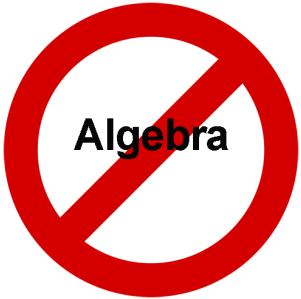
Clasificacion

Walter Sosa-Escudero

Universisad de San Andrés y CONICET

Clasificacion, riesgo y probabilidades

Pacto de damas y caballeros.



Clasificacion

- Admitir a un alumno a un posgrado, en base a sus calificaciones y recomendaciones.
- Dar un credito, en base a antecedentes.
- Paro cardiaco, sobredosis o ataque epileptico, en base a sintomas.
- Spam, personal, social, en base a contenido.

Clasificar Y en base a X. Y es cualitativa. No necesariamente ordenada. Empezamos por el caso binario.

Teoria de la decision

$Y \sim \{0, 1\}, \quad \hat{Y} \sim \{0, 1\}$

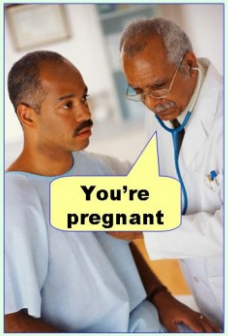
Matriz de confusion:

		Y	
		0	1
$\hat{Y}$	0	vn	fn
	1	fp	vp

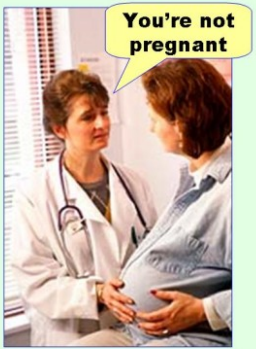
vn = verdadero negativo, fp = falso positivo, fn = falso negativo, vp = verdadero positivo.

Dos acciones ($\hat{Y}$) y dos estados de la naturaleza (Y).

Type I error
(false positive)



Type II error
(false negative)



Riesgo

- Acciones ($\hat{Y}$): $i \in \{0, 1\}$
- Estados de la naturaleza (Y): $j \in \{0, 1\}$
- Probabilidades: $p = Pr(Y = 1|X), 1 - p = Pr(Y = 0|X)$
- Perdida: $\lambda(i, j)$, castiga estar en la casilla i, j.
- Riesgo esperado de la accion i: Castigo esperado.

$$R(i) = (1 - p)\ \lambda(i, 0) + p\ \lambda(i, 1)$$

Idea: elegir accion que minimiza el riesgo esperado. Tenemos que definir $\lambda(i, j)$

Clasificador de Bayes

$$R(i) = (1-p)\ \lambda(i,0) + p\ \lambda(i,1)$$

- *Perdida 0-1*: $\lambda(i,j) = 1\,[i \neq j]$
- $R(0) = p, \quad R(1) = 1-p$

Entonces:

$$\begin{aligned} R(1) &< R(0) \\ 1-p &< p \\ p &> 1/2 \end{aligned}$$

Clasificador de Bayes: elegir el estado mas probable minimiza el riesgo esperado.

- Bajo 'penalidad 0-1', el problema se reduce a encontrar $p = Pr(Y = 1 \,|X)$ y predecir 1 si $p > 0{,}5$ y 0 en caso contrario (clasificador de Bayes).
- Es una probabilidad *condicional*
- Ojo: la penalidad 0-1 es simetrica. ¿Bueno?

Plan de lucha:

1. Tres modelos para $p = Pr(Y = 1 \,|X = x)$: a) logistico, c) vecinos cercanos, c) analisis discriminante, Siempre en el caso binario.
2. Multiples opciones.
3. Perdida asimetrica. Analisis ROC.

Pregunta: ¿Por qué no $p = X\beta$??

- a) ¿Por qué si?, b) Se sale del rango (0,1), c) logit/ analisis discriminante es facil de generalizar al caso de P categorias no ordenadas.

Regresion logistica

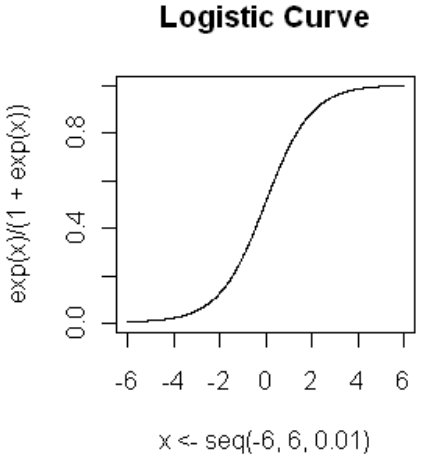
Logit

- $Y =$, variable binaria.
- X vector de P predictores.
- $p = Pr(Y = 1 \,|X)$

Logit

$$p = \frac{e^z}{1 + e^z},$$

$z \equiv X\beta$, β es un vector de P coeficientes.


Logistic Curve
exp(x)/(1 + exp(x))
0.0
0.4
0.8
-6
-4
-2
0
2
4
6
x <- seq(-6, 6, 0.01)

Logit fun: verificar todo

- Asíntotas en 0,1
- Odds ratio: $p/(1-p)$. Interpretacion?
- $$\ln\left(\frac{p}{1-p}\right) = X\beta$$
- $$\beta_k = \frac{\partial \ln\left(\frac{p}{1-p}\right)}{\partial X_k}$$
- Econometria: interes en β (efecto marginal).

Maxima verosimilitud (recordatorio)

$$Pr(Y = y|X) = f(y; \theta)$$

- $f()$ es conocida, no asi sus parametros θ
- Ejemplo: $Y|X \sim$ Poisson(μ), $f(y; \mu) = e^{-\mu}\mu^y/y!$
- $Y_i, i = 1, \ldots, n$ es una muestra iid de tamaño n,

$$Pr(Y_i = 1|X_i) = f(y_i; \theta)$$

Verosimilitud

$$L(\theta; y) = f(y; \theta)$$

Roles cambiados: probabilidad de que y haya ocurrido, para distintos valores de θ.

Verosimilitud muestral: $\mathbf{y} = y_1, y_2, \ldots, y_n$

$$L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta)$$

Si $\mathbf{y}$ es iid

$$L(\theta; \mathbf{y}) = \prod_{i=1}^{n} L(\theta; y_i)$$

(probabilidad conjunta).

Estimador maximo verosimil

$$\hat{\theta} = \text{argmax}_{\theta} \prod_{i=1}^{n} L(\theta; y_i)$$

- Ingenieria reversa.
- Propiedades: hacer el curso de maestria.
- Consistente, asintoticamente normal, asintoticamente eficiente, invariante.

En el caso logit, recordar que

$$p_i = Pr(Y_i = 1|X_i) = \frac{e^{z_i}}{1 + e^{z_i}}, \quad z_i = X_i\beta$$

Entonces

$$L(\beta; \mathbf{y}) = \prod_{y_i=1} p_i \prod_{y_i \neq 1} (1 - p_i)$$

Detalles (maestria): la minimizacion no tiene una forma cerrada. Es un problema computacional.

Resumen

- Observamos $(Y_i, X_i), \;\; i = 1 \ldots, n.$
- Modelo logit

$$p_i = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}}$$

- $\hat{\beta}$ por maxima verosimilitud
- Prediccion

$$\hat{p}_i = \frac{e^{X_i \hat{\beta}}}{1 + e^{X_i \hat{\beta}}}$$

- Clasificacion (Bayes)

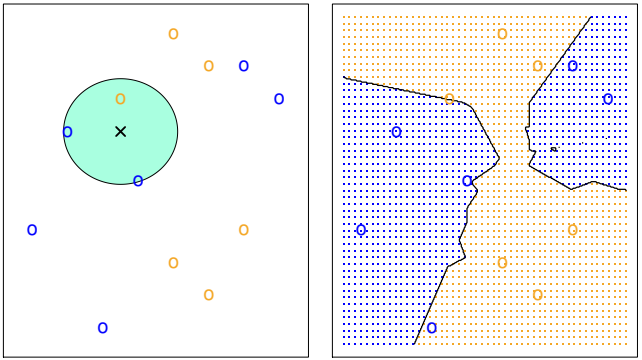
$$\hat{Y}_i = 1 \left[\hat{p}_i > 0{,}5 \right]$$

Vecinos cercanos

Vecinos cercanos

Metodo de fuerza bruta pero efectivo. Detalles cuando veamos metodos no parametricos

- $(Y_i, X_i), i = 1, \ldots, n.$
- Problema: predecir Y dado $X = x_0$.
- $d(X_i, x_0) =$ distancia a x_0.
- K vecinos cercanos a x_0.
- Computar $\sum_{j=1}^{K} 1\left[Y_j\right]/K$
- Predecir de acuerdo a la regla de Bayes (voto por mayoria).



Fuente: James, Witten, Hastie y Tibshirani, 2013, *An Introduction to Statistical Learning*, Wiley, New York.

Cuestiones

- Metodo muy flexible y generalizable.
- Eleccion de K: crucial. Lo discutiremos con detalle mas adelante.
- Eleccion de $d(X_i, x_0)$. No trivial fuera de la distancia euclidea (datos cualitativos). Tambien discutiremos esto mas adelante.
- Mas de un predictor? Simple pero... 'maldicion de la dimensionalidad' (mas adelante).

Analisis discriminante

Maldito Bayes (Regla de Bayes)

$A_i, i = 1, \ldots, K$ sucesos excluyentes y exhaustivos y B otro suceso.

$$P(A_i|B) = \frac{P(B|A_i)\ P(A_i)}{\sum_{i=1}^{k} P(B|A_i)\ P(A_i)}$$

Nuestro problema: conocer $p = Pr(Y = 1|X)$

De acuerdo a la regla de Bayes

$$p = \frac{\kappa \, f_1(x)}{\kappa \, f_1(x) + (1 - \kappa) \, f_0(x)}$$

- $\kappa \equiv Pr(Y = 1)$ (no condicional!)
- $f_1(x) \simeq Pr(X = x|Y = 1)$ (cuidado)
- $f_0(x) \simeq Pr(X = x|Y = 0)$ (cuidado)

Analisis discriminante: estimar κ, f_1, f_0, reemplazar y usar el clasificador de Bayes.

$\kappa \equiv Pr(Y = 1)$. Este es facil

$$\hat{\kappa}_i = \frac{\sum_{i=1}^{n} 1\left[Y_i = 1\right]}{n}$$

(proporcion de 1's en la muestra)

Supongamos que hay un solo predictor y que $X|Y$ es *normal*:

$$f_j(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_j)\right)^2, \; j = 0, 1$$

- El problema se redujo a estimar $\mu_0, \mu_1, \sigma_0^2, \sigma_1^2$
- Facil: partir la muestra en dos, de acuerdo a $Y = 1$ y $Y = 0$ y estimar las medias y varianzas para cada particion

Resumamos

- Necesitamos p_i
- Por Bayes

$$p_i = \frac{\kappa f_1(x_i)}{\kappa f_1(x_i) + (1-\kappa) f_0(x_i)}$$

- Entonces, necesitamos $\kappa, f_1(x_i)$ y $f_0(x_i)$.
- κ es facil (proporcion de 1's)
- Supongamos un abrelatas: un solo predictor, normal.
- Ahora necesitamos $\mu_0, \mu_1, \sigma_0^2, \sigma_1^2$. Tambien facil (partir la muestra)
- Reemplazar en las formulas. Computar $\hat{p}_i$
- Usar el clasificador de Bayes

Abrelatas: normalidad? (no, por ahora).

Un solo predictor? Sea X un vector de p varaibles aleatorias. Normal multivariada

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

- $\mu = E(X)$ (ojo que es un vector)
- $\Sigma = V(x)$ (ojo que es una matriz)

Reemplazar y ya estamos de vuelta en el problema anterior.

Para nuestro problema, con P predictores y bajo normalidad

$$f_j(x) = \frac{1}{(2\pi)^{P/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_j)'\Sigma_j^{-1}(x_j-\mu)\right), \quad j = 0, 1$$

Ahora tenemos que estimar μ_j y Σ_j

- $\hat{\mu}_j$ es el *vector* de medias muestrales en cada particion, $j = 0, 1$.
- $\hat{\Sigma}_j$ es la matriz de varianzas y covarianzas para cada particion.

Discriminacion lineal y cuadratica

Notar que (chequear):

$$p > 1/2 \iff \ln\left(\frac{p}{1-p}\right) > 0$$

Logit con un predictor:

$$\beta_1 + \beta_2 X > 0$$

- Clasificacion: en el espacio de probabilidades.
- Discriminacion: en el espacio de X.
- $\beta_1 + \beta_2 X$ es la *funcion de discriminacion* del logit: lineal.

Analisis discriminante? Caso: un predictor, $\sigma_0^2 = \sigma_1^2$ (igual varianza). Recordar

$$p = \frac{\kappa f_1(x)}{\kappa f_1(x) + (1-\kappa) f_0(x)}$$

Entonces, bajo los supuestos

$$\frac{p}{1-p} = \frac{\kappa f_1(x)}{(1-\kappa) f_0(x)} = \frac{\kappa \exp\left((x-\mu_1)\right)^2}{(1-\kappa) \exp\left((x-\mu_0)\right)^2}$$

Tomando logs

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{\kappa}{1-\kappa}\right) + (x-\mu_1)^2 - (x-\mu_0)^2$$

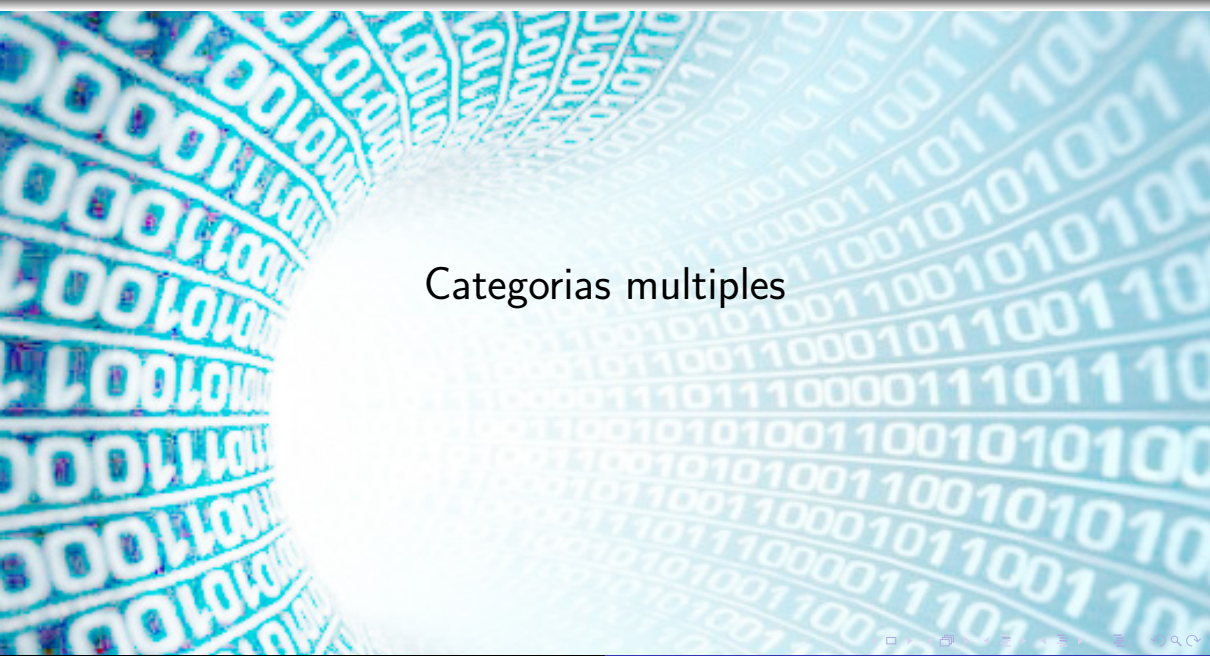
$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \ln\left(\frac{\kappa}{1-\kappa}\right) + (x-\mu_1)^2 - (x-\mu_0)^2 \\
&= \ln\left(\frac{\kappa}{1-\kappa}\right) + \mu_1^2 - \mu_0^2 - 2x(\mu_1 - \mu_0) \\
&= \gamma_1 + \gamma_2\, x,
\end{aligned}$$

$\gamma_1 \equiv \ln(\kappa/(1-\kappa)) + \mu_1^2 - \mu_0^2$, $\gamma_2 \equiv -2(\mu_1 - \mu_0)$

- Bajo igual varianza la funcion de discriminacion es *lineal*.
- Igual varianza: *analisis discriminante lineal*.
- Logit estima directamente γ_1 y γ_2

Chequear (todo facil, algebra)

- Multiples predictores logit: trivial
- Multiples predictores, igual varianza analisis discriminante: mas engorroso, pero trivial.
- Un predictor, varianza variable: la funcion de discriminacion es *cuadratica*: analisis discriminante cuadratico.
- Multiples predictores, varianza variable: mas engorroso, tambien facil.



Categorias multiples

K categorias

- $Y = \{1, 2, \ldots, K\}$
- Ordenadas: bajo, medio, alto.
- No ordenadas: arte, ciencia, deporte.
- Logit y el modelo lineal no se extienden facilmente (menos aun al caso no ordenado).
- K-vecinos mas cercanos: simple. Computar probabilidades para todas las cateogorias, predecir la de la mayoria.

Analisis discriminante: muy simple. Recordar

$$p = \frac{\kappa \ f_1(x)}{\kappa \ f_1(x) + (1 - \kappa) \ f_0(x)}$$

La generalizacion al caso de $k =, 1, \ldots, K$ categorias es

$$p(k) = \frac{\kappa_k \ f_k(x)}{\sum_{j=1}^{K} \kappa_j \ f_j(x)}$$

- $\hat{\kappa}_k$ proporcion (no condicional en cada categoria)
- Ahora para $j = 1, \ldots, K$,

$$f_j(x) = \frac{1}{(2\pi)^{P/2}|\Sigma_j|^{1/2}} \quad \exp\left(-\frac{1}{2}(x-\mu_j)'\Sigma_j^{-1}(x_j-\mu)\right)$$

- Repetir lo anterior para cada una de las K categorias.
- Reemplazar en la regla de Bayes.

Analisis ROC

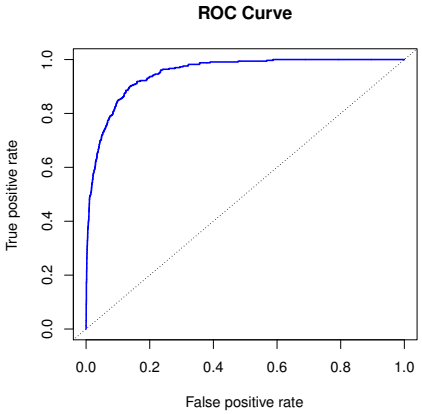
Analisis ROC

Cuadro: My caption

		Y		
		0	1	
$\hat{Y}$	0	vn	fn	$\hat{n}_0$
	1	fp	vp	$\hat{n}_1$
		n_0	n_1	n

Falso positivo (Err I) $= fp/n_0$
Verdadero positivo (1-Err II) $= vp/n_1$

- Estaria bueno que $fp = fn = 0$. Imposible.
- Distintas medidas de error agregado (ver Wikipedia!)
- Accuracy: $tp + tn/n$. Suerte de R^2.
- No puede estar debajo de 0.5. Por que?
- Minimizar un error implica maximizar el otro.
- Trade off: empezar con tipo $1 = 0$ (admito todos, corte 1), bajar progresivamente el corte.
- Curvar ROC: err tipo II vs. 1 - err tipo II.



Fuente: James, Witten, Hastie y Tibshirani, 2013, *An Introduction to Statistical Learning*, Wiley, New York.