1. Conceptos generales de clasificación
2. Clasificador k-vecino más cercano
3. Clasificador Bayesiano óptimo
4. Análisis discriminante lineal (LDA)
5. Clasificadores lineales y el Modelo perceptrón
6. Máquinas de soporte vectorial
7. Regresión logística

7 Regresión logística (RL)

Retomamos el clasificador Bayesiano óptimo, caso binario y $L(x, y) = I(x \neq y)$:

$$f^*(x) = \begin{cases} 0 & \text{si } P(Y = 0|X = x) > P(Y = 1|X = x). \\ 1 & \text{si } P(Y = 1|X = x) \geq P(Y = 0|X = x) \end{cases}$$

Antes, en Análisis Discriminante Lineal usamos:

$$P(Y = 0|X = x) > P(Y = 1|X = x) \quad \leftrightarrow \quad P(X = x|Y = 0)P(Y = 0) > P(X = x|Y = 1)P(Y = 1)$$

y definimos $P(X = x|Y = y)$ y $P(Y = y)$.

Es un enfoque generativo. Problema: tenemos que especificar más de lo que necesitamos.

Ahora: parametrizamos y estimamos directamente:

$$\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}$$

Es un enfoque discriminatorio.

Regresamos a la codificación con $Y = 0, Y = 1$.

Dos casos: x variables nominales o métricas (o una mezcla de ambos).

El caso 1D con x métrico o binario.

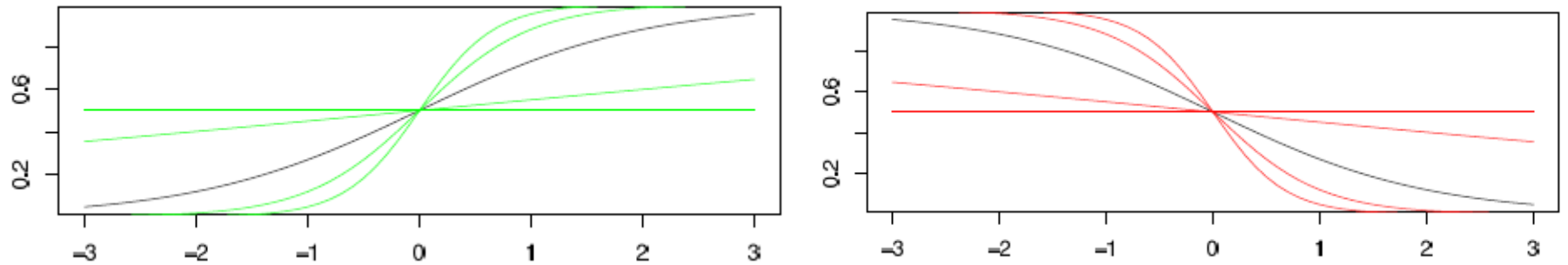
Vamos a suponer que $\exists \alpha, \beta$:

$$\log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \alpha + \beta x.$$

Usando $P(Y = 0|X = x) = 1 - P(Y = 1|X = x)$, lo anterior significa:

$$P(Y = 1|X = x) = \frac{1}{1 + \exp(-\alpha - \beta x)}$$

Interpretación α y β:



Insertar `logit.r`

• Interpretar α:

$$\frac{P(Y=1|X=0)}{P(Y=0|X=0)} = \exp(\alpha).$$

• Interpretar β:

Si x es binario:

$$\frac{\frac{P(Y=1|X=1)}{P(Y=0|X=1)}}{\frac{P(Y=1|X=0)}{P(Y=0|X=0)}} = \exp(\beta).$$

Si x es métrico:

$$\frac{\frac{P(Y=1|X=x+1)}{P(Y=0|X=x+1)}}{\frac{P(Y=1|X=x)}{P(Y=0|X=x)}} = \exp(\beta).$$

Observación final: una alternativa es empezar con:

$$P(Y=1|X=x) = F(\alpha+\beta x),$$

con F una función de distribución acumulativa.

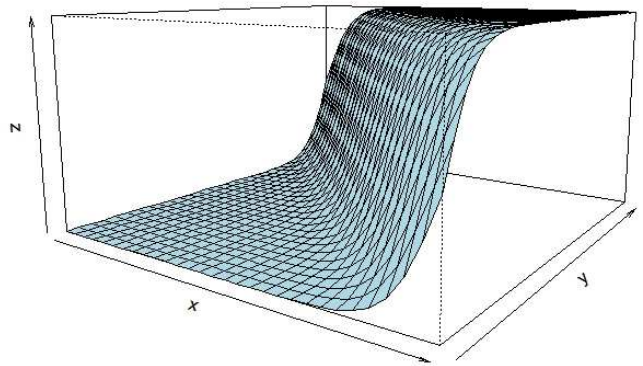
Si es la de una estandar normal: se obtiene *modelo probit*.

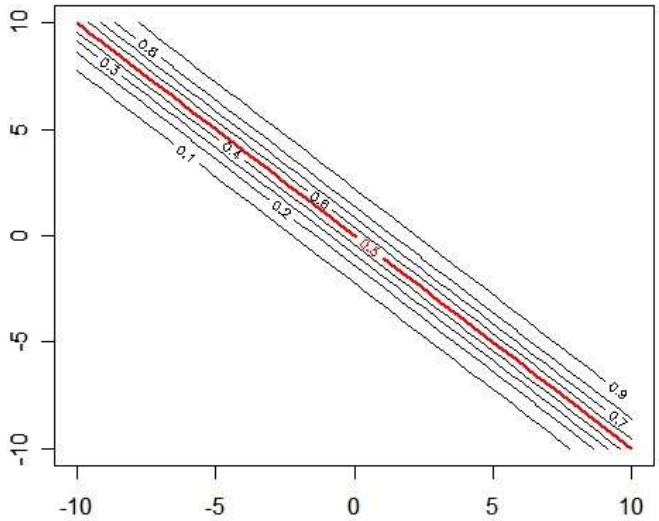
El caso multidimensional con x métrico o binario.

$$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \alpha + \beta^t x.$$

Para simplificar la notación: incluye en vector x la constante 1:

$$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \beta^t x.$$





lineas de contornos: !! paralelas pero no equidistantes!!

para interpretar los parámetros: como antes, fijando todos los predictores menos uno. Insertar `logit2D.r`

El caso con x un vector con valores nominales

$$\log \frac{P(Y=1|X^1=x^1,\cdots,X^m=x^m)}{P(Y=0|X^1=x^1,\cdots,X^m=x^m)} = \alpha + \beta^1_{x_1} + \cdots + \beta^m_{x_m}.$$

Estimación de los parámetros

Define $\pi(x) = P(Y = 1|X = x)$.

La logverosimilitud de $Y|X$ es

$$l(\beta) = \sum \log(\pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}),$$

$$= \sum (y_i \log(\pi(x_i)) - y_i \log(1 - \pi(x_i)) + \log(1 - \pi(x_i))),$$

$$= \sum (y_i \log(\pi(x_i)/(1 - \pi(x_i))) - \log(1/(1 - \pi(x_i)))),$$

$$= \sum (y_i \beta^t x_i - \log(1 + \exp(\beta^t x_i))),$$

Problema: no-lineal en β. Por ejemplo se puede usar Newton-Raphson:

$$\beta^n = \beta^{n-1} - (\frac{\delta^2 l(\beta^{n-1})}{\delta^2 \beta^{n-1}})^{-1}(\frac{\delta l(\beta^{n-1})}{\delta \beta^{n-1}})$$

$$\frac{\delta l(\beta)}{\delta \beta} = \sum_i (y_i x_i - \frac{\exp(\beta^t x_i)}{1 + \exp(\beta^t x_i)} x_i) = \sum_i x_i (y_i - \pi(x_i)),$$

$$\frac{\delta^2 l(\beta)}{\delta^2 \beta} = \sum_i (x_i^t \pi(x_i)(1 - \pi(x_i)) x_i)$$

Si se escribe lo anterior en términos de matrices:

$$\frac{\delta l(\beta)}{\delta \beta} = X^t(Y - \Pi), \frac{\delta^2 l(\beta)}{\delta^2 \beta} = -X^t W X,$$

con $W = Diag(\pi(x_i)(1 - \pi(x_i))$ De este manera:

$$\beta^n = \beta^{n-1} - (X^t W X)^{-1} X^t (Y - \Pi) =$$

$$\beta^n = (X^t W X)^{-1} X^t W (X\beta^{n-1} - W^{-1}(Y - \Pi).$$

es de la forma de regresión lineal ponderada:

$$\beta^n = (X^t W X)^{-1} X^t W Z.$$

Inferencia para los parámetros

Heredamos todas las buenas propiedades de estimadores de máxima verosimilitud.

Si $n \to \infty$:

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, ASE)$$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.86603    0.15447   5.607 2.06e-08 ***
V3          -0.06721    0.12526  -0.537    0.592
V4           0.51243    0.12691   4.038 5.40e-05 ***
V5          -0.14059    0.13409  -1.049    0.294
V6          -0.96718    0.12838  -7.534 4.92e-14 ***
V7          -0.68089    0.16583  -4.106 4.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insertar `logistictodo.r`

Inferencia para modelos

$\Rightarrow$ Likelihood ratio approach.

Toma dos modelos anidados, M_1 y M_2: M_2 es M_1 con algun(os) parámetros θ_γ igual a 0.

Calcula la log verosimilitud (maximizada) para M_1 y M_2: l_1, l_2, entonces bajo $H_o : \theta_\gamma = 0$, si $n \to \infty$:

$$-2(l_2 - l_1) \sim \chi^2_q,$$

con q la diferencia en número de parámetros en θ_γ.

	$X_3=0$				$X_3=1$			
	$X_2=0$		$X_2=1$		$X_2=0$		$X_2=1$	
	$X_1=0$	$X_1=1$	$X_1=0$	$X_1=1$	$X_1=0$	$X_1=1$	$X_1=0$	$X_1=1$
$X_5=0, X_4=0$								
$Y=0$	37	27	51	48	51	55	109	86
$Y=1$	16	11	10	19	24	28	21	25
$X_5=0, X_4=1$								
$Y=0$	16	15	7	6	32	34	30	31
$Y=1$	12	24	13	7	55	39	26	19
$X_5=1, X_4=0$								
$Y=0$	10	8	12	15	2	1	9	5
$Y=1$	9	4	8	9	8	9	4	5
$X_5=1, X_4=1$								
$Y=0$	7	10	7	3	5	2	1	3
$Y=1$	8	4	6	4	10	9	3	6

The variables are: Y whether the student agrees that they will need mathematics in the future (0=agree, 1=disagree); X_1 whether the student assisted to the lectures (0= yes, 1=no), X_2 his/her sex (0=female, 1=male), X_3 his/her type of school (0=suburban, 1=urban), X_4 his/her course preferences (0=mathematics, 1=liberal arts), X_5 his/her future plans (0=college, 1=job).

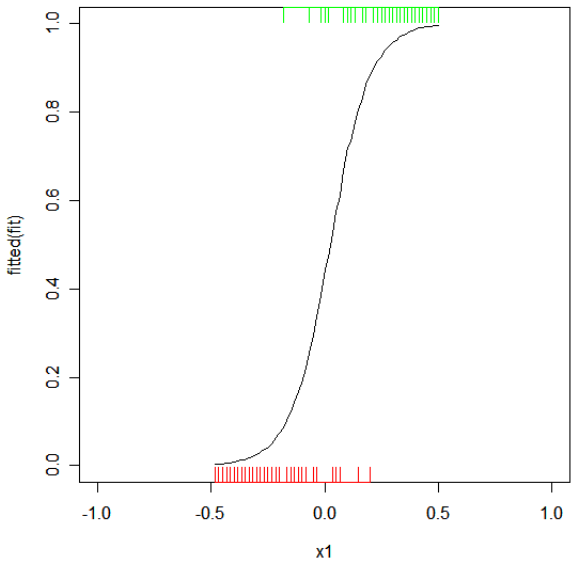
```
Residual deviance:  44.288  on 26  degrees of freedom
```

aquí: M_1 es modelo saturado con el número de parámetros igual al número de distinctos valores de x en la muestra; M_2 es un modelo con 6 parámetros (cf. slide anterior): $q = 2^5 - 6$.

Evaluación de ajuste y búsqueda de modelos

Problema fundamental:

contrario a regresión, lo que se observe $(0/1)$ y lo que se estima $(P(Y|X))$ no viven en el mismo espacio.



Si x es discreta: calcular para cada intervalo porcentaje de 1's esperados y compararlo con el observado.

Si x no es discreta: puede ayudar discretizar en intervalos.

LDA vs RL: ambos definen un modelo lineal para $\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$. La diferencia es como estiman los parámetros!!

Modelar versus predecir

A partir de un modelo de regresión logística, se puede construir un clasificador:

$$\hat{y}(x) = \begin{cases} 1 & \text{si } \log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \alpha + \beta x > 0 \\ 0 & \text{otro caso} \end{cases}$$

Se puede calcular el error de clasificación.

IMPORTANTE:

las preguntas de interés en regresión logística son diferentes a las de los clasificadores que vimos.

Un buen modelo puede predecir (clasificar) bastante pobre.

En regresión logística, el énfasis está en ¿ por qué ? y en tratar de entender la relación entre predictores y respuesta.

En SVM, perceptrón & co. el énfasis está en ¿ como ? clasificar/predecir bien.

¿ Cómo buscar un buen modelo?

Si hay preguntas (información) específicas, hay que usarlas como guía en la construcción.

En dimensiones (muy) altas, típicamente existen varios modelos buenos: difícil de hablar del mejor modelo.

Método más usado: stepwise backward (o forward):

- Empieza con modelo saturado;
- Repite:

  Ajusta modelo RL;

  Busca parámetro menos significativo; iguálalo a 0.

  Si el modelo resultante es adecuado: repite lo anterior con este modelo reducido;

  si no: párate.

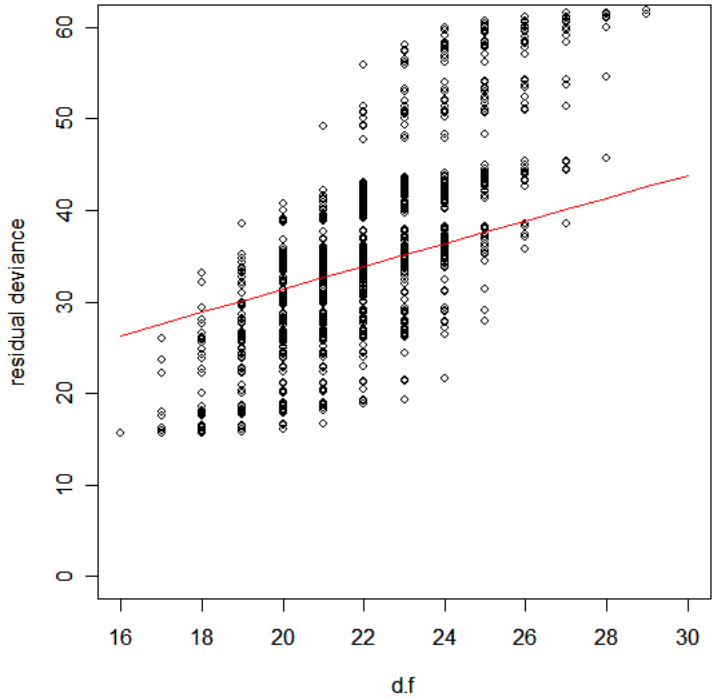
Dos criterios: complejidad y calidad de ajuste; problema multiobjetivo!

¿ Cómo medir?

Degrees of freedom (d.f): entre mayor, más sencillo el modelo.

Residual deviance: entre menor, mejor ajuste.

Para el conjunto data.mat:



RL y regularización

Igual a regresión ridge, se puede introducir un término de regularización en la (log) verosimilitud:

$$-\sum \log(\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}) + \lambda||\beta||_p^2.$$

Las estimaciones obtenidas ya no van a ser asimptóticamente insesgadas.

La regularización puede ser util en casos donde hay problemas con alta variabilidad sobre las estimaciones.

Si $p = 2$, solución es:

$$\beta^n = (X^t W X + \lambda I)^{-1} X^t W Z.$$

En R:

```
library(stepPlr)
n <- 100;p <- 10
x <- matrix(rnorm(n*p),nrow=n)
y <- sample(c(0,1),n,replace=TRUE)
fit <- plr(x,y,lambda=1)
```