set (Lenses).
- Selecionar diferentes formas de visualizar os dados com os seguintes widgets:
- Box Plot, Distribuition, Data Table.
- Usar os widgets dos algoritmos: Logistic Regression e Tree, que serão ligados ao data set.
- O Widget Tree Viewer é utilizado para ver os nós da árvore, os quais indicam a ordem de importância das feactures para a tomada de decisão.

Orange Canvas

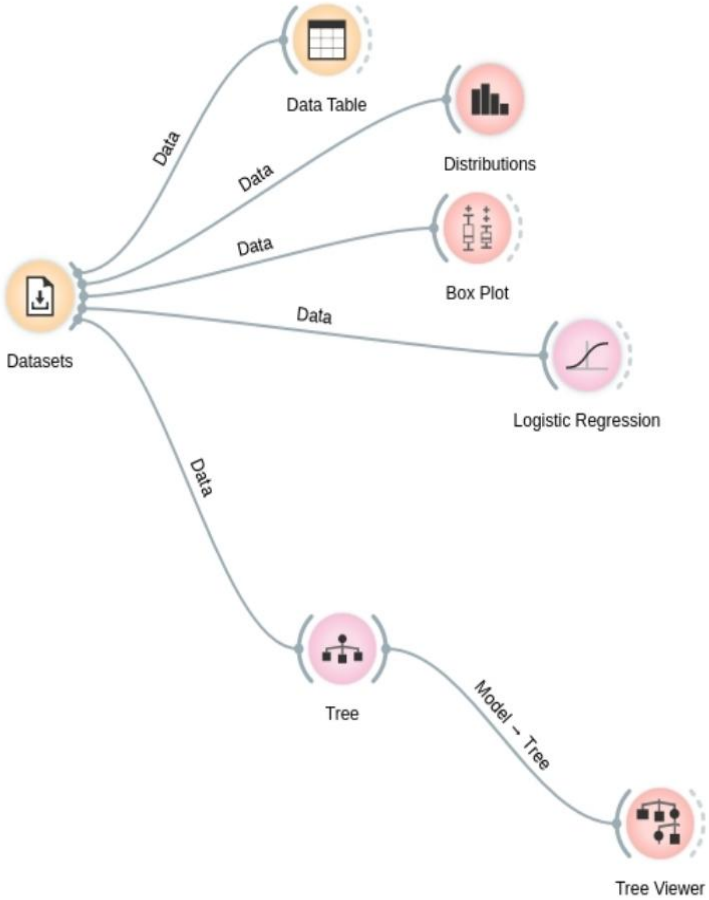
File Edit View Widget Options Help

Data

File
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Colum...
Select Rows
Rank
Correl...
Merge Data
Conca...
Select by Dat...
Transp...
Rando...
Prepro...
Transf...
Impute
Outliers

Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.



Continuação do Processo

- Após inserir os dois algoritmos que serão utilizados, vamos agora selecionar o widget predictions, para fazer uma predição em novos dados.
- Ligar o widget Tree ao widget predictions.
- Assim, teremos a exibição das predições de acordo com os dados estabelecidos.
- Deve-se também ligar o widget Logistic Regression ao widget predictions.
- Então, liga o widget do Data Set ao widget de predictions, para ter a base de dados.

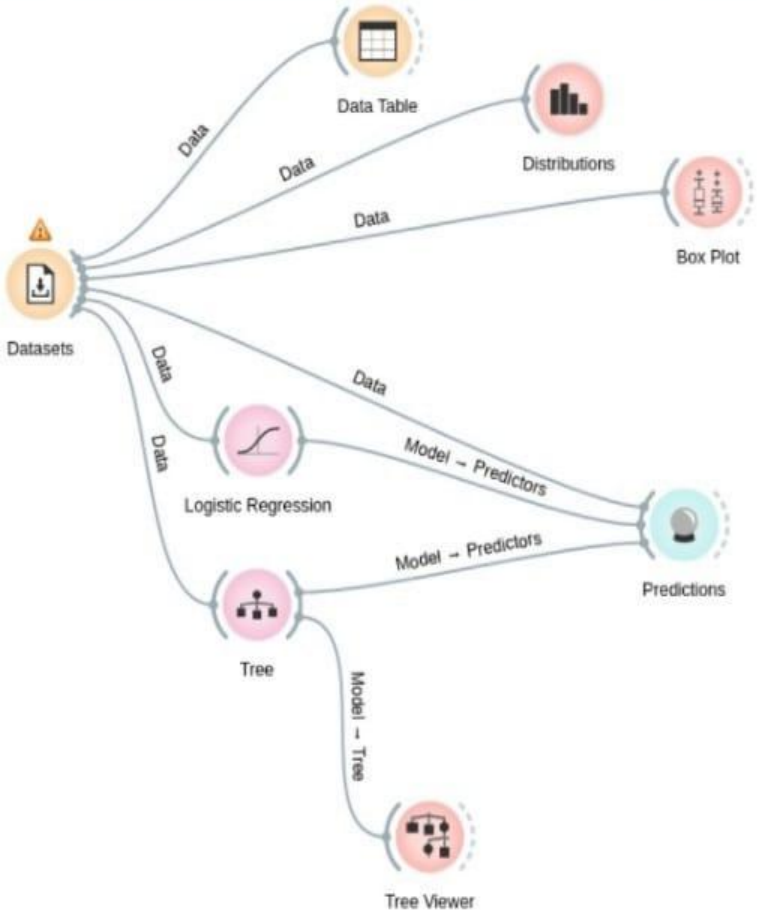
File Edit View Widget Options Help

Data

File
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Colum...
Select Rows
Rank
Correl...
Merge Data
Conca...
Select by Dat...
Transp...
Rando...
Prepro...
Transf...
Impute
Outliers

Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.



Usando novos dados

- Nesse passo, deve-se também criar um arquivo em formato “csv” com alguns dados de acordo com os do problema, os quais informaremos os valores nas 4 feactures e teremos como retorno o label que corresponda ao caso.
- Passos:
- Abrir o dontpad.com/grupogeam10/09.
- Pegar os dados que lá estão e inseri-los em uma planilha. (Obs: os nomes das feactures na planilha devem estar iguais aos do dataset.
- Salvar o arquivo.
- Abrir o widget FILE, no Orange e fazer o carregamento desses dados.
- Exibir os dados em uma tabela, com o widget data table.
- Ligar os dados ao widget predictions

lenses2.ows*

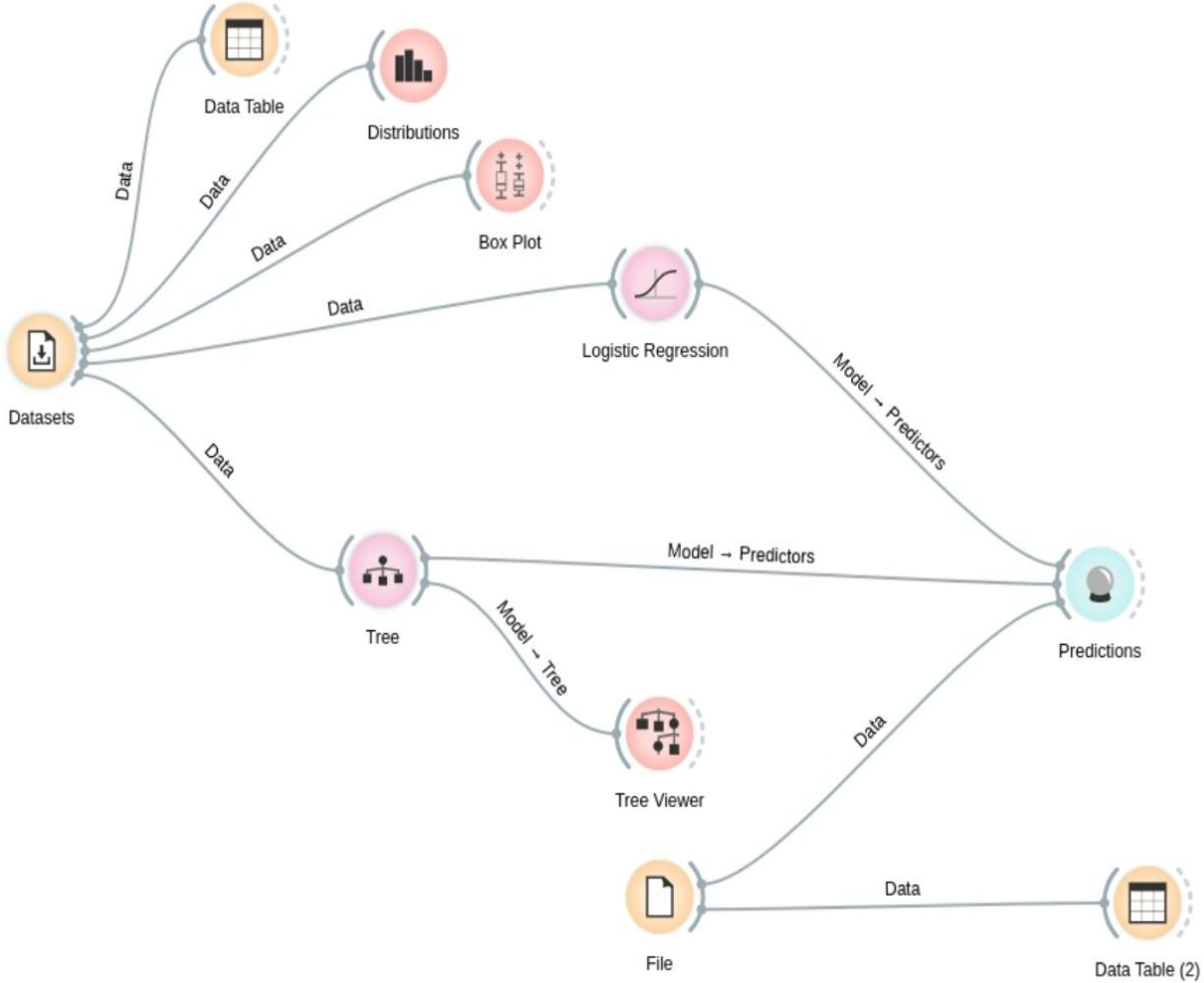
File Edit View Widget Options Help

Data

File
Datasets
SQL Table
Data Table
Paint Data
Data Info
Data Sampler
Select Colum...
Select Rows
Rank
Correl...
Merge Data
Conca...
Select by Dat...
Transp...
Rando...
Prepro...
Transf...
Impute
Outliers

Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.



Testando os Algoritmos

- Agora é hora de usar o widget Test Score, para comparar qual dos dois modelos é mais adequado para a resolução do problema.
- A eficiência é medida pelo Acuraccy.
-

Sampling

- Cross validation
  - Number of folds: 10
  - Stratified
- Cross validation by feature
- Random sampling
  - Repeat train/test: 10
  - Training set size: 70 %
  - Stratified
- Leave one out
- Test on train data
- Test on test data

Target Class

(Average over classes)

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Tree	0.906	0.812	0.821	0.860	0.812
Logistic Regression	0.906	0.750	0.749	0.750	0.750

Conclusão

- Com base na imagem anterior, percebemos que o algoritmo de Árvore é mais eficiente que o de Regressão Logística.
- Um dos fatores é que o de Regressão Logística geralmente é utilizado quando o label é de fator binário, o que não é o caso nesse problema, já que temos 3 possíveis resultados.

Orange Canvas

File Edit View Widget Options Help

Data

File Datasets SQL Table Data Table

Paint Data Data Info Data Sampler Select Colum...

Select Rows Rank Correl... Merge Data

Conca... Select by Dat... Transp... Rando...

Prepro... Transf... Impute Outliers

Test & Score

Cross-validation accuracy estimation.

more...