ANÁLISIS DE DATOS

Ricardo Aler Mur

R

CASOS PRÁCTICOS

CONCENTRACIÓN DE ALGAS

CONCENTRACIÓN DE ALGAS

- Las altas concentraciones de ciertas algas en los ríos tienen consecuencias sobre la fauna y la calidad del agua
- Para predecir la concentración de algas, se tomaron muestras mensuales en diferentes ríos de Europa durante un año
- Para cada muestra, se hacen diversas pruebás quimicas y se analiza que tipo de algas y su cantidad, hay presentes
- Las pruebas químicas son baratas, pero el análisis biológico es caro. Es interesante poder predecir la concentración de algas a partir de las pruebas químicas solamente

CONCENTRACIÓN DE ALGAS

- COIL 1999 international data analysis competition
- 200 muestras para entrenamiento (cada muestra es la media de muestras recogidas durante 3 meses consecutivos)
- 140 muestras para test
- 11 atributos de entrada:
  - 3 nominales: mes, tamaño y velocidad del río

CONCENTRACIÓN DE ALGAS

- Las 8 variables químicas:

- Maximum pH value
- Minimum value of O_2 (oxygen)
- Mean value of Cl (chloride)
- Mean value of NO_3^- (nitrates)
- Mean value of NH_4^+ (ammonium)
- Mean of PO_4^{3-} (orthophosphate)
- Mean of total PO_4 (phosphate)
- Mean of chlorophyll

CONCENTRACIÓN DE ALGAS

- Donde conseguir los datos:

http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR/datasets2.html

- Entrenamiento: Analysis.txt
  - Hay 7 columnas al final, una por cada concentración de tipo de alga (o sea, 7 clases numéricas)
- Test: Eval.txt
- Sols.txt: 7 columnas para los 7 distintos tipos de algas del conjunto de test
- Para evitar tener que teclear el código:
  - Buscar en google: data mining with R case studies
  - http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR/code2.html

CARGANDO LOS DATOS

```
algae <- read.table('Analysis.txt',
        header=F,
        dec='.',
        col.names=c('season','size','speed','mxPH','mnO2
','Cl', 'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',
'a5','a6','a7'),
        na.strings=c('XXXXXXX'))
```

CARGANDO LOS DATOS DE TEST

```
> algaeTest <- read.table('Eval.txt',
        header=F,
        dec='.',
        col.names=c('season','size','speed','mxPH','mnO2','Cl',
'NO3','NH4','oPO4','PO4','Chla'),
        na.strings=c('XXXXXXX'))

>sols = read.table('Sols.txt',
        header=F,
        dec='.',
        col.names=c('a1','a2','a3','a4', 'a5','a6','a7'),
        na.strings=c('XXXXXXX'))
```

CARGANDO LOS DATOS DE TEST

```
>algaeTest = cbind(algaeTest, sols)

>algaeTest
```

	season	size	speed	mxPH	mnO2	Cl	NO3	NH4	oPO4	PO4	Chla	a1	a2	a3	a4	a5	a6	a7
1	summer	small	medium	7.95	5.70	57.3330	2.460	273.333	295.667	380.000	NA	1.2	36.5	1.9	0.0	1.2	0.0	28.0
2	winter	small	medium	7.98	8.80	59.3330	7.392	286.667	33.333	138.000	7.100	1.2	0.0	0.0	0.0	23.2	46.4	0.0
3	summer	small	medium	8.00	7.20	80.0000	1.957	174.286	47.857	113.714	4.500	7.0	23.0	6.5	1.4	21.2	0.0	2.1
4	spring	small	high	8.35	8.40	68.0000	3.026	458.000	45.200	111.800	3.200	1.4	38.2	2.4	0.0	4.8	1.0	1.2
5	spring	small	medium	8.10	13.20	19.0000	0.000	130.000	6.000	40.000	2.000	3.9	55.4	8.4	0.0	0.0	0.0	0.0
6	summer	small	medium	8.37	12.10	12.8500	0.840	15.000	5.000	10.507	13.800	28.4	2.4	0.0	0.0	0.0	0.0	4.6
7	spring	small	high	7.31	9.90	6.0000	1.395	58.750	6.000	16.000	0.800	11.4	1.7	6.1	0.0	2.2	0.0	1.9
8	autumn	small	high	7.91	11.20	5.0000	1.383	6.000	24.333	30.000	32.000	29.7	2.0	2.0	0.0	2.7	3.7	0.0
9	summer	small	high	7.99	10.70	4.0000	1.368	117.000	17.250	44.750	0.800	74.3	1.7	1.4	0.0	1.0	0.0	0.0

CARGANDO LOS DATOS

```
> algae[1:5,]
```

	season	size	speed	mxPH	mnO2	Cl	NO3	NH4	oPO4	PO4	Chla	a1	a2	a3	a4	a5	a6	a7
1	winter	small	medium	8.00	9.8	60.800	6.238	578.000	105.000	170.000	50.0	0.0	0.0	0.0	0.0	34.2	8.3	0.0
2	spring	small	medium	8.35	8.0	57.750	1.288	370.000	428.750	558.750	1.3	1.4	7.6	4.8	1.9	6.7	0.0	2.1
3	autumn	small	medium	8.10	11.4	40.020	5.330	346.667	125.667	187.057	15.6	3.3	53.6	1.9	0.0	0.0	0.0	9.7
4	spring	small	medium	8.07	4.8	77.364	2.302	98.182	61.182	138.700	1.4	3.1	41.0	18.9	0.0	1.4	0.0	1.4
5	autumn	small	medium	8.06	9.0	55.350	10.416	233.700	58.222	97.580	10.5	9.2	2.9	7.5	0.0	7.5	4.1	1.0

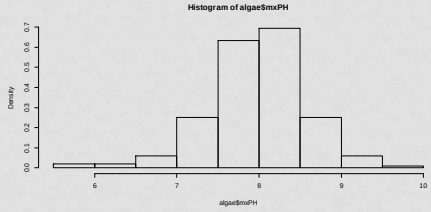
DESCRIPCIÓN INICIAL DE LOS DATOS

```
>summary(algae)
```

- Si la media y la mediana son muy distintas, la distribución está descentrada
- Comprobar si hay muchos Nas
- Hay mas muestras recogidas en invierno

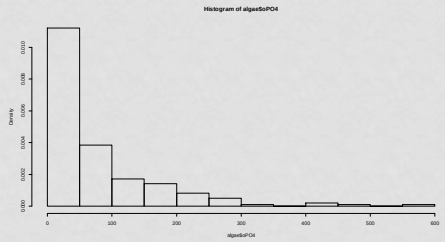
HISTOGRAMA MXPH

```
>hist(algae$mxPH, prob=T)
```



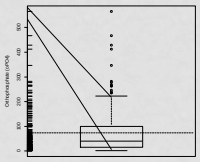
HISTOGRAMA OPO4

>hist(algae$oPO4, prob=T)



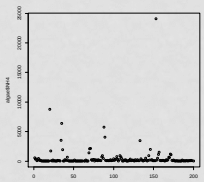
BOXPLOT OPO4

>boxplot(algae$oPO4,ylab='Orthophosphate (oPO4)')
>rug(jitter(algae$oPO4),side=2)
>abline(h=mean(algae$oPO4,na.rm=T),lty=2)



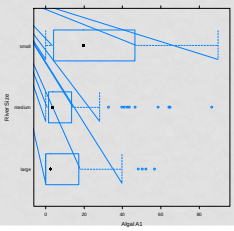
IDENTIFICACIÓN DE OUTLIERS

> plot(algae$NH4,xlab='')
> clicked.lines <- identify(algae$NH4)
> algae[clicked.lines,]



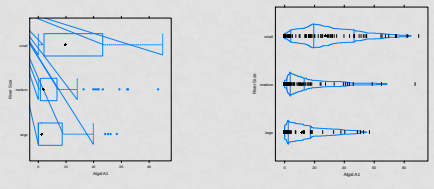
RELACIÓN ATRIBUTO-CLASE

>library(lattice) bwplot(size ~ a1, data=algae,ylab='River Size',xlab='Algal A1')
Vemos que las altas frecuencias de algas a1 están en los ríos pequeños



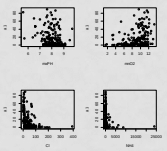
RELACIÓN ATRIBUTO-CLASE

>library(Hmisc)
>bwplot(size ~ a1, data=algae,panel=panel.bpplot, probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='River Size',xlab='Algal A1')



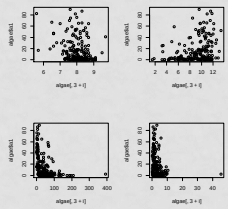
CORRELACIONES VARIABLE-CLASE

> split.screen(c(2,2))
> screen(1); with(algae, plot(mxPH,a1))
> screen(2); with(algae, plot(mnO2,a1))
> screen(3); with(algae, plot(Cl,a1))
> screen(4); with(algae, plot(NH4,a1))



CORRELACIONES VARIABLE-CLASE

> split.screen(c(2,2))
> for (i in 1:4) {screen(i); plot(algae[,3+i],algae$a1)}



ELIMINACIÓN DE NA

- Hay atributos con valores NA
  - Recordar la función is.na()
- Los datos que no tienen ningún NA se pueden obtener con
  - algae=algae[complete.cases(algae),]
  - algaeTest = algaeTest[complete.cases(algaeTest),]

CREAR UN MODELO DE REGRESIÓN LINEAL

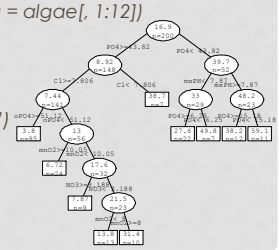
>lm.a1 <- lm(a1 ~ ., data = algae[, 1:12])

- *Si queremos usar menos variables:*
  - "a1 ~ mxPH + NH4"
- Detalles del modelo construido:
  - *summary(lm.a1)*

CREAR UN MODELO DE REGRESION TREES

> install.packages('rpart')
> library(rpart)
> rt.a1 <- rpart(a1 ~ ., data = algae[, 1:12])
> rt.a1
> summary(rt.a1)

> install.packages('DMwR')
> library(DMwR)
> prettyTree(rt.a1)

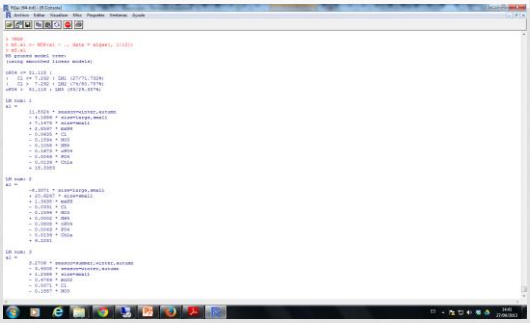


CREAR UN MODEL TREE (PARA REGRESIÓN)

- Usa el algoritmo M5' de Weka:

> install.packages('RWeka')
> library(RWeka)
> ?M5P
> m5.a1 <- M5P(a1 ~ ., data = algae[, 1:12])

M5.A1



EVALUACIÓN DE LOS MODELOS CON EL CONJUNTO DE TEST

- *Mediante error absoluto medio*

```
>lm.predictions.a1 <- predict(lm.a1, algaeTest)
>rt.predictions.a1 <- predict(rt.a1, algaeTest)
>m5.predictions.a1 <- predict(m5.a1, algaeTest)

>(mae.a1.lm <- mean(abs(lm.predictions.a1 - algaeTest$a1)))
>(mae.a1.rt <- mean(abs(rt.predictions.a1 - algaeTest$a1)))
>(mae.a1.m5 <- mean(abs(m5.predictions.a1 - algaeTest$a1)))
```

EVALUACIÓN DE LOS MODELOS CON EL CONJUNTO DE TEST

```
> (mae.a1.lm <- mean(abs(lm.predictions.a1 - algaeTest$a1)))
[1] 11.60885
> (mae.a1.rt <- mean(abs(rt.predictions.a1 - algaeTest$a1)))
[1] 10.83857
> (mae.a1.m5 <- mean(abs(m5.predictions.a1 - algaeTest$a1)))
[1] 9.798872
```