

Seminario1

November 22, 2016

Determinación de los parámetros cosmológicos a partir de métodos de aprendizaje automático

0.0.1 Martín de los Rios, Mariano Dominguez

```
In [11]: #Funciones necesarias para correr todo
#Feature selection por bineado
#{{{
feature.select<-function(x,y,n.feats,x.min,x.max){
  #xbins=(max(x)-min(x))/(n.feats)
  xbin=(x.max-x.min)/(n.feats)

  y.feats<-1:n.feats
  x.feats<-1:n.feats

  dx=0.00001 #Es para evitar que el maximo se quede afuera por redondeo de
  for(i in 1:n.feats){
    #x.feats[i]=min(x)+xbin*(2*(i-1)+1)/2
    x.feats[i]=x.min+xbin*(2*(i-1)+1)/2
    xmin=x.feats[i]-xbin/2
    xmax=x.feats[i]+xbin/2
    y.aux<-y[which(x>=(xmin-dx) & x<=(xmax+dx))]
    if(length(y.aux) > 1){
      y.feats[i]=mean(y.aux)
    } else if (length(y.aux) == 1){
      y.feats[i]= y.aux[1]
    } else if (length(y.aux) == 0){
      y.feats[i]= -9999
    }
  }
}

for(i in 1:n.feats){
  if(y.feats[i] == -9999){
    ysup=-9999
    yinf=-9999
    for(j in 1:(n.feats-1)){
      if(ysup<0 & (i+j) <= n.feats){
        if(y.feats[i+j] > 0){ysup=y.feats[i+j]}
      }
    }
  }
}
```

```

    }
    if(yinf<0 & (i-j) > 0){
      if(y.feats[i-j] > 0){yinf=y.feats[i-j]}
    }
  }
  y.feats[i]=(ysup+yinf)/2
}
}

y.feats[which(y.feats < 0)]=0

mat<-data.frame(x.feats,y.feats)
return(mat)
}
#}}}
#Feature selection por binedo (cmb)
#{{{
feature.select.cmb<-function(x,y,n.feats){
  xbin=(max(x)-min(x))/(n.feats)

  y.feats<-1:n.feats
  x.feats<-1:n.feats

  for(i in 1:n.feats){
    x.feats[i]=min(x)+xbin*(2*(i-1)+1)/2
    xmin=x.feats[i]-xbin/2
    xmax=x.feats[i]+xbin/2
    y.feats[i]=mean(y[which(x>=xmin & x<=xmax)])
  }

  mat<-data.frame(x.feats,y.feats)
}
#}}}

#PCA_mat
#{{{
PCA_mat<-function(dat){
  mat_cov<-cov(dat)
  eig<-eigen(mat_cov)
  mat_cb<-solve(eig$eigenvectors)

  return(mat_cb)
}
#}}}

#validacion
#{{{
validacion<-function(datos,datos_aux){

```

```

    fl=0
    datos_aux<-subset(datos_aux,datos_aux$y.feats > 0)
    for(i in 1:length(datos_aux$y.feats)){
      if((datos_aux$y.feats[i] < min(datos[,4+i]) | (datos_aux$y.feats[i] > ma

    }

    return(fl)
  }
#}}}
#validacion1
#{{{
validacion1<-function(datos,datos_aux){

  flags<-1:length(datos_aux$y.feats)
  flags[]=0
  for(i in 1:length(datos_aux$y.feats)){
    if((datos_aux$y.feats[i] < min(datos[,4+i]) | (datos_aux$y.feats[i] > ma

      flags[i]=1
    }
  }

  return(flags)
}
#}}}
#rename
#{{{
rename<-function(trainset){

  name1<-colnames(trainset)[1]
  trainset_aux<-trainset[, -1]

  name<-1:(length(trainset)-1)
  for(i in 1:length(name)){
    name[i]<-paste('X',toString(i), sep='')
  }

  name<-c(name1, name)
  colnames(trainset)<-name
  return(trainset)
}
#}}}
#rename1
#{{{
rename1<-function(trainset){

  trainset_aux<-trainset

```

```

name<-1:length(trainset_aux)
for(i in 1:length(trainset_aux)){
  name[i]<-paste('X',toString(i),sep='')
}

colnames(trainset)<-name
return(trainset)
}
#}}}}
library('latex2exp')
library('caret')

```

Loading required package: lattice

Loading required package: ggplot2

Attaching package: 'ggplot2'

The following object is masked from 'package:randomForest':

margin

0.1 Técnicas de Aprendizaje automático

- Aprendizaje Supervisado
- Aprendizaje no supervisado

0.1.1 Aprendizaje Supervisado

Random Forest

—
—

Support Vector Machine

Redes Neuronales

Ejemplo simple de aprendizaje supervisado

In [8]: `summary(iris)`

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300
Median :5.800	Median :3.000	Median :4.350	Median :1.300
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199

```

3rd Qu.:6.400    3rd Qu.:3.300    3rd Qu.:5.100    3rd Qu.:1.800
Max.      :7.900    Max.      :4.400    Max.      :6.900    Max.      :2.500
Species
setosa      :50
versicolor:50
virginica   :50

```

```

In [9]: #Cargamos la libreria
library('randomForest')

#Separamos en set de entrenamiento y testeo
ind<-sample(x=1:length(iris$Species),replace=F,size = 120)
set_entrenamiento<-iris[ind,]
set_testeo<-iris[-ind,]

#Entrenamos un modelo de aprendizaje automático
model<-randomForest(Species~., data=set_entrenamiento)

#Predecimos la especie para el set de testeo
especie_predicha<-predict(model,newdata = set_testeo)

```

```

In [13]: confusionMatrix(data = especie_predicha,reference = set_testeo$Species)$ta

```

	Reference		
Prediction	setosa	versicolor	virginica
setosa	12	0	0
versicolor	0	8	1
virginica	0	1	8

0.1.2 Aprendizaje No Supervisado

Mixtura de Gaussianas

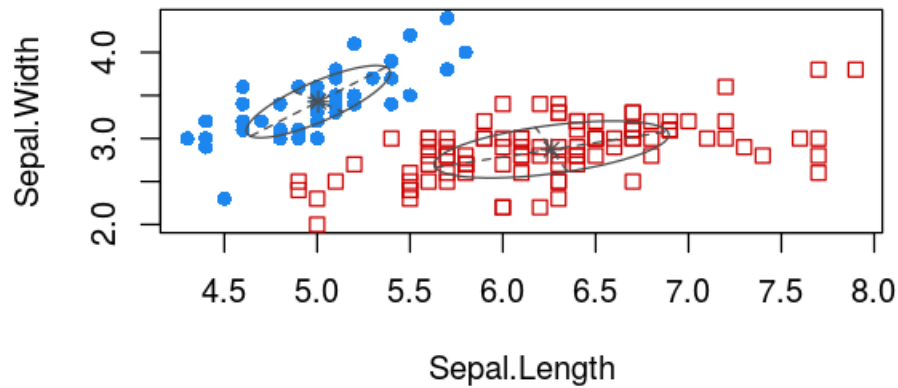
Análisis de Componentes Principales

Ejemplo simple de aprendizáje no supervisado

```

In [28]: #Cargo la libreria
library('mclust')
#Entreno el modelo
Mclust(data = iris[,-5],G = 1:4)->model
#Visualizo los resultados
options(repr.plot.width=5, repr.plot.height=3)
plot.Mclust(model,what = 'classification',dimens = c(1,2),main='')

```



0.1.3 El aprendizaje automático en la Astronomía.

- Ann-z: Estimación de redshift a partir de datos fotométricos utilizando redes neuronales.
- Técnicas de aprendizaje automático aplicadas a simulaciones cosmológicas para la construcción de catálogos simulados (1303.1055, 1510.07659, 1510.06402).
- Estimación de masas de cúmulos de galaxias utilizando SDM (1509.05409).
- Clasificación morfológica de galaxias utilizando aprendizaje automático supervisado (1005.0390).

PAPER BRUNO Y JUAN

MeSsI Clasificación de cúmulos de galaxias en proceso de fusión utilizando random forest.

0.1.4 Sobre los parámetros cosmológicos y como medirlos.

Supernovas JLA (1512.07869) Fondo Cósmico de Microondas (1502.01589)

Estimación de los parámetros cosmológicos (1401.4064)

||| |:—|—:| ||| *Distribución Angular de los parámetros cosmológicos medidos con supernovas (1512.07869).*

Problemas

- Estimar la distribución angular de los parámetros cosmológicos utilizando datos de supernovas -> Muchos pixeles con pocos datos

- La estimación de los parámetros cosmológicos a través de datos del CMB requiere la utilización de cadenas de Markov Monte-Carlo -> Mucho Tiempo de Cómputo
- Para estimar los intervalos de confianza, en los diferentes planos, utilizando varios conjuntos de datos se necesita conocer la matriz de covarianza entre los conjuntos de datos -> **La estimación de dicha matriz de covarianza es un tema de estudio actual**

0.1.5 Estimación de la distribución angular de los parámetros cosmológicos mediante métodos de aprendizaje automático aplicados a datos del CMB y de supernovas.

Soluciones

- Utilizando los datos del CMB hay una mayor cantidad de datos en todos los pixeles.
- Una vez que se construyó el conjunto de entrenamiento y se entrenó el algoritmo de aprendizaje automático, la estimación de los parámetros cosmológicos no requiere mucho tiempo de cómputo.
- No es necesario conocer la matriz de covarianza entre los diferentes conjuntos de datos.

Espectros de potencia generados por CAMB

Mixtura de Gaussianas

Random Forest $\Delta\Omega_m * h^2 = 0.002$ $\Delta\Omega_k = 0.007$ $\Delta\Omega_b * h^2 = 0.005$ $\Delta H_0 = 2.8$

Support Vector Machine $\Delta\Omega_m * h^2 = 0.0009$ $\Delta\Omega_k = 0.006$ $\Delta\Omega_b * h^2 = 0.0002$ $\Delta H_0 = 2.8$

Conjunto de supernovas simuladas

Random Forest $\Delta\Omega_m * h^2 = 0.01$ $\Delta\Omega_k = 0.05$ $\Delta\Omega_b * h^2 = 0.003$ $\Delta H_0 = 1.4$

Support Vector Machine $\Delta\Omega_m * h^2 = 0.01$ $\Delta\Omega_k = 0.06$ $\Delta\Omega_b * h^2 = 0.003$ $\Delta H_0 = 2.3$

Random Forest $\Delta\Omega_m * h^2 = 0.003$ $\Delta\Omega_k = 0.005$ $\Delta\Omega_b * h^2 = 0.0007$ $\Delta H_0 = 0.5$

Support Vector Machine $\Delta\Omega_m * h^2 = 0.001$ $\Delta\Omega_k = 0.004$ $\Delta\Omega_b * h^2 = 0.0002$ $\Delta H_0 = 0.27$

0.1.6 Distribución angular de los parámetros cosmológicos

- Planck + Sn tipo Ia (jlca)
- Support Vector Machine
- Healpix

0.2 Conclusiones y Trabajo futuro

- Desarrollamos un método de aprendizaje automático que permite calcular los parámetros cosmológicos de una manera mucho más eficiente que la tradicional.
- Este método tiene la ventaja de poder utilizar a la vez los datos de supernovas y del cmb.
- Gracias a la velocidad de dicho método, pudimos calcular los parámetros cosmológicos en diferentes sectores del cielo.