

O emprego de R na detección das características máis influentes na clasificación de pacientes infectados por COVID-19 en Galicia

VII Xornada de Usuarios de R en Galicia

Laura Davila Pena, Balbina Casas Méndez, Ignacio García Jurado

15 de outubro de 2020



Problemas de clasificación

Introdución

Problema de clasificación

Un **problema de clasificación** consiste en predicir o valor dunha variable resposta cualitativa para un ou máis individuos, facendo uso dos valores que xa coñecemos de certas variables categóricas (ou atributos) de tales individuos.

Problema de clasificación

Un **problema de clasificación** consiste en predicir o valor dunha variable resposta cualitativa para un ou máis individuos, facendo uso dos valores que xa coñecemos de certas variables categóricas (ou atributos) de tales individuos.

Predicións → Coñecemento obtido a través dunha mostra de individuos con valores coñecidos dos atributos e resposta.

Problema de clasificación

Un **problema de clasificación** consiste en predicir o valor dunha variable resposta cualitativa para un ou máis individuos, facendo uso dos valores que xa coñecemos de certas variables categóricas (ou atributos) de tales individuos.

Predicións → Coñecemento obtido a través dunha mostra de individuos con valores coñecidos dos atributos e resposta.



Machine learning

Problemas de clasificación

Clasificadores

- Moitos clasificadores, ademais de clasificar, permiten avaliar a importancia que os diversos atributos tiveron na clasificación dun individuo concreto.
- En Strumbelj & Kononenko (2010) introdúcese un procedemento xeral para avaliar dita importancia.
- Este procedemento baséase no valor de Shapley para xogos cooperativos.



STRUMBELJ, E. & KONONENKO, I. (2010) An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11, 1–18.

Importancia de atributos na clasificación

Strumbelj & Kononenko (2010)

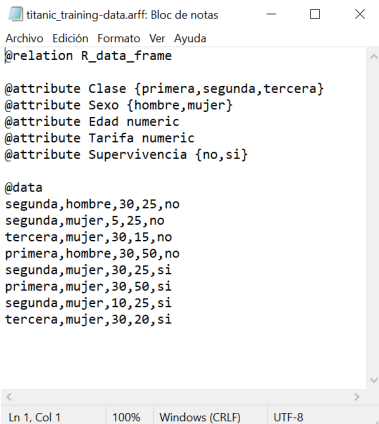
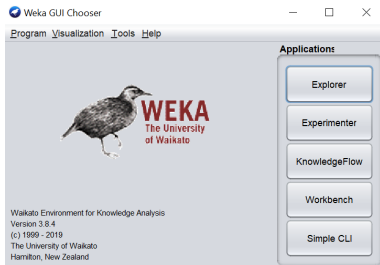
- 1 Consideramos un individuo $\mathbf{x} = (x_1, \dots, x_n)$ onde x_i é o valor do atributo i .
- 2 Dado un subconxunto de atributos, S , calcúlase a diferenza entre a predición cando só coñecemos eses valores do individuo, cuxos atributos pertencen a dito subconxunto, e a predición cando non se coñece ningún atributo.

$$\Delta(S) = \frac{1}{|\mathcal{A}_{N \setminus S}|} \sum_{y \in \mathcal{A}_{N \setminus S}} f_c(\tau(x, y, S)) - \frac{1}{|\mathcal{A}_N|} \sum_{y \in \mathcal{A}_N} f_c(y)$$
$$\tau(x, y, S) = (z_1, \dots, z_n) \text{ con } z_i = \begin{cases} x_i & \text{se } i \in S \\ y_i & \text{se } i \notin S \end{cases}$$

- 3 Calculamos o valor de Shapley do xogo anterior: cada coordenada representa a influencia dese atributo na clasificación.

Exemplo sinxelo: con Weka

Titanic



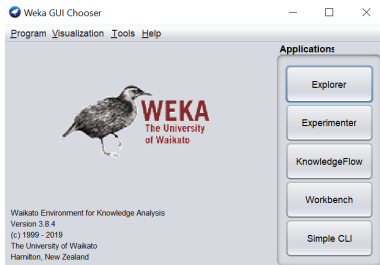
```
@relation R_data_frame

@attribute Clase {primera,segunda,tercera}
@attribute Sexo {hombre,mujer}
@attribute Edad numeric
@attribute Tarifa numeric
@attribute Supervivencia {no,si}

@data
segunda,hombre,30,25,no
segunda,mujer,5,25,no
tercera,mujer,30,15,no
primera,hombre,30,50,no
segunda,mujer,30,25,si
primera,mujer,30,50,si
segunda,mujer,10,25,si
tercera,mujer,30,20,si
```

Exemplo sinxelo: con Weka

Titanic



```
Archivo Edición Formato Ver Ayuda
@relation R_data_frame

@attribute Clase {primera,segunda,tercera}
@attribute Sexo {hombre,mujer}
@attribute Edad numeric
@attribute Tarifa numeric
@attribute Supervivencia {no,si}

@data
segunda,hombre,30,25,no
segunda,mujer,5,25,no
tercera,mujer,30,15,no
primera,hombre,30,50,no
segunda,mujer,30,25,si
primera,mujer,30,50,si
segunda,mujer,10,25,si
tercera,mujer,30,20,si
```

$x = (\text{primera}, \text{mujer}, 30, 50)$

Exemplo sinxelo: con Weka

Titanic

The screenshot shows the Weka Explorer interface. The 'Classifier' tab is active, displaying a 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' model. The 'Test options' section has 'Supplied test set' selected. The 'Classifier output' window shows the following results:

```
Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances      8          100 %
Incorrectly Classified Instances    0           0 %
Kappa statistic                     1
Mean absolute error                  0.2165
Root mean squared error              0.2444
Relative absolute error              43.3 %
Root relative squared error          48.8721 %
Total Number of Instances           8

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area
1,000      1,000    0,000    1,000     1,000    1,000     1,000    1,000
1,000      0,000    1,000     1,000    1,000     1,000    1,000    1,000
Weighted Avg.   1,000    0,000    1,000     1,000    1,000     1,000    1,000

=== Confusion Matrix ===

 a b  <-- classified as
 4 0 | a = no
 0 4 | b = si
```

Exemplo sinxelo: con Weka

Titanic

$$S \in \{\{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \\ \{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\}$$

```
titanic_coal-1.arff: Bloc de notas
Archivo Edición Formato Ver Ayuda
@relation R_data_frame

@attribute Clase {primera,segunda,tercera}
@attribute Sexo {hombre,mujer}
@attribute Edad numeric
@attribute Tarifa numeric
@attribute Supervivencia {no,si}

@data
primera,hombre,5,15,?
primera,hombre,5,20,?
primera,hombre,5,25,?
primera,hombre,5,50,?
primera,hombre,10,15,?
primera,hombre,10,20,?
primera,hombre,10,25,?
primera,hombre,10,50,?
primera,hombre,30,15,?
primera,hombre,30,20,?
primera,hombre,30,25,?
primera,hombre,30,50,?
primera,mujer,5,15,?
primera,mujer,5,20,?
primera,mujer,5,25,?
primera,mujer,5,50,?
primera,mujer,10,15,?
```

```
titanic_coal-24.arff: Bloc de notas
Archivo Edición Formato Ver Ayuda
@relation R_data_frame

@attribute Clase {primera,segunda,tercera}
@attribute Sexo {hombre,mujer}
@attribute Edad numeric
@attribute Tarifa numeric
@attribute Supervivencia {no,si}

@data
primera,mujer,5,50,?
primera,mujer,10,50,?
primera,mujer,30,50,?
segunda,mujer,5,50,?
segunda,mujer,10,50,?
segunda,mujer,30,50,?
tercera,mujer,5,50,?
tercera,mujer,10,50,?
tercera,mujer,30,50,?
```

Exemplo sinxelo: con Weka

Titanic

The screenshot displays the Weka Explorer application window. The main interface is in the 'Classifier' tab, showing the 'RandomForest' classifier selected. The 'Test options' section includes radio buttons for 'Use training set', 'Supplied test set' (selected), 'Cross-validation', and 'Percentage split'. The 'Classifier output' area shows a summary of the classification results, including 'Time taken to train', 'Summary', 'Correctly Classified Instances', 'Incorrectly Classified Instances', 'Kappa statistic', and 'Mean absolute error'. A 'Test Instances' dialog box is open, showing 'Relation: None' and 'Instances: None'. An 'Abrir' (Open) dialog box is also open, showing a list of files in the 'Konenenko' directory, including 'titanic_coal-1.arff' through 'titanic_coal-N.arff' and 'titanic_instancia.arff'. The 'Abrir' dialog box has 'Nombre de archivo:' set to 'titanic_coal-1.arff' and 'Archivos de tipo:' set to 'Arff data files (*.arff)'. The 'Abrir' button is highlighted.

Exemplo sinxelo: con Weka

Titanic

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Test options

Use training set

Supplied test set

Cross-validation Folds 10

Percentage split % 66

(Nom) Supervivencia

Classifier output

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	8	100	%
Incorrectly Classified Instances	0	0	%
Kappa statistic	1		
Mean absolute error	0.2165		
Root mean squared error	0.2444		
Relative absolute error	43.3	%	
Root relative squared error	48.8721	%	
Total Number of Instances	8		

=== Detailed Accuracy By Class ===

	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
0,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
0,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
0,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Result list (right-click for options)

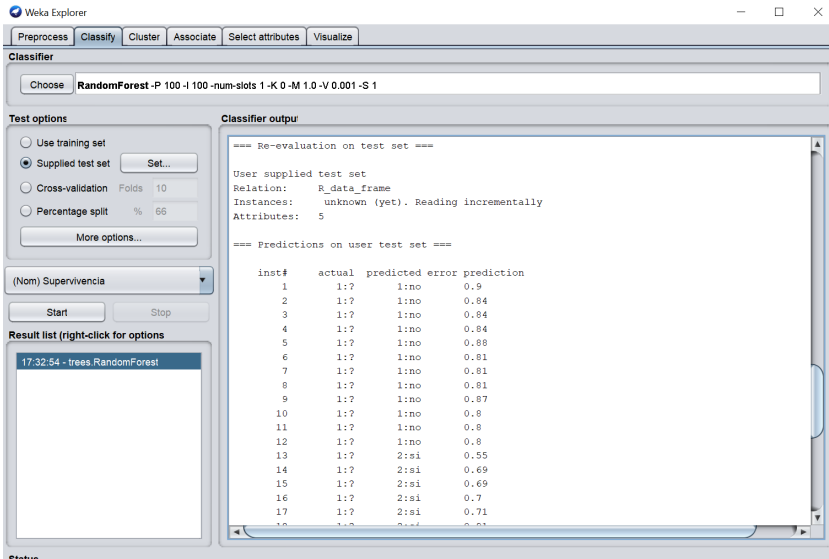
17:32:54 - trees.RandomForest

- View in main window
- View in separate window
- Save result buffer
- Delete result buffer(s)
- Load model
- Save model
- Re-evaluate model on current test set
- Re-apply this model's configuration
- Visualize classifier errors
- Visualize tree
- Visualize margin curve
- Visualize threshold curve

Status

Exemplo sinxelo: con Weka

Titanic



The screenshot shows the Weka Explorer interface. The 'Classifier' tab is active, displaying the 'RandomForest' classifier with the following options: `-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1`. Under 'Test options', 'Supplied test set' is selected. The 'Classifier output' window shows the results of a re-evaluation on a test set.

Test options:

- Use training set
- Supplied test set Set...
- Cross-validation Folds 10
- Percentage split % 66

More options...

Classifier output:

```
=== Re-evaluation on test set ===  
User supplied test set  
Relation:      R_data_frame  
Instances:     unknown (yet). Reading incrementally  
Attributes:    5  
  
=== Predictions on user test set ===
```

inst#	actual	predicted	error	prediction
1	1:?	1:no	0.9	
2	1:?	1:no	0.84	
3	1:?	1:no	0.84	
4	1:?	1:no	0.84	
5	1:?	1:no	0.88	
6	1:?	1:no	0.81	
7	1:?	1:no	0.81	
8	1:?	1:no	0.81	
9	1:?	1:no	0.87	
10	1:?	1:no	0.8	
11	1:?	1:no	0.8	
12	1:?	1:no	0.8	
13	1:?	2:si	0.55	
14	1:?	2:si	0.69	
15	1:?	2:si	0.69	
16	1:?	2:si	0.7	
17	1:?	2:si	0.71	
18	1:?	2:si	0.81	

Result list (right-click for options):

- 17:32:54 - trees.RandomForest

Exemplo sinxelo: con Weka

Titanic

The screenshot shows the Weka Explorer interface. The 'Classifier' tab is active, displaying a 'RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' model. The 'Test options' section is set to 'Supplied test set'. The 'Classifier output' window shows the results of a re-evaluation on a test set, including a table of predictions and error rates. Two specific instances are highlighted with red arrows and text: instance 5 (a man) and instance 13 (a woman).

Test options

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66
-

Classifier output

```
=== Re-evaluation on test set ===  
User supplied test set  
Relation: R_data_frame  
Instances: unknown (yet). Reading incrementally  
Attributes: 5  
  
=== Predictions on user test set ===
```

inst#	actual	predicted	error	prediction
1	1:?	1:no	0.9	→ (primera, hombre, 5, 15)
2	1:?	1:no	0.84	
3	1:?	1:no	0.84	
4	1:?	1:no	0.84	
5	1:?	1:no	0.88	
6	1:?	1:no	0.81	
7	1:?	1:no	0.81	
8	1:?	1:no	0.81	
9	1:?	1:no	0.87	
10	1:?	1:no	0.8	
11	1:?	1:no	0.8	
12	1:?	1:no	0.8	
13	1:?	2:si	0.55	→ (primera, mujer, 5, 15)
14	1:?	2:si	0.69	
15	1:?	2:si	0.69	
16	1:?	2:si	0.7	
17	1:?	2:si	0.71	
18	1:?	2:si	0.81	

Exemplo sinxelo: con R

Titanic

$$\Delta(S) = \frac{1}{|\mathcal{A}_{N \setminus S}|} \sum_{y \in \mathcal{A}_{N \setminus S}} f_c(\tau(x, y, S)) - \frac{1}{|\mathcal{A}_N|} \sum_{y \in \mathcal{A}_N} f_c(y)$$

```
> library(Rweka)
> RF <- make_Weka_classifier("weka/classifiers/trees/RandomForest")
> modelo_rf <- RF(muestra_weka$Supervivencia ~ ., data = muestra_weka)

> library(ggm)
> S <- powerset(1:dim(X)[2], nonempty=T, sort=T)

> predict(modelo_rf, newdata=<X_S[[i]]>, type = c("class"))
> v[[i]] <- sum(pred[[i]][,5]==classlabel)/dim(pred[[i]])[1] - factor_fixo

> library(GameTheoryAllocation)
> Shapley <- Shapley_value(unlist(v), game = "profit")
```

- Temos un conxunto de 10454 pacientes de Galicia infectados con COVID-19 dende o 6 de marzo de 2020 ata o 7 de maio de 2020.
- O obxectivo é estudar a influencia de varias características/atributos dos pacientes en tres variables resposta binarias de especial interese:
 - Necesidade de hospitalización.
 - Necesidade de ingreso en UCI.
 - Falecemento.
- Os atributos considerados son:
 - Idade: 0 (0-49 anos); 1 (50-64 anos); 2 (65-79 anos); 3 (80 anos en adiante).
 - Sexo: 0 (muller); 1 (home).
 - Patoloxías cardíacas: 0, 1, 2.
 - Patoloxías respiratorias: 0, 1, 2.
 - Patoloxías metabólicas: 0, 1, 2.
 - Patoloxías urinarias: 0, 1.

Consideramos o seguinte xogo:

$$v_x(S) = \frac{1}{|\mathcal{A}_N \setminus S|} \sum_{y \in \mathcal{A}_N \setminus S} f_c^P(\tau(x, y, S))$$

Consideramos o seguinte xogo:

$$v_x(S) = \frac{1}{|\mathcal{A}_N \setminus S|} \sum_{y \in \mathcal{A}_N \setminus S} f_c^p(\tau(x, y, S))$$

```
> predict(modelo_rf, newdata=<X_S[[i]]>, type = c("probability"))
```

Consideramos o seguinte xogo:

$$v_x(S) = \frac{1}{|\mathcal{A}_{N \setminus S}|} \sum_{y \in \mathcal{A}_{N \setminus S}} f_c^p(\tau(x, y, S))$$

```
> predict(modelo_rf, newdata=<X_S[[i]]>, type = c("probability"))
```

- 1 Para cada un dos atributos, j , e o seu valor, a_j , fixamos a submostra \mathcal{M}_{a_j} cos individuos que teñen esas características.
- 2 Calculamos o xogo v_x para cada individuo x de \mathcal{M}_{a_j} .
- 3 Calculamos o valor de Shapley do xogo v_x , $\phi(v_x)$.
- 4 Promediamos os valores de Shapley, obtendo a nosa medida de influencia $I_j^\phi = \frac{1}{|\mathcal{M}_{a_j}|} \sum_{x^i \in \mathcal{M}_{a_j}} \phi(v_{x^i})$.

$$T^\Phi := \sum_{k \in N} I_k^\Phi.$$

A cantidade T^Φ pertence a $[0, 1]$ e pódese interpretar como unha estimación da probabilidade de que a correspondente resposta dun individuo con atributo j igual a a_j sexa positiva.

Observación

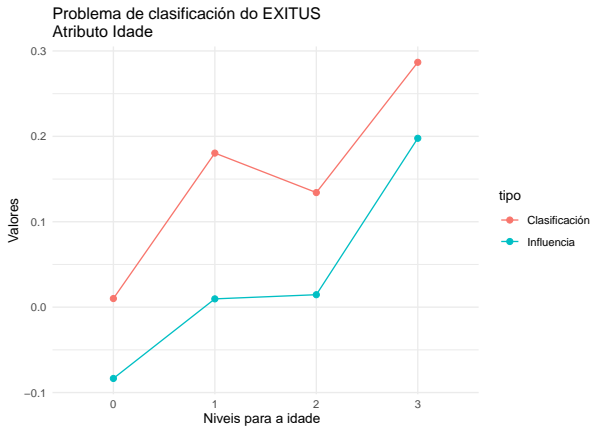
Nótese que I_j^Φ é a parte correspondente ao atributo j cando repartimos a cantidade T^Φ entre todos os atributos.

Deste xeito, a evolución dos números $\{I_j^\Phi\}$ e $\{T^\Phi\}$ é moi ilustrativa da influencia que os distintos valores de j teñen na resposta.

Por exemplo, se para un determinado valor observamos que ambos valores son próximos, é a vez que T^Φ é cercano a 1, podemos concluir que os individuos co atributo j igual a a_j teñen unha alta probabilidade de ser clasificados como positivos, e que iso débese principalmente ao atributo j .

COVID-19

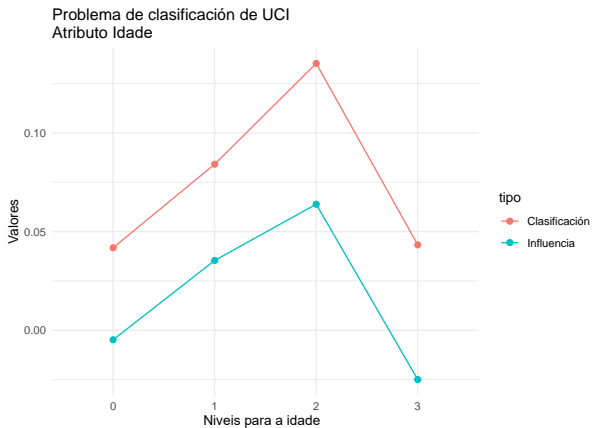
Idade - falecemento



```
> library(ggplot2)
```

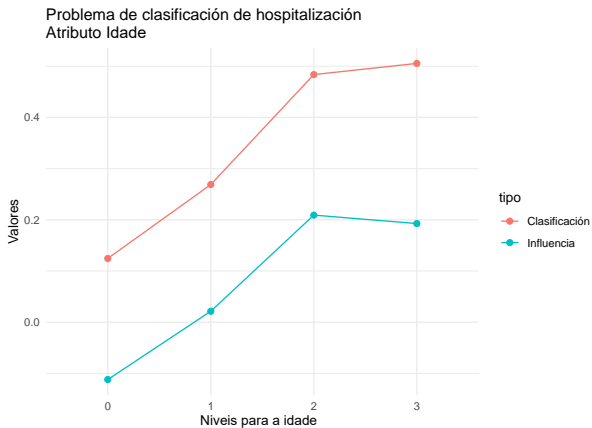
COVID-19

Idade - necesidade de ingreso en UCI



COVID-19

Idade - necesidade de hospitalización



O emprego de R na detección das características máis influentes na clasificación de pacientes infectados por COVID-19 en Galicia

VII Xornada de Usuarios de R en Galicia

Laura Davila Pena, Balbina Casas Méndez, Ignacio García Jurado

15 de outubro de 2020

