Mineria de Datos

Clasificacion Supervisada: Regresion Logistica

Dr. Edgar Acuna
Departmento de Matematicas

Universidad de Puerto Rico- Mayaguez

academic.uprm.edu/eacuna

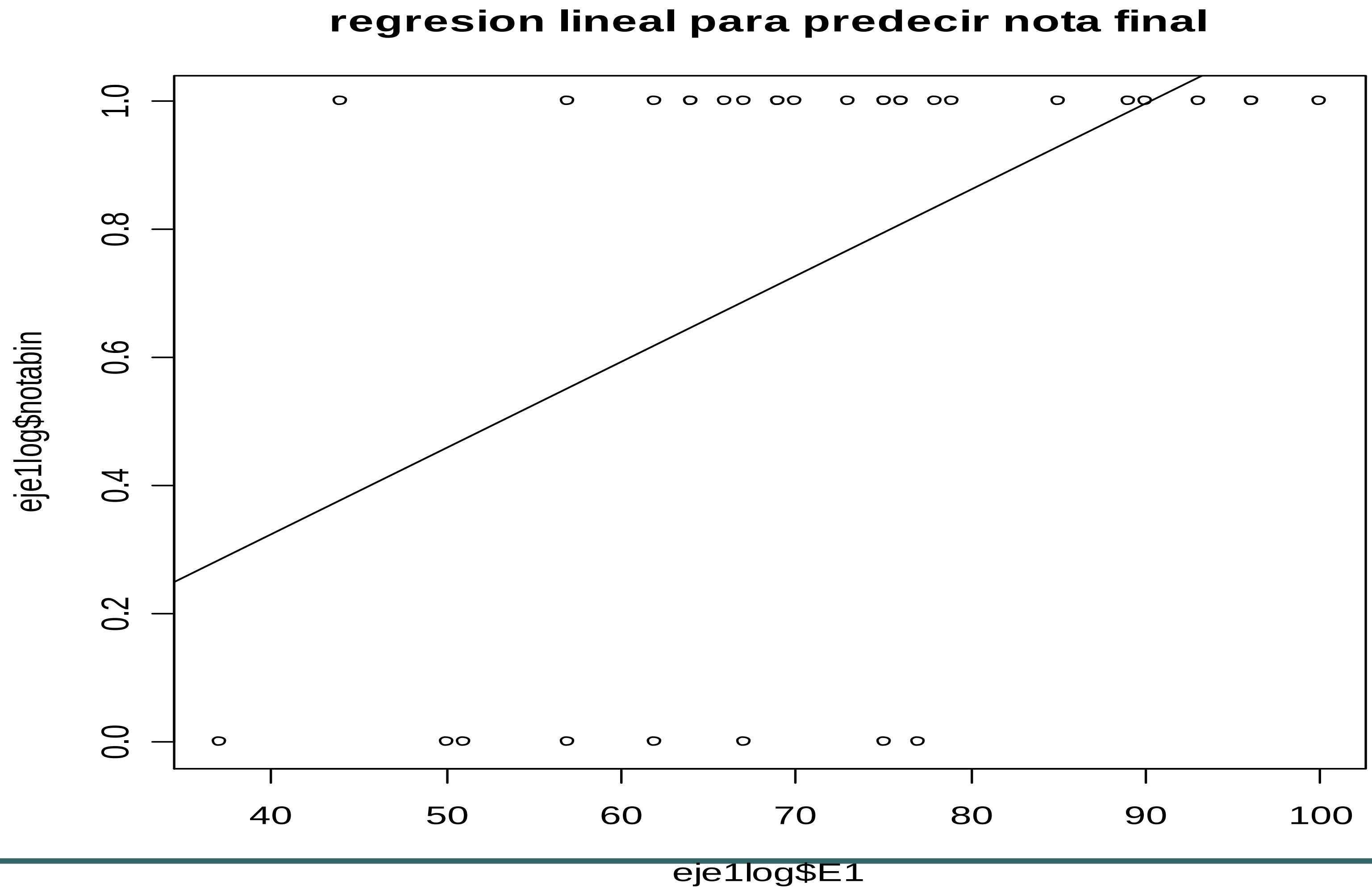
Consideraremos por ahora que solo tenemos dos clases. O sea, que nuestro conjunto de datos consiste de una muestra de tamaño $n=n_1+n_2$, n_1 observaciones son de la clase C_1 y n_2 son de la clase C_2.

Para cada observación $\mathbf{x}_j$ se introduce una variable binaria *Y* que vale 1 si ella es de la clase C_1 y vale 0 si la observacion pertenece a la clase C_2 .

La variable Y tiene una *probabilidad a priori* π_1 de que Y sea igual 1.

EJEMPLO DE PREDICCION DE LA NOTA FINAL USANDO SOLO EX1

```
> > eje1log[,c(1,4)]
   E1 notabin
1   96      1
2   96      1
3  100      1
4   93      1
5   90      1
6   75      1
7   75      1
8   64      1
9   89      1
10  57      1
11  70      1
12  44      1
13 …………
```

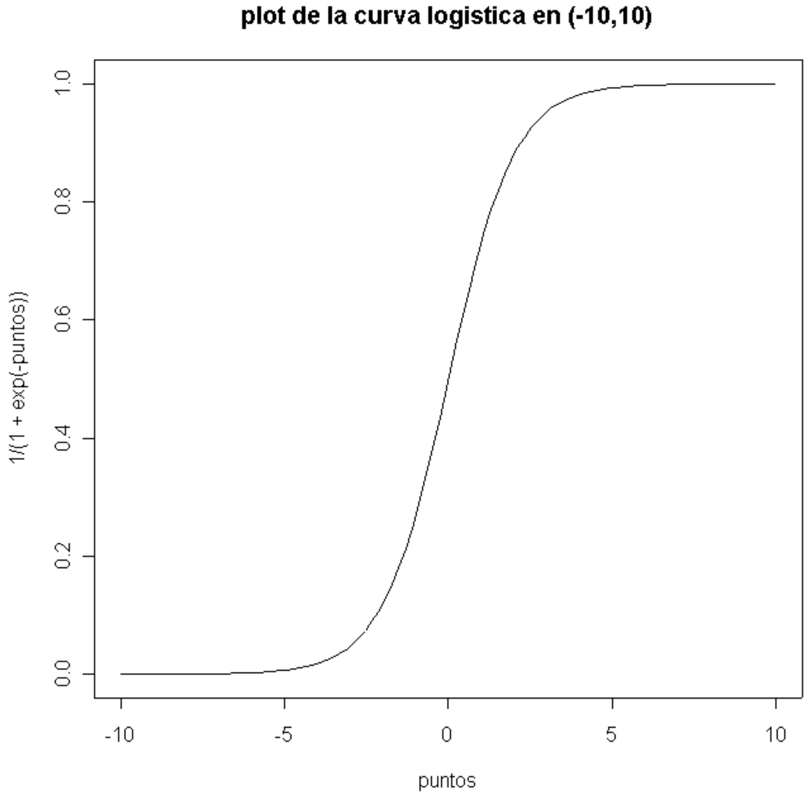


regresion lineal para predecir nota final
eje1log$notabin
0.0
0.2
0.4
0.6
0.8
1.0
40
50
60
70
80
90
100
eje1log$E1

La Curva Logística

Funcion de distribución logística

$$F(x) = \frac{1}{(1+e^{-x})}$$



En el modelo logístico (Berson, 1944) se asume que

$$\log\left(\frac{f(x/C_1)}{f(x/C_2)}\right) = \alpha + \beta' x$$

Aqui **x** es un vector p-dimensional de variables predictoras y β es un vector de p parámetros. Entre las distribuciones que satisfacen esta suposicion esta la Normal multivariada con igual matriz de covarianza por clase.

Por otro lado sea p=P(Y=1/x) la probabilidad a posteriori de que y sea igual a 1, entonces se puede notar que:

$$\frac{p}{1-p} = \frac{\dfrac{\pi_1 f(x/C_1)}{f(x)}}{\dfrac{\pi_2 f(x/C_2)}{f(x)}} = \frac{\pi_1 f(x/C_1)}{\pi_2 f(x/C_2)}$$

donde $$\frac{p}{1-p}$$

es llamado la razón de apuestas (***odds ratio***). Tomando logaritmos en ambos lados se obtiene

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) + \log\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}$$

Assumiendo distribucion normal multivariada de las predictoras se obtiene.

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) + (\mu_1 - \mu_2)'\Sigma^{-1}[x - (1/2)(\mu_1 + \mu_2)]$$

que puede ser escrita de la forma

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta'\mathbf{x}$$

o de la forma

$$p = \frac{\exp(\alpha + \beta'\mathbf{x})}{1+\exp(\alpha + \beta'\mathbf{x})}$$

esta ecuación es llamada regresión logística y

$$\log\left(\frac{p}{1-p}\right)$$

es llamado la transformacion ***logit***.

Los estimados $\tilde{\alpha}$ y $\tilde{\beta}$ son aquellos que maximizan la función de maxima versomilitud (o su logaritmo) correspondiente al modelo. Para una muestra de tamano 1 esta dado por

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(X_1 = u_1 \cdots X_m = u_m \mid Y = v)$$

Para una muestra de tamano n esto es equivalente a maximizar

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n_1} \frac{\exp(\alpha + \mathbf{x_i}'\boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{x_i}'\boldsymbol{\beta})} \cdot \prod_{j=n_1+1}^{n} \frac{1}{1 + \exp(\alpha + \mathbf{x_j}'\boldsymbol{\beta})}$$

En la solucion de este proceso de maximizacion se aplican métodos iterativos tales como Newton-Raphson o minimos cuadrados reponderados iterativos. Sin embargo algunas veces hay problemas de convergencia, generado cuando hay separabilidad ente las clases.

Regresion logistica para clasificacion

Para efectos de clasificación la manera más facil de discriminar es considerar que si p>0.5 entonces la obervación pertenece a la clase que uno está interesado. Pero algunas veces esto puede resultar injusto sobre todo si se conoce si una de las clases es menos frecuente que la otra.

Métodos alternos son:

i) Plotear en la misma grafica, el porcentaje de observaciones que poseen el evento (o sean que pertenecen al grupo 1) y que han sido correctamente clasificadas (Sensitividad) versus distintos niveles de probabilidad y el porcentaje de observaciones de la otra clase que han sido correctamente clasificadas (especifidad) versus los mismos niveles de probabilidad anteriormente usados. La probabilidad que se usará para clasificar las observaciones se obtienen intersectando las dos curvas.

ii)Usar la curva ROC (receiver operating characteristic curva). En este caso se grafica la sensitividad versus (1-especifidad)100%, y se coge como el p ideal aquel que está más cerca a la esquina superior izquierda, o sea al punto (0,100).

En **R** para hacer discriminación logística con dos clases se usa la función **glm** (modelo lineal general) con la opción **family=binomial,** Tambien se puede usar la funcion **lrm** de la librería Design.

Example: Bupa[1]

```
bupalog=glm(V7~.,data=bupa1,family=binomial)
phat=bupalog$fit
b=as.numeric(names(phat[phat>=.5]))
nobs=345
clases=rep(0,nobs)
clases[b]=1
mean(clases!=bupa1[,7])
[1] 0.2956522
```

Haciendo la clasificacion usando la sensitividad y la especificidad

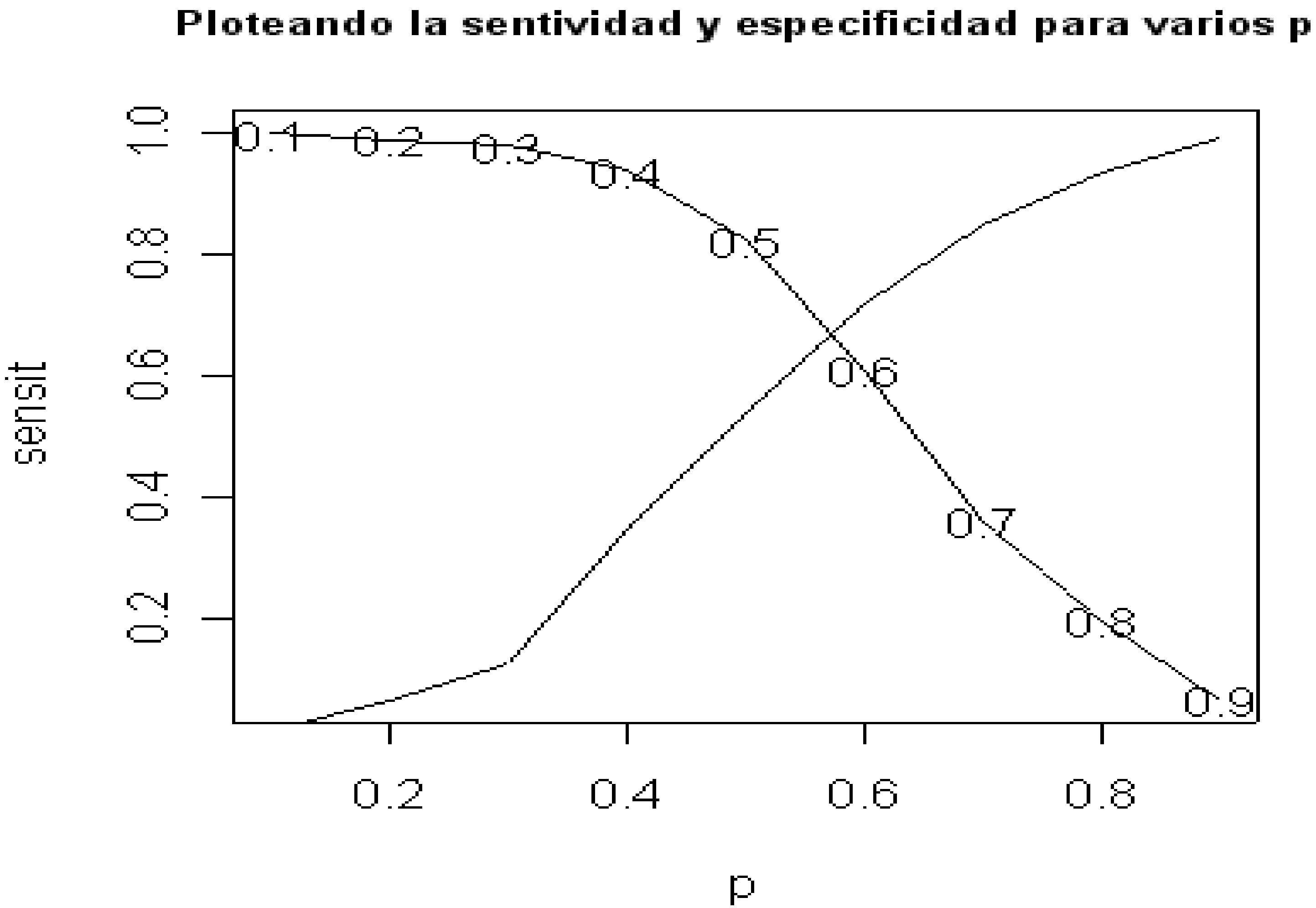
```
 p=seq(.1,.9,length=9)
sensit=rep(0,9)
especif=rep(0,9)
```

Example: Bupa[2]

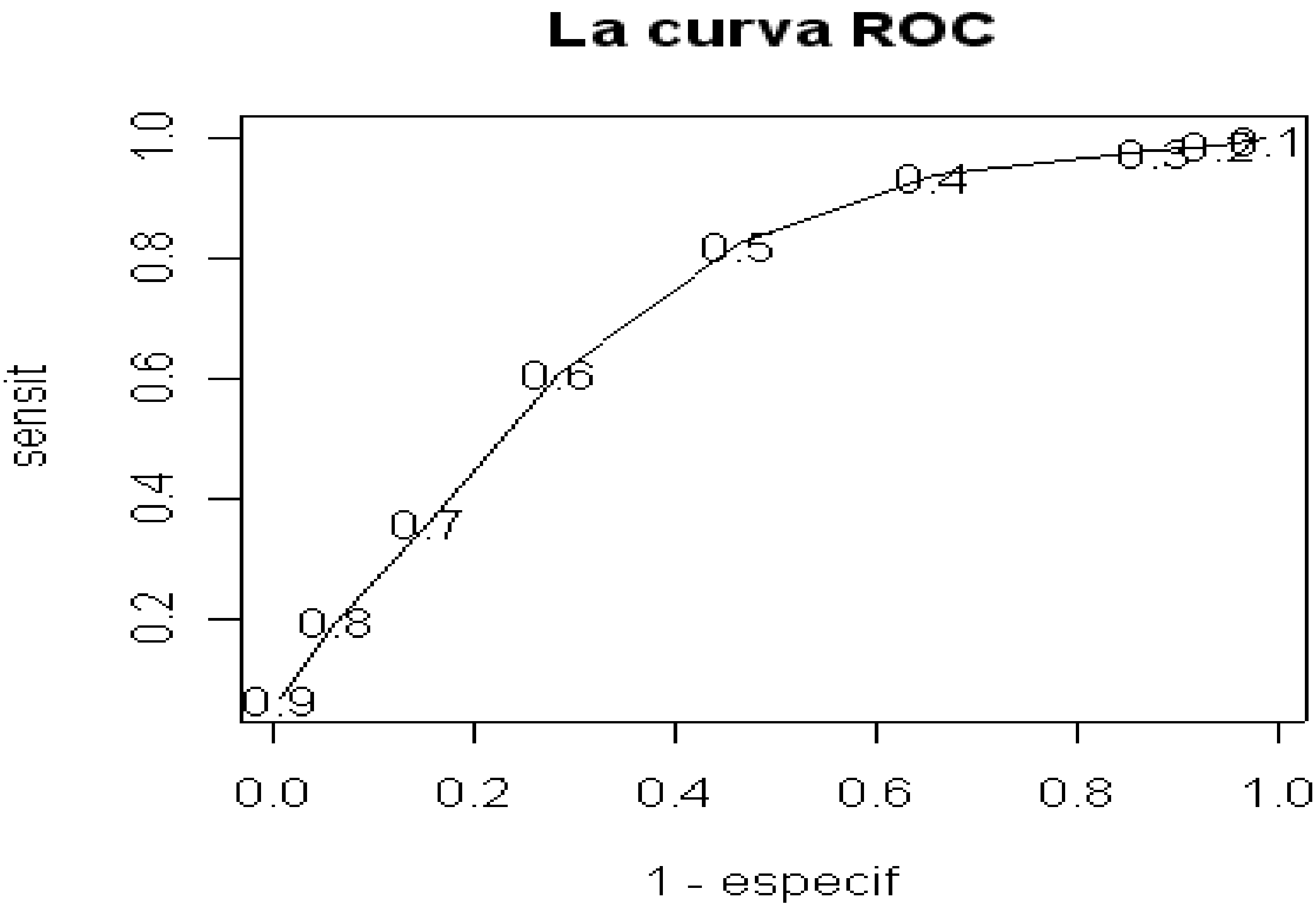
```
for(j in 1:9) {
clases1=rep(0,nobs)
for(i in 1:nobs)
{if(phat[i]>=p[j]){clases1[i]=1} }
 tempo=cbind(bupa1[,7],clases1)
positivo=tempo[tempo[,1]==1,]
 negativo=tempo[tempo[,1]==0,]
 sensit[j]=mean(positivo[,1]==positivo[,2])
 especif[j]=mean(negativo[,1]==negativo[,2]) }
 tabla=cbind(p,sensit,especif)
cat("Sensitividad y especifidad para varios valores de p\n")
```

Example: Bupa[3]

```
> print(tabla)
       p sensit    especif
 [1,] 0.1  1.000 0.01379310
 [2,] 0.2  0.990 0.06206897
 [3,] 0.3  0.980 0.12413793
 [4,] 0.4  0.940 0.34482759
 [5,] 0.5  0.825 0.53793103
 [6,] 0.6  0.610 0.71724138
 [7,] 0.7  0.360 0.84827586
 [8,] 0.8  0.195 0.93793103
 [9,] 0.9  0.065 0.99310345
plot(p,sensit,type="l");lines(p,especif); text(p,sensit,labels=p);
    title("Ploteando la sentividad y especificidad para varios p")
```



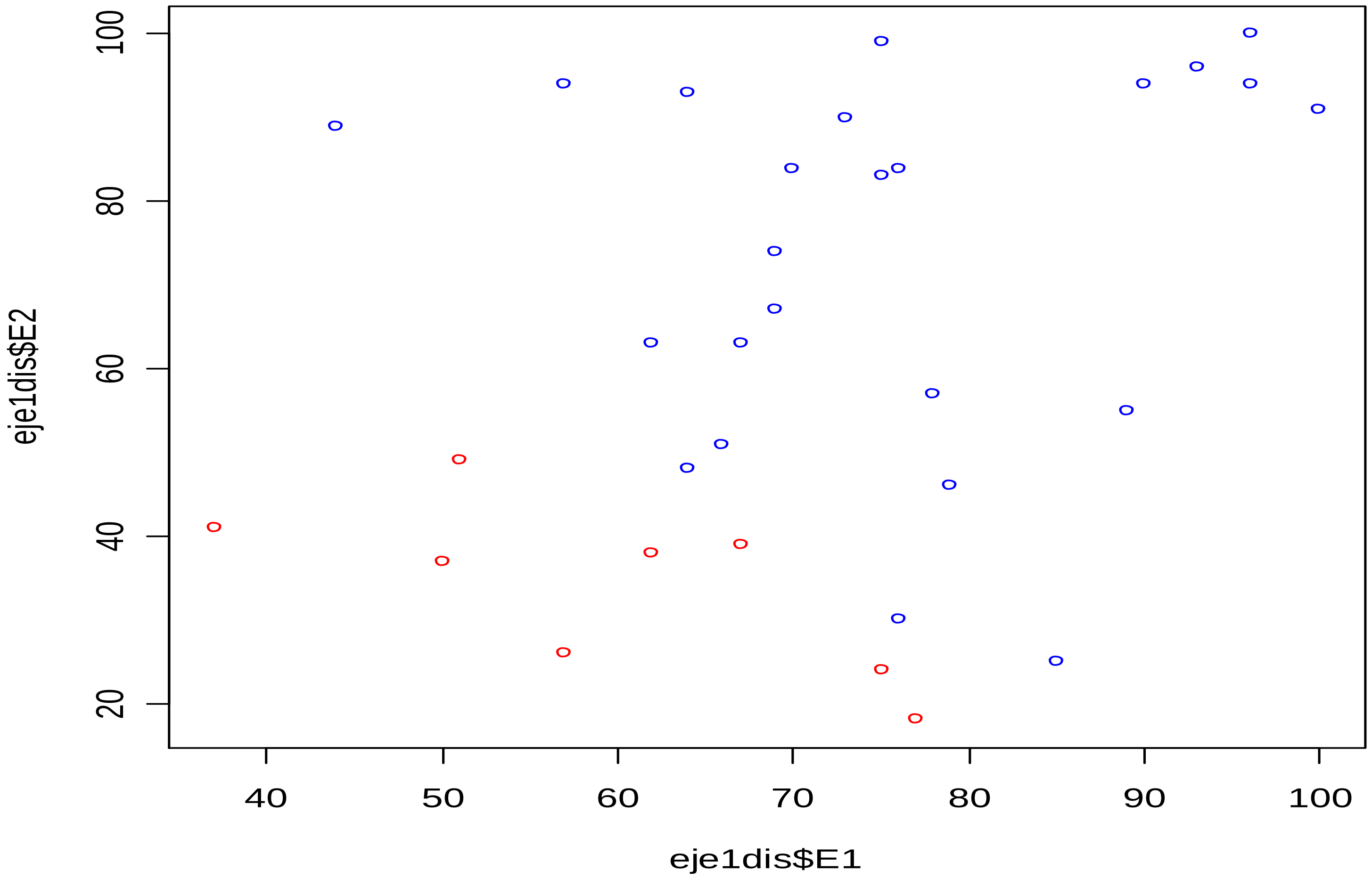
plot(p,sensit,type="l") ;lines(p,especif) ;text(p,sensit,labels=p);
title("Ploteando la sentividad y especificidad para varios p")


La curva ROC
sensit
1 - especif
0.0
0.2
0.4
0.6
0.8
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

#Notar que para p=.57 la curva esta mas cerca a la esquina superior izquierda

```
> #Clasificacion final
> clasesf<-rep(0,nobs)
> for(i in 1:nobs)
+ {if(phat[i]>=0.57){clasesf[i]<-1}
+ }
> ratef<-mean(clasesf!=bupa1[,7])
> cat("la tasa de mala clasificacion optima es=",ratef,"\n")
la tasa de mala clasificacion optima es= 0.3246377
>
```

Total Separability



Regresion logistica en R

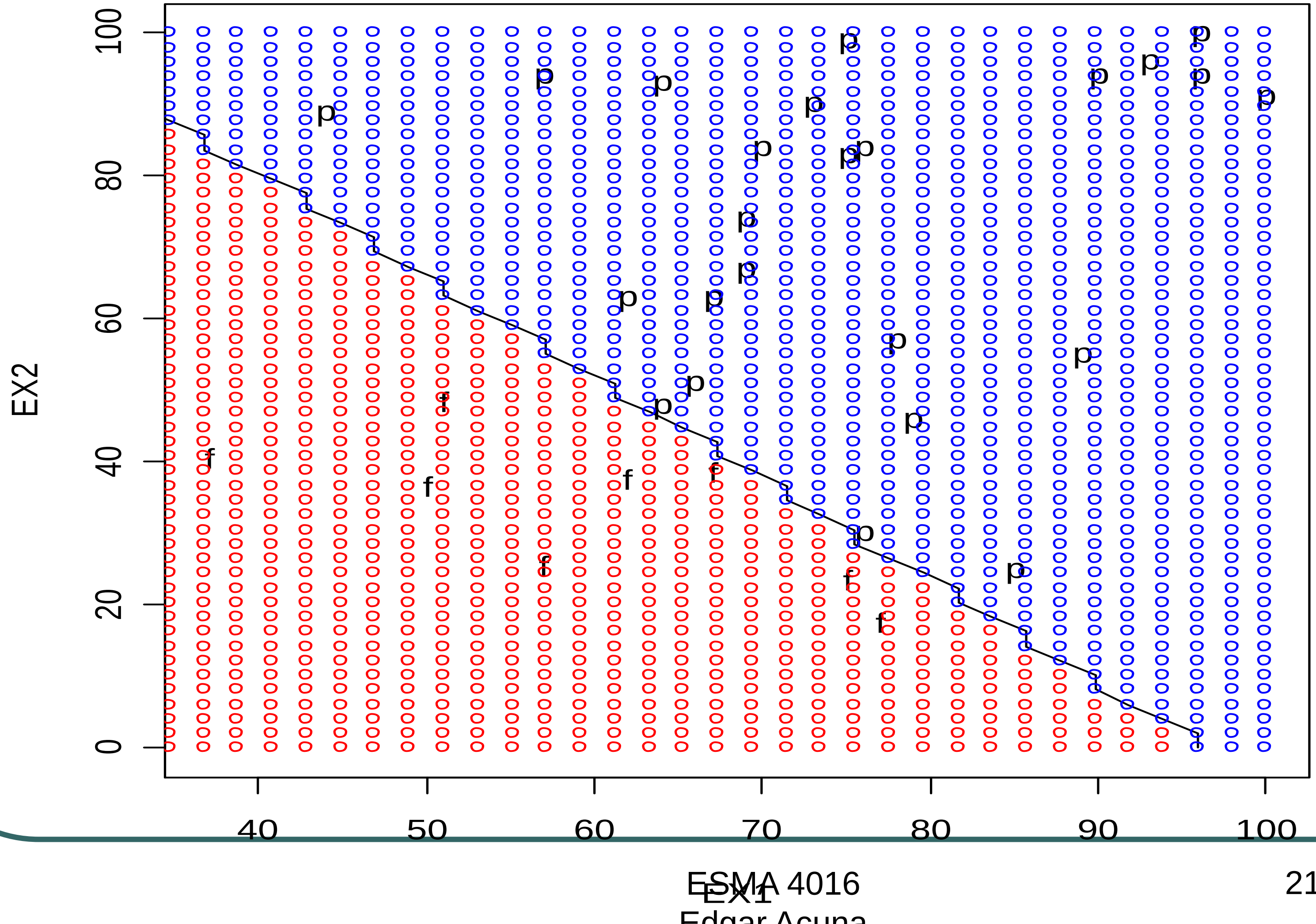
```
> notaslog=glm(notabin~E1+E2,data=eje1log,family=binomial)
Warning messages:
1: In glm.fit(x = X, y = Y, weights = weights, start = start, etastart =
etastart,  :
  algorithm did not converge
2: In glm.fit(x = X, y = Y, weights = weights, start = start, etastart =
etastart,  :
  fitted probabilities numerically 0 or 1 occurred
> notaslog
Call:  glm(formula = notabin ~ E1 + E2, family = binomial, data = eje1log)
Coefficients:
(Intercept)          E1           E2
   -1233.64       13.08         8.67

Degrees of Freedom: 31 Total (i.e. Null);  29 Residual
Null Deviance:      35.99
Residual Deviance: 1.791e-08    AIC: 6
```

Regresion logistica en R

```
>  notaslog$fit
         1           2            3            4            5            6 
1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 
         7           8            9           10           11           12 
1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 
        13          14           15           16           17           18 
1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 
        19          20           21           22           23           24 
1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 
        25          26           27           28           29           30 
2.220446e-16 2.220446e-16 4.525405e-09 2.220446e-16 2.220446e-16 2.220446e-16 
        31          32 
2.220446e-16 2.220446e-16
```

Hay una probabilidad 1 de que los 24 primeros estudiantes aprueban las clases y una probabilidad 0 de que los 8 ultimos aprueben las clases. Esto ocurre porque hay total separability.

Grafica de las fronteras de logistica



Multiclass logistic regression

Fue discutida en detalle por Engel (1988) y es Tambien llamada regresion logistica politomica. Si se tuvieran K clases entonces los odds ratios son tomados con respecto a una de ellas, por decir la k-esima clase. Asi,

Log[P(Y=j/x)/P(Y=K/x)]=$\alpha_j+\beta'_j\mathbf{x}$ para j=1,2,….K-1

De donde,

$$\sum_{j=1}^{K-1} P(Y = j / x) = P(Y = K / x)\exp(\alpha j + \beta'_j x)$$

y usando el hecho que

$$\sum_{j=1}^{K} P[Y = j / \mathrm{x}] = 1$$

Multiclass logistic regression

Se tiene que

$$p_j = P[Y = j/\mathrm{x}] = \frac{\exp(\alpha_j + \beta_j^{'}\mathrm{x})}{1 + \sum_{j=1}^{K-1} \exp(\alpha_j + \beta_j^{'}\mathrm{x})}$$

Para j=1,2,…K-1, y,

$$p_K = P[Y = K/\mathrm{x}] = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\alpha_j + \beta_j^{'}\mathrm{x})}$$

En el area de neural networks estas funciones son llamadas softmax. El objecto x es asignado a la clase j* tal que j*=argmax(p_j). En el algoritmo el vector de K clases es transformado en una matriz de k columnas con entradas 0 y 1 y en el output son reemplazadas por las probabilidades posteriores. Para hacer discriminación logística con varias clases se usa la función **Multinom** de la libreria **nnet.**

Example: Vehicle

```
> cv10log(vehicle,10)
los estimados del error en cada repeticion son
 [1] 0.1914894 0.2033097 0.1950355 0.1985816 0.1926714
    0.1985816 0.1973995
 [8] 0.2044917 0.2033097 0.1973995
 El estimado del error por VC en el numero de repeticiones
    dados es
[1] 0.1982270
>
```

Example: Iris

library(nnet)
> cv10lda(my.iris,repet=10)
[1] 0.02133333
➢cv10log(my.iris,repet=10)
los estimados del error en cada repeticion son
[1] 0.04000000 0.02000000 0.02666667 0.03333333 0.03333333 0.03333333
[7] 0.02666667 0.02000000 0.02666667 0.02000000
El estimado del error por VC en el numero de repeticiones dados es
[1] 0.028

Si la suposicion de normalidad se cumple entonces el discriminante lineal es mejor que la regresion logistica.
La regresion logistica se relaciona con otro clasificador llamado maquina de soporte vectorial (SVM).