Aprendizaje Automático aplicado a la Paleotonlogía

Leonardo Moreno
Sebastián Vallejo

Facultad de Ciencias

Montevideo - 9 de Setiembre, 2014

El problema de las capas

Deseamos predecir la capa marron o verde. Contamos con una muestra de 547 de la capa marrón y 186 en capa verde.

```
df = read.csv("data_v6c.csv",header = TRUE, sep = ";")  #
    read csv file
head(df)
  marr n..verde estados rodado trampling fractura
1             1       1      1         0        1
2             1       1      0         0        0
3             1       1      0         0        0
4             1       1      0         0        0
  Grupos.de.Voorhies agrupaci n.anat mica capa Procesar
1                  3                    5    M        1
2                  3                    5    M        1
3                  3                    5    M        1
4                  1                    4    M        1
```

El problema de las capas (cont)

Vamos a predecir la capa con estado, rodado, trampling y fractura:

```
df$capa=as.factor(df$capa)
dp=df[df$Procesar == 1,]
dp1= data.frame(dp$estados,dp$rodado,dp$trampling,
    dp$fractura,dp$capa)
summary(dp1)
```

Modelo SVM: Support Vector Machines

```
modelo1 <- svm(dp.capa ~ . , data=dp1,cost = 10, gamma =
    0.1,cross=10)
summary(modelo1)

10-fold cross-validation on training data:
Total Accuracy: 79.94543

pred1 <- predict(modelo1, dp1)
(acc <- table(pred1, dp1$dp.capa))

pred1   M   V
    M 516 107
    V  31  79
```

Modelo Regresión Logística

Utilizamos la función glm con ciertos parámetros:

```
modelologi = glm(dp.capa ~ ., family=binomial(logit), data=
    dp1)
summary(modelologi)

predlogi <- predict(modelologi, dp1)
resulogi=dp1
resulogi$pred1=predlogi
```

Modelo Regresión Logística (cont)

Modelo obtenido

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.9266     0.3051 -12.869  < 2e-16 ***
dp.estados     1.2671     0.1898   6.674 2.48e-11 ***
dp.rodado     -0.1957     0.1252  -1.563    0.118
dp.trampling   0.4641     0.1017   4.563 5.04e-06 ***
dp.fractura    1.1340     0.2316   4.896 9.76e-07 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *
    0.05    .    0.1          1
```

Modelo Regresión Logística (cont)

Matriz de confusión y tasa de error :Predice mejor en la capa marrón que en la verde.

```
prob=predict(modelologi,type=c("response"),dp1)
prob=predict(modelologi,type=c("response"),dp1)
dp1$prob=prob
confusion<-table(prob>0.5,dp1$dp.capa)
> confusion
          M   V
  FALSE 515 112
  TRUE   32  74

errorrate<-sum(diag(confusion))/sum(confusion)
errorrate
[1] 0.8035471
```

Conclusión

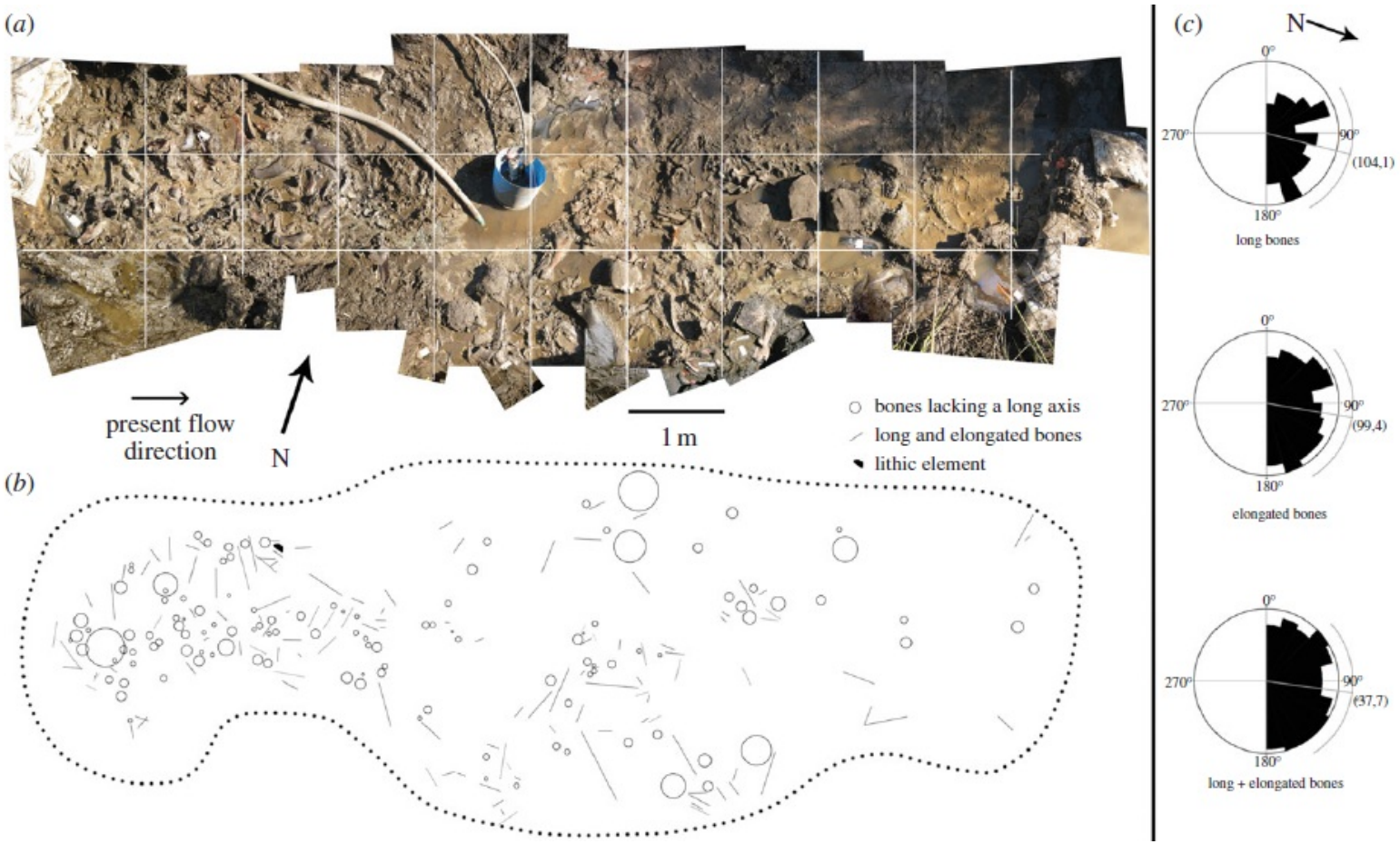
Si bien se obtiene una tase de error baja, con ambos modelos, predice mejor en la capa marrón que en la verde.
En particular la capa verde la predice MUY MAL.

Tareas en proceso:

- Tratamiento de datos no balanceados.
- Utilizar más columnas para la predicción
- Utilizar árboles de decisión

Problema de la orientación de los huesos

Contamos con una lista que contiene los valores de los ángulos de elementos asociados al yacimiento de Arroyo Vizcaíno.

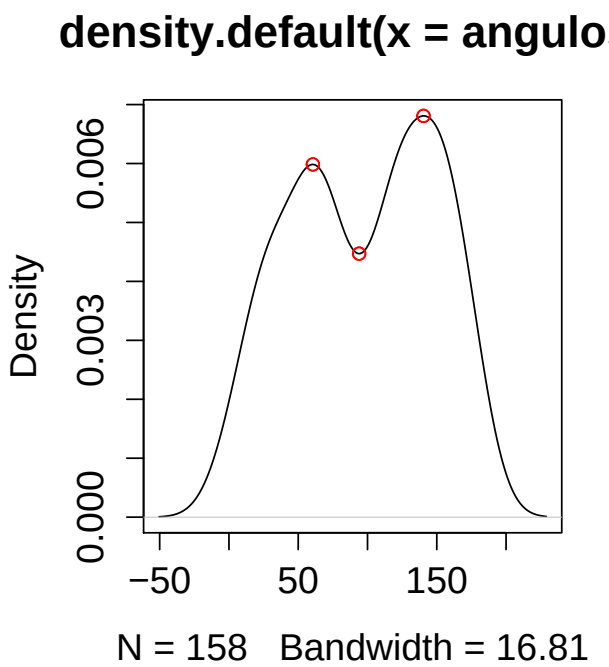


Problema de la orientación de los huesos (cont)

Utilizaremos una técnica de aprendizaje no supervisado para formar grupos (clústering). Tenemos una sola variable lo que hace especial al problema.

Es no supervisado porque nuestra muestra tiene datos, pero no sabemos a que grupo pertenece cada elemento.

Estimación no paramétrica de la densidad



Observamos que la densidad es bimodal. Vamos a probar una hipótesis en que los huesos no están distribuídos en forma uniforme.

Selección de un modelo paramétrico

Lo aproximaremos con una mezcla de densidades normales. Vamos a estudiar cuantos grupos es probable que contengan:

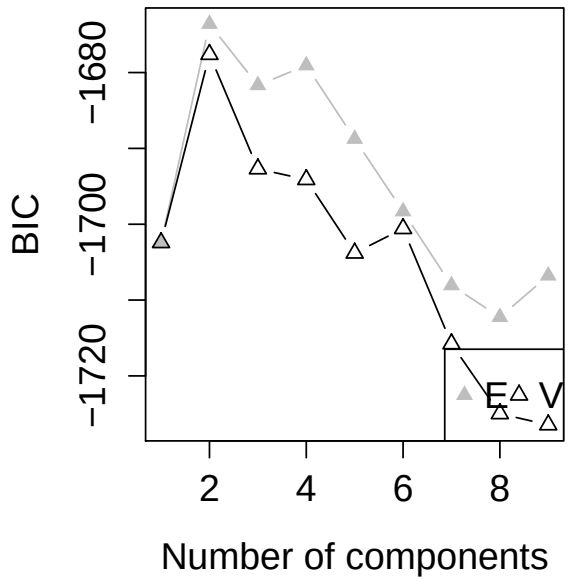
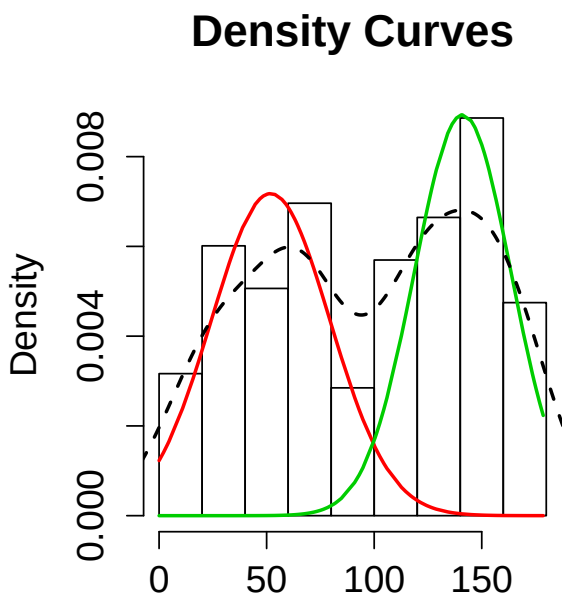


Figure: Elegir modelo: Cantidad de grupos .

Resultados obtenidos

Obtenemos una mezcla de dos normales con las siguientes características (diferencia de media 90 aprox):

```
$ lambda     : num [1:2] 0.496 0.504
$ mu         : num [1:2] 51.9 141.1
$ sigma      : num [1:2] 27.6 22.5
```



Código en lenguaje R y comentarios

```
#cargo datos
datahueso2 = read.csv("datasauce2.csv")  # read csv file
head(datahueso2)
angulos=datahueso2$Angulo180

#una aproximacion por kernels de la densidad
plot(density(angulos))
```

Código en lenguaje R y comentarios

```
#ahora vamos a probar la libreria mixtools, tambien por
    mezcla de normales
#mixtools paper: http://www.jstatsoft.org/v32/i06/paper
library(mixtools)

mezcla = normalmixEM(angulos)
plot(mezcla,which=2)
lines(density(angulos), lty=2, lwd=2)

pi<-mezcla$lambda[1]
mu1<-mezcla$mu[1]
mu2<-mezcla$mu[2]
sigma1<-mezcla$sigma[1]
sigma2<-mezcla$sigma[2]
```