

Universidad de Santiago de Chile  
Facultad de Ingeniería  
Depto. de Ingeniería Informática



## *Análisis de Datos* *Capítulo IV* *“Análisis Discriminante”*

Profesor: Dr. Max Chacón

### **Objetivos:**

(Repaso: Dominar los métodos de clasificación basados en razón de probabilidades (logística))

- Presentar métodos de clasificación paramétrico y no paramétrico más conocidos.
- Presentar un método de clasificación basados en discriminación lineal.
- Comprender el método no paramétrico de discriminación (clasificación y regresión) mas simple.
- Comprender la metodología de evaluación de la clasificación binaria.



## Repaso Regresión Logística

Aquí se considera el caso binario, esto es la variable dependiente o respuesta  $y$  es una variable binaria  $y \in \{0, 1\}$ .

Es posible también generalizar a  $m$  clases.

La relación con el modelo es mediante la probabilidad de ocurrencia de esta variable  $p = Pr(y=1) \in [0, 1]$  o  $1-p = Pr(y=0)$ .

Para realizar la estimación de parámetros de este problema no-lineal, se requiere maximizar la probabilidad conjunta.

$$Pr(y_i=1 \wedge y_j=0) \quad \forall i, j.$$

Principio de *máxima verosimilitud*.



El estimador de *verosimilitud* para esta probabilidad es:

$$\prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \exp \left[ \sum_{i=1}^n (y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)) \right]$$

La función a maximizar es logaritmo del estimador de verosimilitud.

$$l(y_i, \pi_i) = \sum_{i=1}^n (y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i))$$

$$\frac{\partial l(y_i, \pi_i)}{\partial \beta_i} = 0 \quad \text{con} \quad \pi_i = \frac{1}{1 + e^{-x_i^T \beta}}$$



Esta ecuación resulta ser trascendental y se debe recurrir a métodos iterativos como el método de optimización de Newton, variación del método de Newton-Raphson ( $x_{n+1} = x_n - f(x_n)/f'(x_n)$ ) para ecuaciones trascendentales.

$$H^{(n-1)} \bar{\beta}^n = H^{(n-1)} \bar{\beta}^{(n-1)} + \bar{J}^{(n-1)}$$

con  $n$  el índice de las iteraciones,

- $H$  es la matriz de segundas derivadas (Hessiana) de  $l$ ,

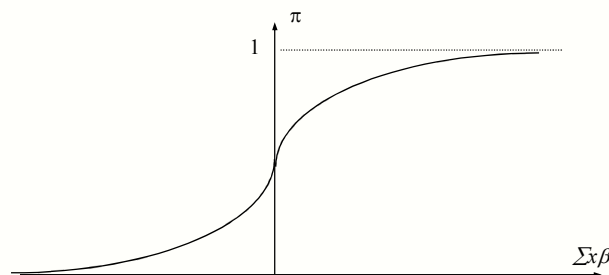
$$h_{jk} = \left[ \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right]$$

- $J$  es el vector de las primeras derivadas (Jacobiana) de  $l$ ,

$$j_j = \frac{\partial l}{\partial \beta_j}$$



Este modelo aproxima la combinación lineal de las variables independientes  $x$  y los parámetros  $\beta$  a la probabilidad  $p$  mediante una función sigmoide.



Éste es uno de los modelos mas utilizados cuando la variable de salida es binaria como es el caso de la mortalidad, predicción de la probabilidad de falla de una máquina o equipo, o probabilidad de ocurrencia de eventos en general, cuando influyen muchas causas.



### - Interpretación de la regresión logística

En el caso lineal los coeficientes  $\beta_i$  representan el incremento de la variable respuesta  $y$ , para un aumento unitario del predictor  $x_i$ .

En la regresión logística el coeficiente puede ser interpretado como el cambio en la función logística para un incremento unitario en el predictor. Considere:

$$\frac{\pi(x_i = 1)}{1 - \pi(x_i = 1)} = \frac{e^{(\beta_0 + \beta_1 x_1 \dots \beta_i \dots \beta_p x_p)} / (1 + e^{(\beta_0 + \beta_1 x_1 \dots \beta_i \dots \beta_p x_p)})}{1 / (1 + e^{(\beta_0 + \beta_1 x_1 \dots \beta_i \dots \beta_p x_p)})}$$

$$\frac{\pi(x_i = 1)}{1 - \pi(x_i = 1)} = e^{(\beta_0 + \beta_1 x_1 \dots \beta_i \dots \beta_p x_p)}$$

Aplicando lo mismo al caso en que la variable  $x_i$  es cero.



Se tiene

$$\frac{\pi(x_i = 0)}{1 - \pi(x_i = 0)} = \frac{e^{(\beta_0 + \beta_1 x_1 \dots \beta_{i-1} + \beta_{i+1} \dots \beta_p x_p)} / (1 + e^{(\beta_0 + \beta_1 x_1 \dots \beta_{i-1} + \beta_{i+1} \dots \beta_p x_p)})}{1 / (1 + e^{(\beta_0 + \beta_1 x_1 \dots \beta_{i-1} + \beta_{i+1} \dots \beta_p x_p)})}$$

$$\frac{\pi(x_i = 0)}{1 - \pi(x_i = 0)} = e^{(\beta_0 + \beta_1 x_1 \dots \beta_{i-1} + \beta_{i+1} \dots \beta_p x_p)}$$

La razón entre estas dos proporciones se denomina razón de probabilidades o razón de chances (*odds ratio*).

$$odds\ ratio = \frac{\pi(x_i = 1)}{1 - \pi(x_i = 1)} \frac{1 - \pi(x_i = 0)}{\pi(x_i = 0)} = \frac{e^{(\beta_0 + \beta_1 x_1 \dots \beta_i \dots \beta_p x_p)}}{e^{(\beta_0 + \beta_1 x_1 \dots \beta_{i-1} + \beta_{i+1} \dots \beta_p x_p)}}$$

$$OR = Odds\ Ratio = e^{(\beta_i)}$$

Esto representa la influencia (o la pendiente) de cada variable en la probabilidad final.

También es de interés el numerador del *OR*, denominado *Razón de Riesgo*:

$$RR = \frac{\pi(x_i = 1)}{1 - \pi(x_i = 1)} = e^{(\beta_0 + \beta_1 x_1 \dots \beta_i \dots \beta_p x_p)}$$

Cuanto crece el riesgo (se multiplica) al considerar  $x_i$ .



### 7.4.2. Evaluación regresión logística

#### -Evaluación de los coeficientes.

Similar al caso de la regresión lineal, es posible contrastar (docimar) la hipótesis de que un coeficiente aislado es distinto de 0, y sigue una distribución normal de media 0 y varianza 1.

El contraste se realiza utilizando la el estadístico de Wald por el cociente entre el valor del coeficiente ( $\hat{\beta}_i$ ) y su correspondiente error estándar.

Esto es:  $H_0 : \hat{\beta}_i = 0$

$$H_a : \hat{\beta}_i \neq 0 \quad Z_{Wald} = \frac{\hat{\beta}_i}{Err\ Est(\hat{\beta}_i)}$$

El cual sigue una distribución normal. Para aceptar  $H_0$  Se quiere:

$$P(|z| < z_{Wald}) < \alpha$$



#### -Evaluación del modelo

Similar al caso de la regresión lineal donde se mide la eficiencia de la regresión por la razón entre los errores cuadrados de la regresión y la suma de errores cuadrados totales, que resulta en una distribución  $F$ .

Análogo al anterior aquí se compara la razón de la verosimilitud del modelo saturado (con todos los predictores) y el modelo nulo (con sólo la intercepción):

$$G^2 = 2 \ln \left( \frac{L}{L_0} \right) = 2(\ln L - \ln L_0) \quad \ln L_0 = n_1 \ln n_1 - n_0 \ln n_0 - n \ln n$$

$$\begin{array}{ll} \text{Con } H_0: & \beta_1 = \beta_2 \dots \beta_p = 0 \\ H_a: & \text{los } \beta_i \neq 0 \end{array}$$

Se denomina el test de la razón de verosimilitud y se obtiene usando una distribución de  $\chi^2$  con  $p$  grados de libertad.



**Ej: Modelo para predecir enfermedad cardiaca, datos de Cliveland**

Se cuenta con 296 casos obtenidos de la ciudad de Cliveland, con 14 atributos (originalmente 76). 13 características que inciden o son causa de enfermedad cardiaca y la clase que corresponde a enfermedad o no. Los atributos son:



N°	Nombre	Característica	N°	Nombre	Característica
1	age	Edad (años)	8	thalach	Infartos previos
2	sex	Sexo (0,1)	9	exang	Angina inducida por ejercicio (0,1)
3	cp	Dolor Pectoral (1-4)	10	oldpeak	Depresión segm. ST, por ejercicio
4	trestbps	Presión Sanguínea (mmHg)	11	slope	Pendiente segm. ST por ejerc. (1-3)
5	chol	Colesterol Sérico (mg/dl)	12	ca	N° vasos coloreados por fluoroscopia (0-3)
6	fb	Glucosa ayunas (0,1)	13	thal	Defectos cardiacos en ejercicio (1-3)
7	restecg	ECG reposo (0-2)	14	num	Diagnostico de enfermedad (0,1)

**Solución:**

Se transforman los datos en formato .rtff, se leen con el programa KNIME y se realiza regresión logística con atributo **num** como la clase

Cleveland-14-heart-disease Logistic Regression, **KNIME**

Log-likelihood = **-93.5797**, Number of iterations = 30, Logit 50

Las variables mas significantivas tienen  $p < 0.05$ .

Logit	Variable	Coeff.	Std. Err.	z-score	P> z
1	age	0.0152	0.025	0.6061	0.5444
2	sex=male	-1.672	0.5451	-3.0674	<b>0.0022</b>
3.1	cp=atyp_angina	0.7182	0.5676	1.2653	<b>0.2058</b>
3.2	cp=non_anginal	1.7939	0.5083	3.5293	<b>0.0004</b>
3.3	cp=typ_angina	2.0447	0.6718	3.0437	<b>0.0023</b>
4	trestbps	-0.0225	0.0114	-1.9823	<b>0.0474</b>
5	chol	-0.0042	0.0041	-1.0278	0.304
6	fb=t	0.591	0.6128	0.9644	0.3349
7.1	restecg=normal	0.4175	0.3879	1.0761	0.2819
7.2	restecg=st_t_wave_abnormality	-0.4738	2.4689	-0.1919	0.8478



Valor p, continuación:

Logit	Variable	Coeff.	Std. Err.	z-score	P> z
8	thalach	0.018	0.0112	1.6034	0.1088
9	exang=yes	-0.775	0.4444	-1.7439	<b>0.0812</b>
10	oldpeak	-0.3676	0.2316	-1.5872	0.1125
11.1	slope=flat	-0.6771	0.8479	-0.7985	0.4246
11.2	slope=up	0.5866	0.9185	0.6387	0.523
12	ca	-1.3642	0.2857	-4.7747	<b>1.80E-6</b>
13.1	thal=normal	-0.0423	0.791	-0.0535	0.9574
13.2	thal=reversible_defect	-1.4582	0.7779	-1.8744	<b>0.0609</b>
	Constant	2.8052	2.8829	0.973	0.3305

Observar Log-likelihood =  $\ln(L)$  = **-93.5797**

Modelo 0:  $n_0=160$ ,  $n_1=136$ ,  $n=296$

$G=2(-93.58-(160\ln 160+136\ln 136-296\ln 296))$

$G=2(-93.58+1540.44)=2893.72$ . Dócima  $p(\chi^2_{12})=0.000$

**Modelo muy bueno**  $p < 0.05$ .

Pero hay que ver los test de coeficientes,  $Z_{\text{wald}}$ , Eliminar variables donde  $p_{\text{wald}} > 0.05$ .



Usando **KNIME** para las 6 variables con  $p < 0.05$ :

Cleveland-6-heart-disease Logistic Regressio

Log-likelihood = -108.2666, Number of iterations = 30

Logit	Variable	Coeff.	Std. Err.	z-score	P> z
1	sex=male	-1.0866	0.4425	-2.4554	0.0141
2.1	cp=atyp_angina	1.3645	0.5185	2.6316	0.0085
2.2	cp=non_anginal	1.7781	0.4467	3.9807	6.87E-5
2.3	cp=typ_angina	1.645	0.5924	2.7771	0.0055
3	trestbps	-0.0188	0.0098	-1.9238	<b>0.0544</b>
4	exang=yes	-1.1993	0.3933	-3.0495	0.0023
5	ca	-1.2246	0.2228	-5.4976	3.85E-8
6.1	thal=normal	0.7039	0.71	0.9913	<b>0.3215</b>
6.2	thal=reversible_defect	-1.0013	0.7029	-1.4245	<b>0.1543</b>
	Constant	3.7425	1.6127	2.3207	0.0203

Observar Log-likelihood =  $\ln(L)$  = **-108.5797**

Modelo 0:  $n_0=160$ ,  $n_1=136$ ,  $n=296$

$G=2(-108.58+1540.44)=2863.72$ . Dócima  $p(\chi^2_5)=0.000$

**Modelo muy bueno**  $p < 0.05$ .



Usando **Weka** para el mismo archivo de 6 variables

Logistic Regression with ridge parameter of 1.0E-8 Coefficients...		Odds Ratios...	
Variable	Class <50	Variable	Class <50
sex	-1.0866	sex	0.3374
cp=typ_angina	0.6388	cp=typ_angina	<b>1.8942</b>
cp=asympt	-1.0063	cp=asympt	0.3656
cp=non_anginal	0.7719	cp=non_anginal	<b>2.1638</b>
cp=atyp_angina	0.3582	cp=atyp_angina	1.4308
trestbps	-0.0188	trestbps	0.9814
exang	-1.1993	exang	0.3014
ca	-1.2246	ca	0.2939
thal=fixed_defect	0.1175	thal=fixed_defect	1.1247
thal=normal	0.8214	thal=normal	<b>2.2737</b>
thal=reversable_defect	-0.8838	thal=reversable_defect	0.4132
Intercept	4.6312		

OR indica que por cada incremento de las variables cp="dolor pectoral" y thal="problemas cardiacos en ejercicios", el riesgo de estar con una afección cardiaca aumenta al doble.

Usando **Weka** para el mismo archivo de 6 variables

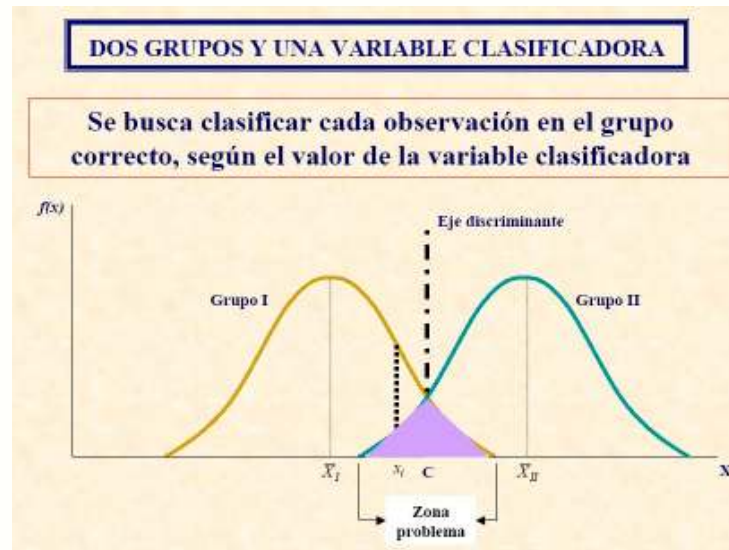
Summary		Detailed Accuracy By Class						
Correctly Classified Instances	249	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
84.1216 %		0.875	0.199	0.838	0.875	0.856	0.915	<50
Incorrectly Classified Instances	47	0.801	0.125	0.845	0.801	0.823	0.915	>50_1
15.8784 %		Weighted Avg.						
Kappa statistic	0.6791	0.841	0.165	0.841	0.841	0.841	0.915	
Mean absolute error	0.227							
Root mean squared error	0.3354							
Relative absolute error	45.7027 %							
Root relative squared error	67.2958 %							
Total Number of Instances	296							

## Confusion Matrix

a	b	←classified as
140	20	a = <50
27	109	b = >50_1



#### 4.1. Idea básica del discriminante de Fisher



Hipótesis: Las distribuciones sólo se diferencian por su localización (igual forma y varianza)

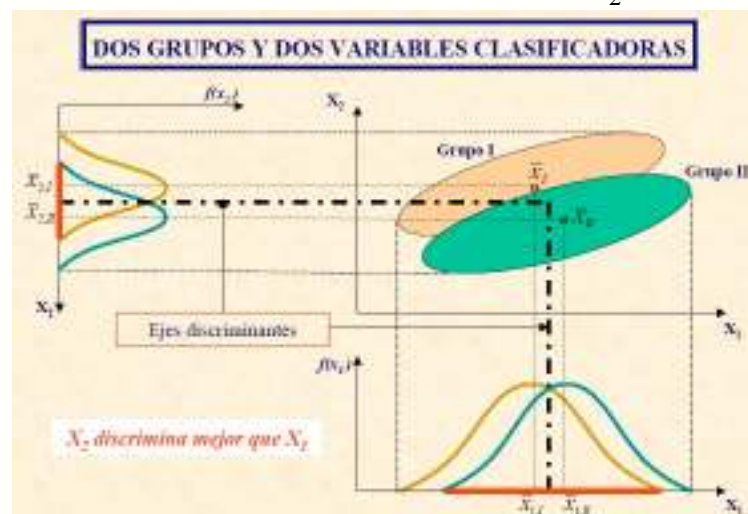
Se trata de minimizar los errores de clasificación

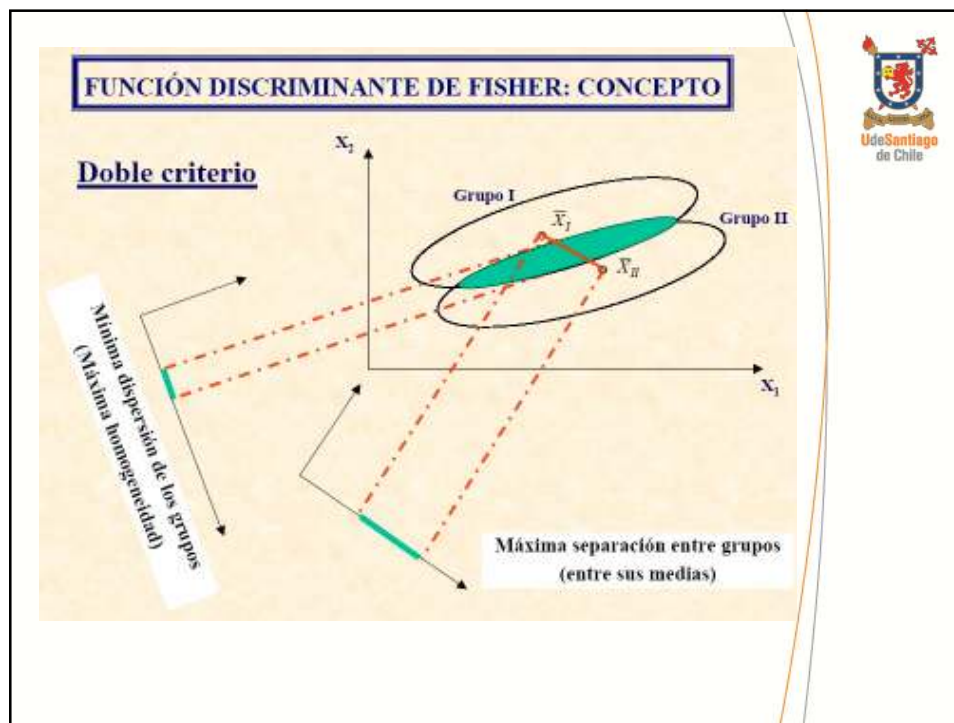
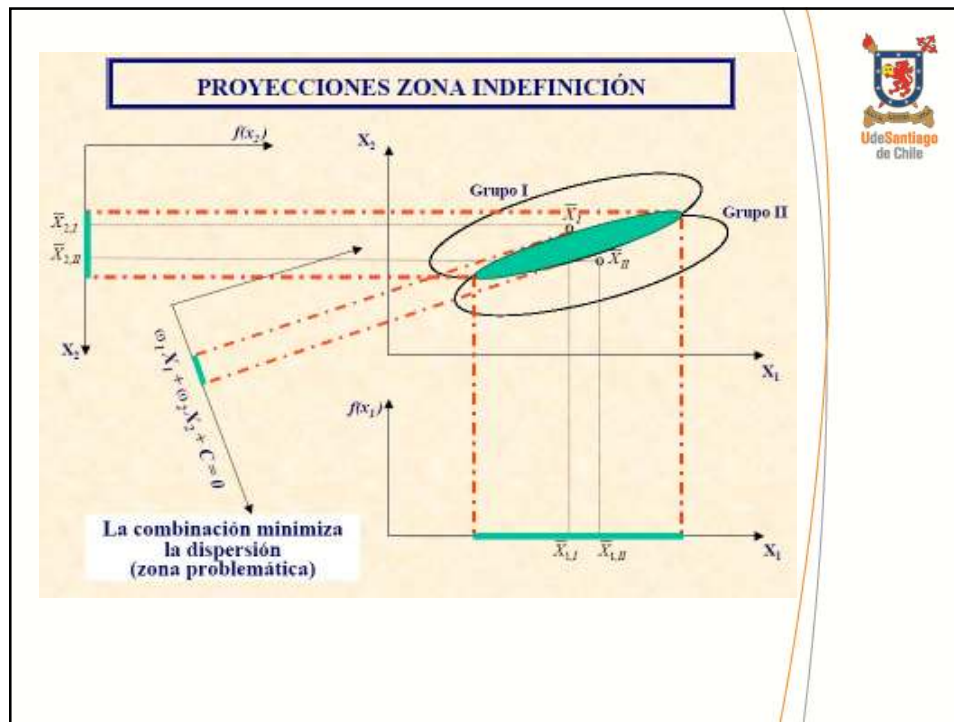
Si  $x_i < C$  se clasifica en el grupo I

Si  $x_i > C$  se clasifica en el grupo II

El punto  $C$  se denomina punto de corte discriminante:

$$C = \frac{\bar{X}_I + \bar{X}_{II}}{2}$$





#### 4.2. Discriminante Lineal de Fisher (matemática)

Para discriminar entre poblaciones se pretende separar poblaciones mediante funciones lineales, para las cuales se les debe determinar los parámetros de las funciones lineales  $\beta$ .

Este método no asume restricciones respecto de las distribuciones, acepto que las varianzas-covarianzas sean homocedástica, esto es: las matrices de varianzas-covarianzas deben ser aproximadamente iguales en cada grupo.



Sea  $X$  la matriz de datos de la muestra de entrenamiento, que incluye sólo las variables independientes (excluye la columna con el factor que indica la clase a la pertenece la población).

La dimensión de la matriz  $X$  es de  $n$  filas y  $p$  columnas  $X_{n \times p}$ . Con  $m$  grupos.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix}$$

La matriz  $X$  será agrupada de acuerdo a  $m$  grupos que separan la población

$$y_1 = \beta_{10} + \beta_{11}x_1 + \dots + \beta_{1p}x_p$$

...

$$y_m = \beta_{m0} + \beta_{m1}x_1 + \dots + \beta_{mp}x_p$$



$X_k$  es la sub-matriz de la población  $I_k$  que corresponden a las observaciones de la población  $\mathcal{P}_k$ .

La varianza total se puede descomponer en cada grupo:

$$\text{cov}(x_j, x_l) = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)$$

La media en cada grupo de la variable  $x_j$  será:  $\bar{x}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} x_{ij}$

Donde  $I_k$ , corresponde a cada grupo  $k=1, \dots, m$ .

La media total de la variable  $x_j$  será el promedio de las sumas de:  $\sum_{i \in I_k} x_{ij} = n_k \bar{x}_{kj}$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} = \frac{1}{n} \sum_{k=1}^m \sum_{i \in I_k} x_{ij} = \frac{1}{n} \sum_{k=1}^m n_k \bar{x}_{kj} = \sum_{k=1}^m \frac{n_k}{n} \bar{x}_{kj}$$



Así cada uno de los términos de la covarianza se puede separar

$$\text{cov}(x_j, x_l) = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)$$

$$(x_{ij} - \bar{x}_j) = (x_{ij} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j)$$

$$(x_{il} - \bar{x}_l) = (x_{il} - \bar{x}_{kl}) + (\bar{x}_{kl} - \bar{x}_l)$$

$$\text{cov}(x_j, x_l) = \frac{1}{n} \sum_{k=1}^m \sum_{i \in I_k} (x_{ij} - \bar{x}_{kj})(x_{il} - \bar{x}_{kl}) + \sum_{k=1}^m \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kl} - \bar{x}_l)$$

$$\text{cov}(x_j, x_l) = d(x_j, x_l) + e(x_j, x_l)$$

La covarianza total se puede dividir en la covarianza dentro de los grupos  $d(x_j, x_l)$  y la covarianza entre los grupos  $e(x_j, x_l)$ .

$$t(x_j, x_l) = d(x_j, x_l) + e(x_j, x_l) \quad \text{matricialmente } T = D + E$$



Las funciones lineales están dadas por  $\hat{y}_k = \vec{\beta}_k^T \vec{x}$

Condicionando a que la primera ( $y_1$ ) maximiza el cociente entre la suma de cuadrados entre grupos y la suma de cuadrados dentro de los grupos, en la muestra de entrenamiento.

La segunda maximiza lo mismo, pero en el espacio ortogonal a  $\beta_1$ , la tercera igual pero en el espacio ortogonal a  $\beta_2$  y así hasta el numero de clases.

En general  $y_k$  es la combinación lineal de  $x_1 \dots x_p$  la mayor discriminación posible entre los grupos después de  $y_{k-1}$  tal que  $\text{corr}(y_k, y_j) = 0$ , para  $j = 1, \dots, (k-1)$ .



El desarrollo es similar a la regresión lineal, pero la variable dependiente  $y$  es categórica con  $k$  categorías.

Para que el método funcione se requieren al menos dos grupos y al menos dos casos en cada grupo.

Entonces el número de variables discriminantes debe ser menor que en numero de casos menos 2;  $p < (n-2)$ .

Ninguna variable discriminante debe ser función de otra.

El numero de funciones discriminantes debe ser el mínimo entre el numero de variables y el numero de grupos menos 1.

Se requiere determinar los  $\beta$  de  $\hat{y}_k = \vec{\beta}_k^T \vec{x}$  forma tal que la varianza entre los grupos sea mayor, respecto de la varianza total.



La varianza de los  $\vec{y}_k = \vec{\beta}^T X$ , puede ser calculada mas fácilmente al considerar las medias cero  $E[\vec{y}_k] = 0$  :

$$\text{var}(\vec{y}_k) = E[\vec{y}_k \vec{y}_k^T] = E[\vec{\beta}_k^T X X^T \vec{\beta}_k] = \vec{\beta}_k^T E[XX^T] \vec{\beta}_k$$

Pero

$$E[XX^T] = \begin{bmatrix} \text{cov}(x_1, x_1) & \dots & \text{cov}(x_1, x_p) \\ \dots & \text{cov}(x_j, x_l) & \dots \\ \text{cov}(x_p, x_1) & \dots & \text{cov}(x_p, x_p) \end{bmatrix}$$

Por lo tanto:  $\text{var}(\vec{y}_k) = \vec{\beta}_k^T T \vec{\beta}_k = \vec{\beta}_k^T D \vec{\beta}_k + \vec{\beta}_k^T E \vec{\beta}_k$

Como se requiere que:  $\max_{\beta} \{ \vec{\beta}_k^T E \vec{\beta}_k \}$  en relación a la varianza total, esto es

$$\max_{\beta} \left\{ \frac{\vec{\beta}_k^T E \vec{\beta}_k}{\vec{\beta}_k^T T \vec{\beta}_k} \right\}$$



Si se considera la razón de varianzas una función

homogénea, entonces maximizar la razón  $\max_{\beta} \left\{ \frac{\vec{\beta}_k^T E \vec{\beta}_k}{\vec{\beta}_k^T T \vec{\beta}_k} \right\}$  es equivalente a:

$$\max_{\beta} \{ \vec{\beta}_k^T E \vec{\beta}_k \} \quad \text{sujeto a} \quad \vec{\beta}_k^T T \vec{\beta}_k = 1$$

Al igual que el caso de las componentes principales, se obtiene el lagrangeano aumentado, aplicando multiplicadores de Lagrange:

$$L(\vec{\beta}_k) = \vec{\beta}_k^T E \vec{\beta}_k + \lambda (\vec{\beta}_k^T T \vec{\beta}_k - 1) \quad \text{se obtiene } \beta_k \text{ de } \frac{\partial L(\vec{\beta}_k)}{\partial \vec{\beta}_k} = 0$$

$$2E\vec{\beta}_k - 2\lambda T\vec{\beta}_k = 0 \quad E\vec{\beta}_k = \lambda T\vec{\beta}_k \Rightarrow$$

Si pre multiplicamos por  $\vec{\beta}_k^T$  y considerando que  $\vec{\beta}_k^T T \vec{\beta}_k = 1$

$$\vec{\beta}_k^T E \vec{\beta}_k = \lambda \vec{\beta}_k^T T \vec{\beta}_k = \lambda$$



Si se considera el mayor vector característico asociado al mayor valor característicos  $\lambda_1$ , se tendrá el máximo *poder discriminante*.

El valor característico  $\lambda_i$  asociado a la función discriminante  $y_i$ , indica la proporción de la varianza total explicada por las  $m$  funciones discriminantes.

Para obtener mas funciones se continua obteniendo los vectores característicos de la matriz  $\mathbf{T}^T \mathbf{E}$  asociado a los valores característicos elegidos en orden decreciente.

La suma de los valores característicos  $\sum_{j=1}^q \lambda_j$  hasta la  $q$  corresponde a la varianza explicada por estas funciones.

Así el porcentaje de varianza explicado por cada  $y_i$  del total de la varianza hasta  $q$  será:

$$100 \frac{\lambda_i}{\sum_{i=1}^q \lambda_j}$$



### - Evaluación análisis discriminante lineal

Analizar los estadísticos:

- $F$  de Snedecor para análisis discriminante.

– $\lambda$  de Wilks denominado también U-estadístico.

-Matriz de confusión.



#### 4.2.1. Tablas de contingencia análisis ROC

Considere un *clasificador simple* cuyo objetivo es clasificar el patrón  $\bar{x}$  como perteneciente a una clase o no.

La respuesta del clasificador será  $T^+$  para clasificar el patrón en la clase  $C$  y será  $T^-$  para el caso en que no pertenece a la clase  $C$  (se clasifica como  $\bar{C}$ ).

En estas condiciones pueden existir dos tipos diferentes de errores, los cuales se observan en la siguiente tabla de doble entrada, denominada tabla de contingencia.



		<i>Realidad</i>		<i>Total</i>
<i>Clasificado</i>	$T^+$	$VP$	$FP$	$VP+FP$
	$T^-$	$FN$	$VN$	$FN+VN$
	<i>Total</i>	$VP+FN$	$FP+VN$	$n$





Verdaderos

*VP*: Verdaderos Positivos

*VN*: Verdaderos Negativos

Errores

*FN*: Falsos Negativos (tipo I)

*FP*: Falsos Positivos (tipo II)

La exactitud total del modelo es la cantidad de verdaderos dividido por el total:

$$\text{Exactitud} = (VP + VN) / n$$

$$\text{Error} = (FP + FN) / n$$




Los indicadores de las *bondades del sistema* se obtienen calculando proporciones por columnas (características del clasificador)

Las *predicciones del sistema* se obtienen por filas (características de las predicciones)

Clasificado	Control		
	<i>T<sup>+</sup></i>	<i>VP</i> <i>FP</i>	$VP / (VP + FP)$
	<i>T<sup>-</sup></i>	<i>FN</i> <i>VN</i>	$VN / (FN + VN)$
		$VP / (VP + FN)$ $FP / (FP + VN)$	
		$FN / (VP + FN)$ $VN / (FP + VN)$	



Las proporciones de interés son las siguientes



Nombre	Significado	Probab.	Estimación
<b>Sensibilidad</b> (Prop. $VP$ )	Frecuencia de los positivos de la clase $C$	$p[T^+ C]$	$VP/(VP+FN)$
Pro. Fal. Neg. (Prop. $FN$ )	Frecuencia de los negativos de la clase $C$	$p[T C]$	$FN/(VP+FN)$
<b>Especificidad</b> (Prop. $VN$ )	Frecuencia de los negativos de la clase $\bar{C}$	$p[T \bar{C}]$	$VN/(FP+VN)$
Pro. Fal. Pos. (Prop. $FP$ )	Frecuencia de los Positivos de la clase $\bar{C}$	$p[T^+ \bar{C}]$	$FP/(FP+VN)$
<b>Valor Predictivo Positivo</b> (VPP)	Frecuencia de la clase $C$ con resultados positivos del Sis.	$p[C T^+]$	Requiere $P[C]$ $VP/(VP+FP)$ Requiere $P[C]$
<b>Valor Predictivo Negativo</b> (VPN)	Frecuencia de la clase $\bar{C}$ con resultados negativos del Sis.	$p[\bar{C} T]$	$VN/(FN+VN)$
<b>Prevalencia</b>	Frecuencia de la clase $C$ en la población total ( $U$ )	$P[C]$	Evaluación Independiente

**Nota:** Prop.  $FN = P[T|C] = 1 - \text{Sensibilidad}$   
 Prop.  $FP = P[T^+|\bar{C}] = 1 - \text{Especificidad}$

### **Características del Sistema Clasificador** (proporciones en columnas)

El mejor clasificador es aquel en que los falsos son cero  
 $\Rightarrow \text{Sensibilidad} = \text{Especificidad} = 1$

La *Sensibilidad* indica la bondad del clasificador para detectar los casos que pertenecen a la clase  $C$ .

La *Especificidad* indica la bondad del clasificador para detectar los casos que no pertenecen a la clase  $C$  (esto es  $\in \bar{C}$ ).

Estas características indican las bondades del clasificador pero no aportan indicación de la probabilidad que tiene un nuevo caso de ser clasificado correctamente en el futuro.

### Características de Predicción (proporciones en filas)

El  $VPP$  ( $\hat{p}(C/T^+)$ ) y  $VPN$  ( $\hat{p}(\bar{C}/T^-)$ ) se pueden estimar por las proporciones de las filas de la tabla de contingencia pero estos valores corresponden a probabilidades sólo cuando la muestra es similar a la población.

Si se cuenta con la *prevalencia*  $P[C]$ , es posible corregir estas medidas.

Usando la definición de probabilidad condicional

$$P[C|T^+] = P[C \cap T^+] / P[T^+]; \quad P[\bar{C}|T^-] = P[\bar{C} \cap T^-] / P[T^-]$$

y la probabilidad de  $P[T^+] = P[C \cap T^+] + P[\bar{C} \cap T^+]$

$$P[T^+] = P[T^+|C]P[C] + P[T^+|\bar{C}]P[\bar{C}]$$

$$P[T^-] = P[T^-|C]P[C] + P[T^-|\bar{C}]P[\bar{C}]$$



Se puede calcular estas probabilidades en función de las proporciones del clasificador

$$VPP = P[C|T^+] = \frac{P[T^+|C]P[C]}{P[T^+|C]P[C] + P[T^+|\bar{C}]P[\bar{C}]} = \frac{1}{1 + \left( \frac{P[\bar{C}]}{P[C]} \frac{(1 - \text{Especificidad})}{\text{Sensibilidad}} \right)}$$

$$VPN = P[\bar{C}|T^-] = \frac{P[T^-|\bar{C}]P[\bar{C}]}{P[T^-|C]P[C] + P[T^-|\bar{C}]P[\bar{C}]} = \frac{1}{1 + \left( \frac{P[\bar{C}]}{P[C]} \frac{\text{Especificidad}}{(1 - \text{Sensibilidad})} \right)^{-1}}$$

Aquí se observa claramente que estas probabilidades dependen de la *prevalencia* en la población.

$$P[C] = 1 - P[\bar{C}].$$

De esta forma el  $VPP$  determina la probabilidad de que un sujeto pertenezca a la clase  $C$  dado que el sistema lo clasificó como  $T^+$ .



### ➤ La Curva de Calibración.

(Receiver-Operating Characteristic, ROC)

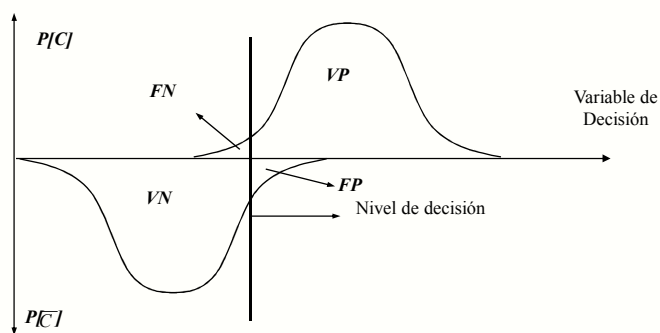
Todo el desarrollo anterior fue realizado considerando que el sistema de clasificación decide por  $T^+$  o  $T^-$  en un nivel fijo.

Supóngase ahora que para decidir la clasificación de un patrón se tiene en el sistema un parámetro  $q$  (por ejemplo a una probabilidad que varía entre  $[0-1]$ ), que al variar produce un cambio en la clasificación del sistema.

Si  $\theta=0,5$  se obtendrán valores de  $VP$ ,  $FP$ ,  $FN$  y  $VN$ , si se varía  $\theta=0.6$ , se obtendrá otra tabla de valores, la cual puede resultar una clasificación mejor que la anterior, logrando así una especificidad y sensibilidad mayor.



Una representación para ver como varían los cambios al variar el nivel de decisión es:



Para tener una representación más adecuada se puede recurrir a la tabla de contingencia.



De las características del test (columnas) se sabe que:

$$\begin{aligned} \text{Sensibilidad} + \text{Prop. FP} &= 1 \\ \text{especificidad} + \text{Prop. FN} &= 1. \end{aligned}$$

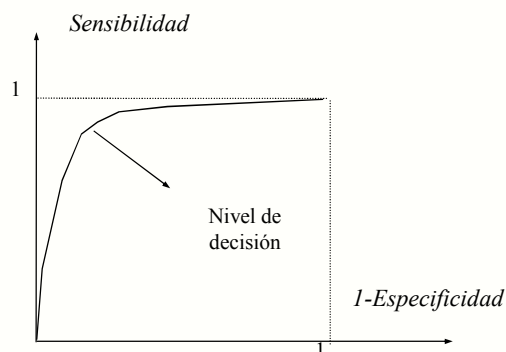
Entonces sólo se requiere dos de estas variables para representar el test y cada nivel de decisión será un punto en el plano.

Al variar continuamente el nivel de decisión se tiene una curva que representa las incidencias del nivel sobre el clasificador.

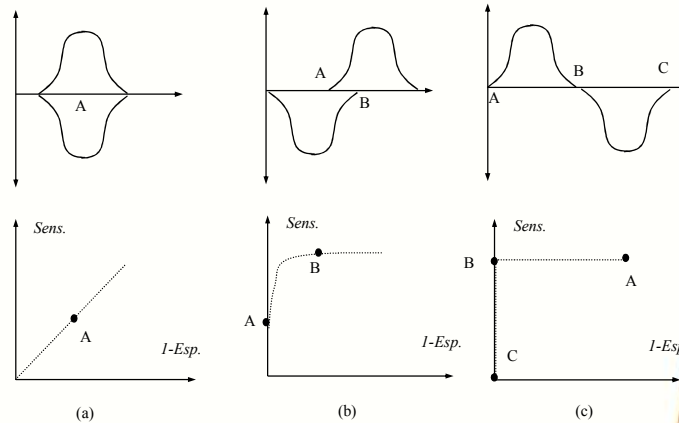
Se usa como variables normalizadas la *Sensibilidad* en las ordenadas y la *Prop. FN* o *1- Especificidad* en las abscisas.



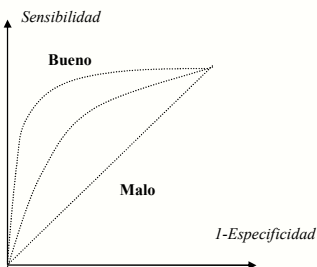
### Curva característica de calibración del sistema (Curva ROC)



Para cada posición de la distribución de probabilidades se tendrán diferentes curvas a medida que se desplaza el nivel de decisión.



La curva (c) corresponde al clasificador perfecto y el nivel de decisión se debe situar en el punto B. La curva (a) no posee ninguna utilidad y el de la curva (b) corresponde a un clasificador real.



El problema es determinar para una curva dada ¿cual es el nivel de decisión ideal?

En general será el punto donde se maximiza la *Sensibilidad* y la *Especificidad*.

En una curva real, en general, el mejor punto de operación se logra en el punto de máxima curvatura.

Pero se puede observar que existe un compromiso entre la *Sensibilidad* y *Especificidad*, esto es, se puede aumentar una en perjuicio de la otra.

Este compromiso dependerá del tipo de clasificación que se requiere realizar.

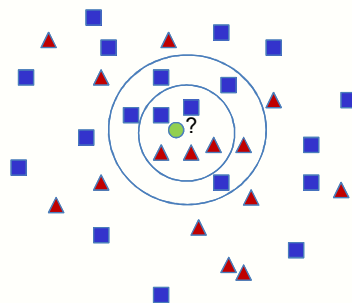
Suponga un sistema que realiza diagnóstico de una enfermedad (examen para detecta una infección) para la cual el tratamiento en pacientes sanos es inocuo, se puede privilegiar la *Sensibilidad* antes de la *Especificidad*.

En cambio, si el tratamiento para la enfermedad que diagnostica el sistema es muy riesgoso para el paciente, es necesario balancear la *Sensibilidad* y *Especificidad*.



### 4.3. Métodos no paramétricos (K-Vecinos mas Cercanos; K Nearest Neighbours)

**Idea Básica:**



Circulo 1:  $k=5$  asignación ▲

Circulo 2:  $k=10$  asignación ■



### El algoritmo solo posee dos etapas:

#### - Etapa de entrenamiento:

Se requiere un conjunto de entrenamiento suficiente de ejemplos previamente etiquetados. Cada  $\langle \vec{x}, f(\vec{x}) \rangle$  ejemplo posee la clasificación  $f(\vec{x}) \in \mathcal{G} = \{0, 1, \dots, m\}$

#### -Etapa de Clasificación:

Cada nuevo ejemplo  $\vec{x}_q$  es clasificado de acuerdo a la proximidad de los  $k$  vecinos obtenidos del grupo de entrenamiento.

$$f(\vec{x}_q) \leftarrow \arg \max_{e \in \mathcal{G}} \sum_{i=1}^k \delta(e, f(\vec{x}_i))$$

Donde  $\delta(a, b) = 1$ , si  $a = b$ , y 0 en otro caso.



Note que la elección de  $k$  determina la forma de clasificación.

- Si  $k$  es un valor pequeño, el algoritmo clasifica localmente de forma que el ruido también es incorporado en la clasificación

- Si  $k$  es un valor grande Se evita el ruido, pero se introduce sesgo, pues la mayoría de las veces, el resultado será la clase mayoritaria.

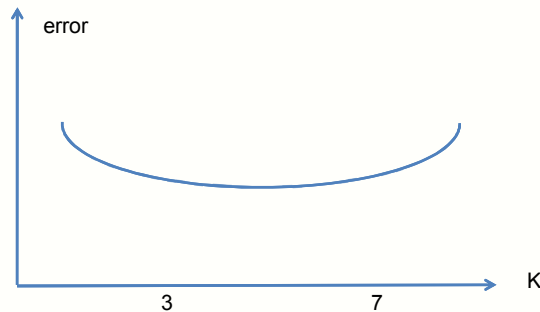
- Que pasa cuando existe la misma cantidad de vecinos? Se utilizan las probabilidades a priori (clase mayoritaria) si son iguales se escoge al azar.

- Que valor de  $K$  se usa?





Si se realiza una curva de error de clasificación, dependiendo del numero de vecinos se puede observar que un numero pequeños y muy grandes llevan a elevar el error.



Existen varias otras alternativas:

- Con rechazo: Se exigen garantías, umbral o mayoría absoluta.
- Distancia mínima: calcular la distancia solo a los casos mas cercanos al centriode de cada clase



### - K vecinos ponderado

Una forma simple controlar el numero de vecinos que participan en la elección de la clase es ponderar la clasificación del nuevo caso por un peso.

$$f(\vec{x}_q) \leftarrow \arg \max_{e \in \mathcal{G}} \sum_{i=1}^n \omega_i \delta(e, f(\vec{x}_i))$$

Donde el ponderador es la propia medida de proximidad o el inverso de la distancia:

$$\omega_i = \frac{1}{\|\vec{x}_q - \vec{x}_i\|}$$

De esta manera no es necesario determinar un numero específico de  $k$ , pues los ejemplos mas distantes no contribuirán significativamente a la determinación de la clase o promedio.

- Usar información mutua como ponderador.

