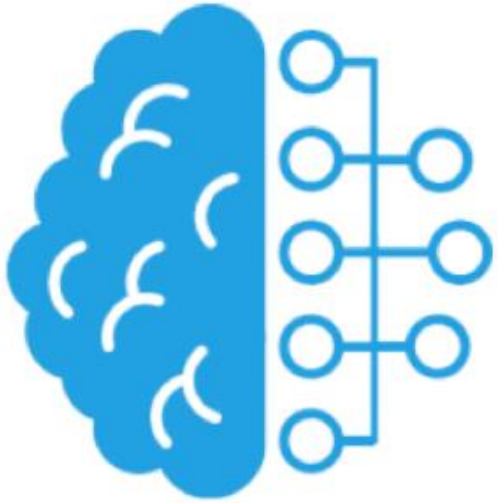
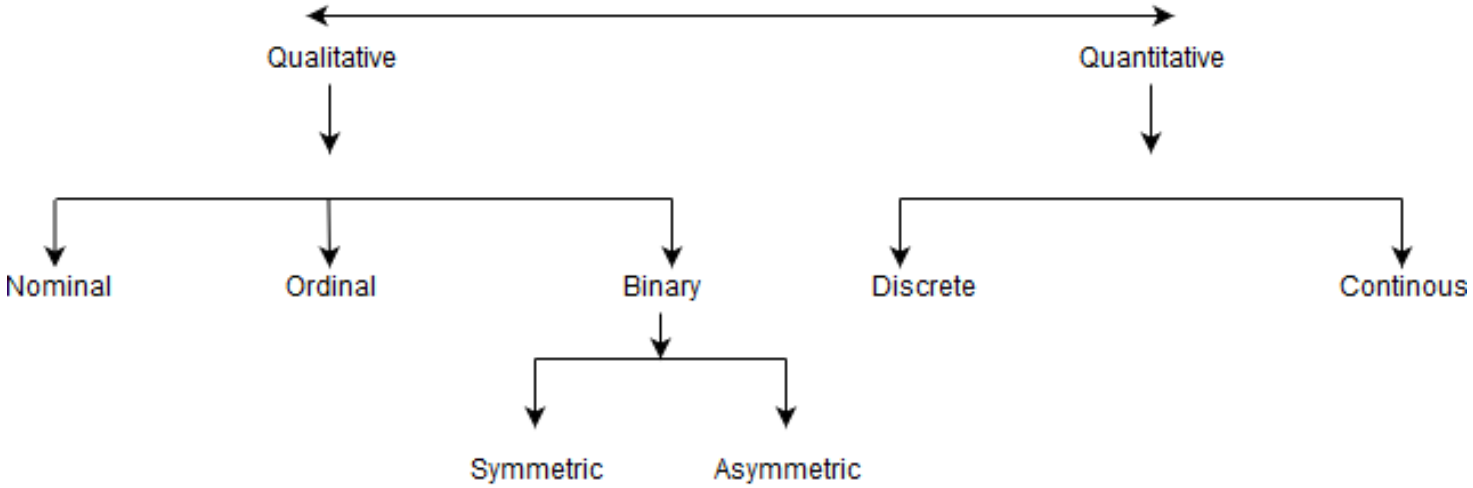
# Aprendizaje de Máquina
## Clase 2

Claudio Delrieux – DIEC - UNS
cad@uns.edu.ar



# Repaso de conceptos básicos

Tipos de atributos (escalares). Se pueden convertir de unos a otros cuando es necesario de acuerdo al tipo de análisis.



# Repaso de conceptos básicos

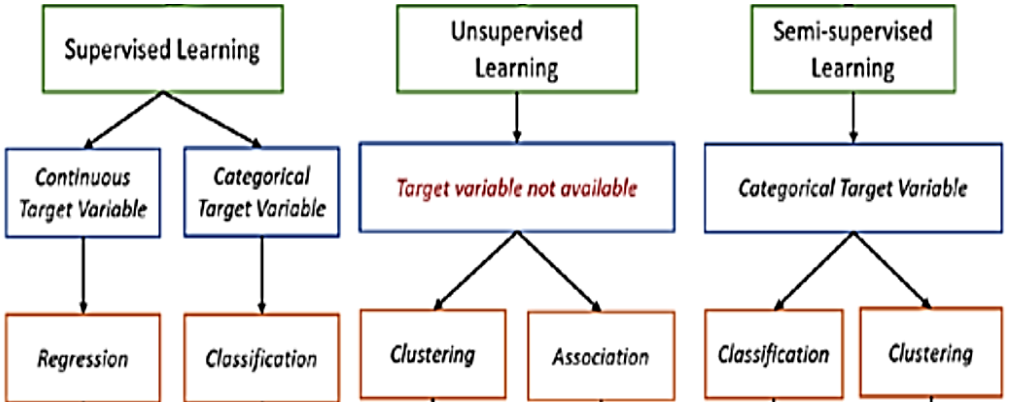
Atributos (variables) target y tipos de modelos.

Atributo

Target

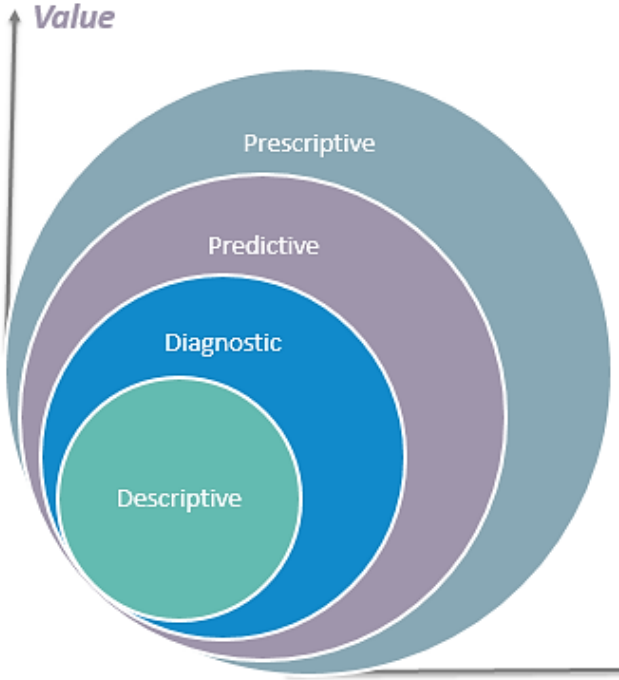
Registro

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Userid | Name | Age | Height | City | Gender | Device | Qualification | Industry |
| 2 | 10000 | WCOCM | 12 | 167 | Mumbai | Male | Tablet | Intermediate | Automobile |
| 3 | 10001 | JRBAL | 71 | 138 | Mumbai | Female | Playstation | PostGraduate | Agriculture |
| 4 | 10002 | DPGYN | 18 | 81 | Mumbai | Female | MobileAndroid | Graduate | Fashion |
| 5 | 10003 | PNSQX | 41 | 90 | Mumbai | Male | SmartTV | PostGraduate | Chemical |
| 6 | 10004 | WQOPX | 16 | 56 | Mumbai | Female | MobileAndroid | HighSchool | Legal |
| 7 | 10005 | VWHDV | 19 | 97 | Mumbai | Male | MobileAndroid | Graduate | Hardware |
| 8 | 10006 | DKOEZ | 35 | 111 | Pune | Female | SmartTV | PostGraduate | Construction |
| 9 | 10007 | CAQUF | 57 | 153 | Pune | Male | MobileAndroid | Graduate | Legal |
| 10 | 10008 | GFDTY | 72 | 90 | Noida | Female | MobileAndroid | PostGraduate | Legal |
| 11 | 10009 | CUFNN | 44 | 139 | Noida | Female | SmartTV | Graduate | Textile |
| 12 | 10010 | OKBHU | 39 | 172 | Pune | Female | Playstation | PostGraduate | Information |
| 13 | 10011 | OKOWE | 12 | 68 | Delhi | Male | Playstation | Intermediate | Energy |
| 14 | 10012 | MTQGW | 32 | 145 | Pune | Female | Playstation | PostGraduate | Construction |
| 15 | 10013 | NUNHT | 31 | 106 | Mumbai | Female | MobileAndroid | Graduate | FinancialServices |
| 16 | 10014 | TBIWC | 13 | 52 | Pune | Female | Desktop | Intermediate | Fashion |



# Repaso de conceptos básicos

## 4 types of Data Analytics




## What is the data telling you?

**Descriptive:** ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

**Diagnostic:** ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

**Predictive:** ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

**Prescriptive:** ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

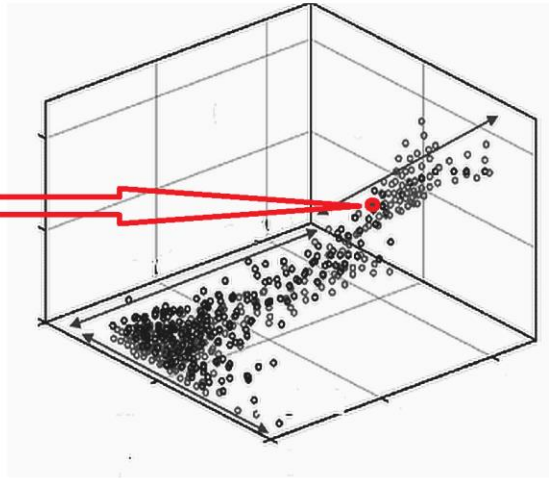
# Repaso de conceptos básicos

El espacio de atributos consiste en «vectorizar» cada registro.

**Atributo**

**Registro**

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Userid | Name | Age | Height | City | Gender | Device | Qualification | Industry |
| 2 | 10000 | WCOCM | 12 | 167 | Mumbai | Male | Tablet | Intermediate | Automobile |
| 3 | 10001 | JRBAL | 71 | 138 | Mumbai | Female | Playstation | PostGraduate | Agriculture |
| 4 | 10002 | DPGYN | 18 | 81 | Mumbai | Female | MobileAndroid | Graduate | Fashion |
| 5 | 10003 | PNSQX | 41 | 90 | Mumbai | Male | SmartTV | PostGraduate | Chemical |
| 6 | 10004 | WQOPX | 16 | 56 | Mumbai | Female | MobileAndroid | HighSchool | Legal |
| 7 | 10005 | VWHDV | 19 | 97 | Mumbai | Male | MobileAndroid | Graduate | Hardware |
| 8 | 10006 | DKOEZ | 35 | 111 | Pune | Female | SmartTV | PostGraduate | Construction |
| 9 | 10007 | CAQUF | 57 | 153 | Pune | Male | MobileAndroid | Graduate | Legal |
| 10 | 10008 | GFDTY | 72 | 90 | Noida | Female | MobileAndroid | PostGraduate | Legal |
| 11 | 10009 | CUFNN | 44 | 139 | Noida | Female | SmartTV | Graduate | Textile |
| 12 | 10010 | OKBHU | 39 | 172 | Pune | Female | Playstation | PostGraduate | Information |
| 13 | 10011 | OKOWE | 12 | 68 | Delhi | Male | Playstation | Intermediate | Energy |
| 14 | 10012 | MTQGW | 32 | 145 | Pune | Female | Playstation | PostGraduate | Construction |
| 15 | 10013 | NUNHT | 31 | 106 | Mumbai | Female | MobileAndroid | Graduate | FinancialServices |
| 16 | 10014 | TBIWC | 13 | 52 | Pune | Female | Desktop | Intermediate | Fashion |



# Repaso de conceptos básicos

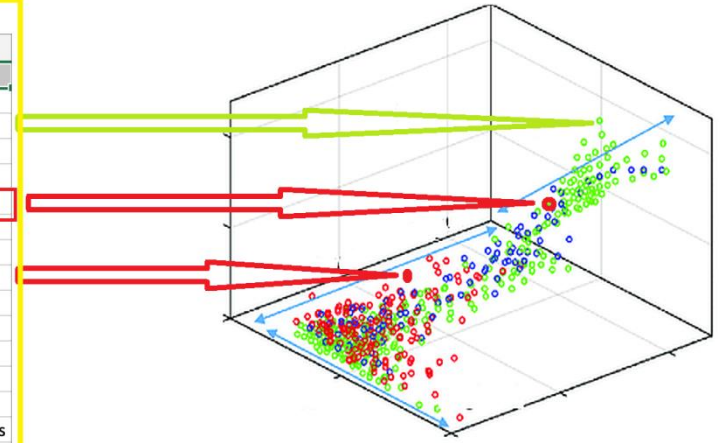
La expectativa es que esta vectorización permita extraer conocimiento de los datos.

**Atributo**

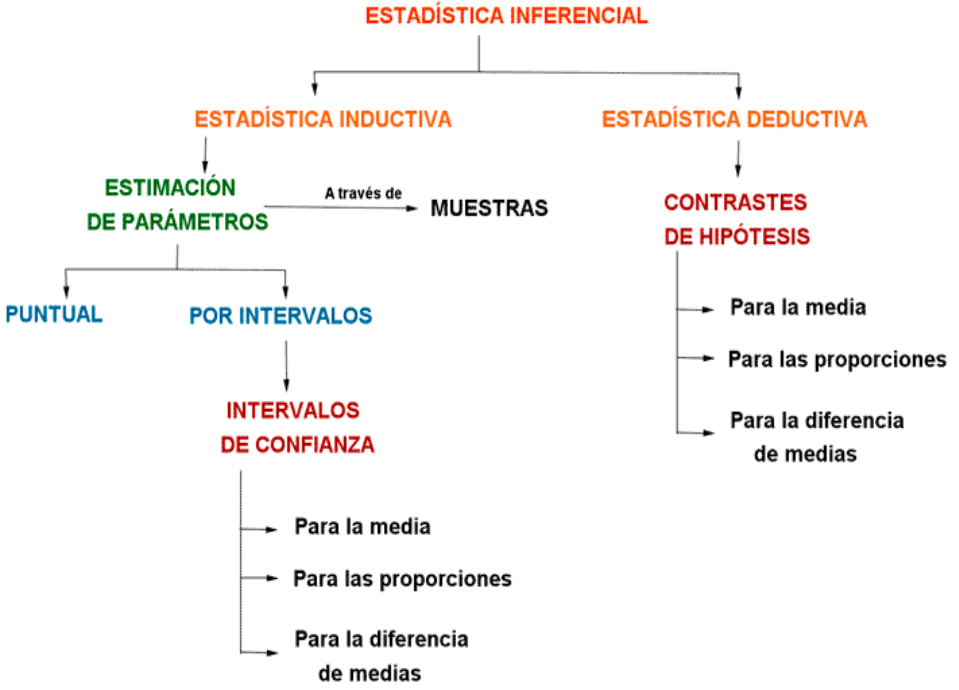
**Target cualitativo**

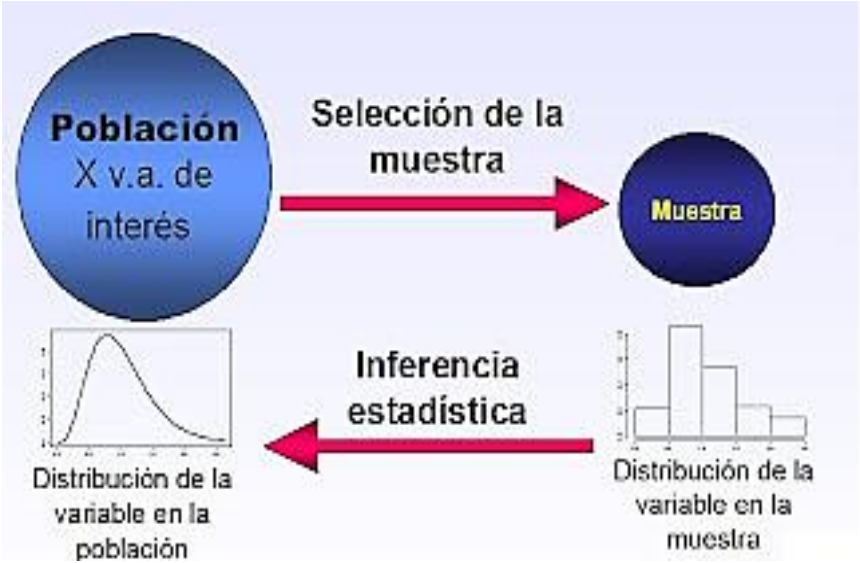
**Registro**

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Userid | Name | Age | Height | City | Gender | Device | Qualification | Industry |
| 2 | 10000 | WCOCM | 12 | 167 | Mumbai | Male | Tablet | Intermediate | Automobile |
| 3 | 10001 | JRBAL | 71 | 138 | Mumbai | Female | Playstation | PostGraduate | Agriculture |
| 4 | 10002 | DPGYN | 18 | 81 | Mumbai | Female | MobileAndroid | Graduate | Fashion |
| 5 | 10003 | PNSQX | 41 | 90 | Mumbai | Male | SmartTV | PostGraduate | Chemical |
| 6 | 10004 | WQOPX | 16 | 56 | Mumbai | Female | MobileAndroid | HighSchool | Legal |
| 7 | 10005 | VWHDV | 19 | 97 | Mumbai | Male | MobileAndroid | Graduate | Hardware |
| 8 | 10006 | DKOEZ | 35 | 111 | Pune | Female | SmartTV | PostGraduate | Construction |
| 9 | 10007 | CAQUF | 57 | 153 | Pune | Male | MobileAndroid | Graduate | Legal |
| 10 | 10008 | GFDTY | 72 | 90 | Noida | Female | MobileAndroid | PostGraduate | Legal |
| 11 | 10009 | CUFNN | 44 | 139 | Noida | Female | SmartTV | Graduate | Textile |
| 12 | 10010 | OKBHU | 39 | 172 | Pune | Female | Playstation | PostGraduate | Information |
| 13 | 10011 | OKOWE | 12 | 68 | Delhi | Male | Playstation | Intermediate | Energy |
| 14 | 10012 | MTQGW | 32 | 145 | Pune | Female | Playstation | PostGraduate | Construction |
| 15 | 10013 | NUNHT | 31 | 106 | Mumbai | Female | MobileAndroid | Graduate | FinancialServices |
| 16 | 10014 | TBIWC | 13 | 52 | Pune | Female | Desktop | Intermediate | Fashion |



# Repaso de conceptos básicos: estadística

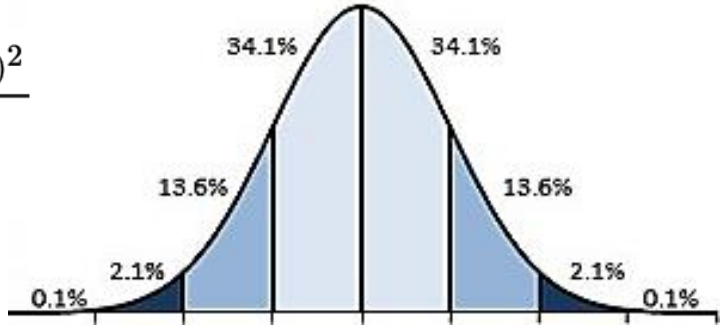


# Repaso de conceptos básicos

**Estimadores univariados**: tendencia central, dispersión, momentos.
**Estadistica paramétrica**. Teorema del límite central. Distribución normal. Kernel. Estandarización. Normalización. Otras distros.
**Estadística no paramétrica**. Percentiles. Rango intercuartil.

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$




Ver: https://github.com/manlio99/Materia-de-aprendizaje/blob/master/4_DataWrangling/8_outlier_detection.ipynb

# Clasificación Estadística Paramétrica

Nuestro objetivo es predecir una categoría o clase a partir de la distribución de probabilidades conocida de los atributos de cada clase.

Por lo tanto, es necesario conocer *a priori* la *prevalencia* de los posibles valores de la variable **nominal** a predecir, y conocer las distribuciones de probabilidad de los atributos en cada una de las clases.

Para simplificar la notación denominamos $w_i$ a las $i$ clases nominales, y $\boldsymbol{x}$ al vector de atributos (variables independientes).

# Clasificación Estadística Paramétrica

Algunas definiciones previas:

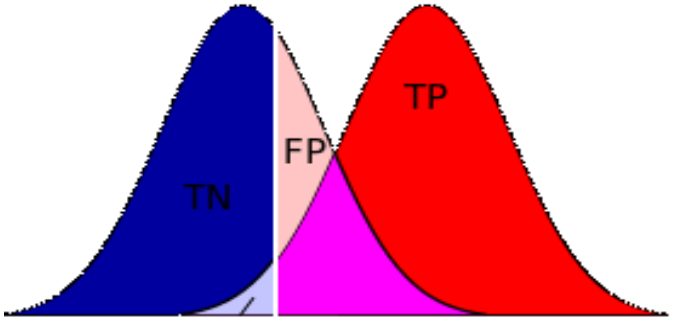
- $p(\omega_i)$ es la probabilidad *a priori* de ocurrencia de un evento de la clase $\omega_i$. Normalmente estas probabilidades se conocen o se pueden estimar.
- $p(\mathbf{x}|\omega_i)$ es la probabilidad condicional, o función de densidad probabilística, (también "verosimilitud") de que se observe un patrón $\mathbf{x}$ cuando el evento es de la clase $\omega_i$.
- $p(\omega_i|\mathbf{x})$ es la probabilidad *a posteriori*, de que el evento sea perteneciente a la clase $\omega_i$ cuando el patrón observado fue $\mathbf{x}$.

# Clasificación Estadística Paramétrica: Regla del mínimo error

la regla de decisión correspondiente resulta:

$$\text{Asignar } \mathrm{x} \rightarrow \omega_j \quad \Longleftrightarrow \quad P(\omega_j|\mathrm{x}) = \max_{1 \leq k \leq m} P(\omega_k|\mathrm{x})$$

Observar que si asumimos que se asigna $\mathrm{x} \rightarrow \omega_j$, la probabilidad condicional de error $\epsilon(\mathrm{x})$ viene dada por



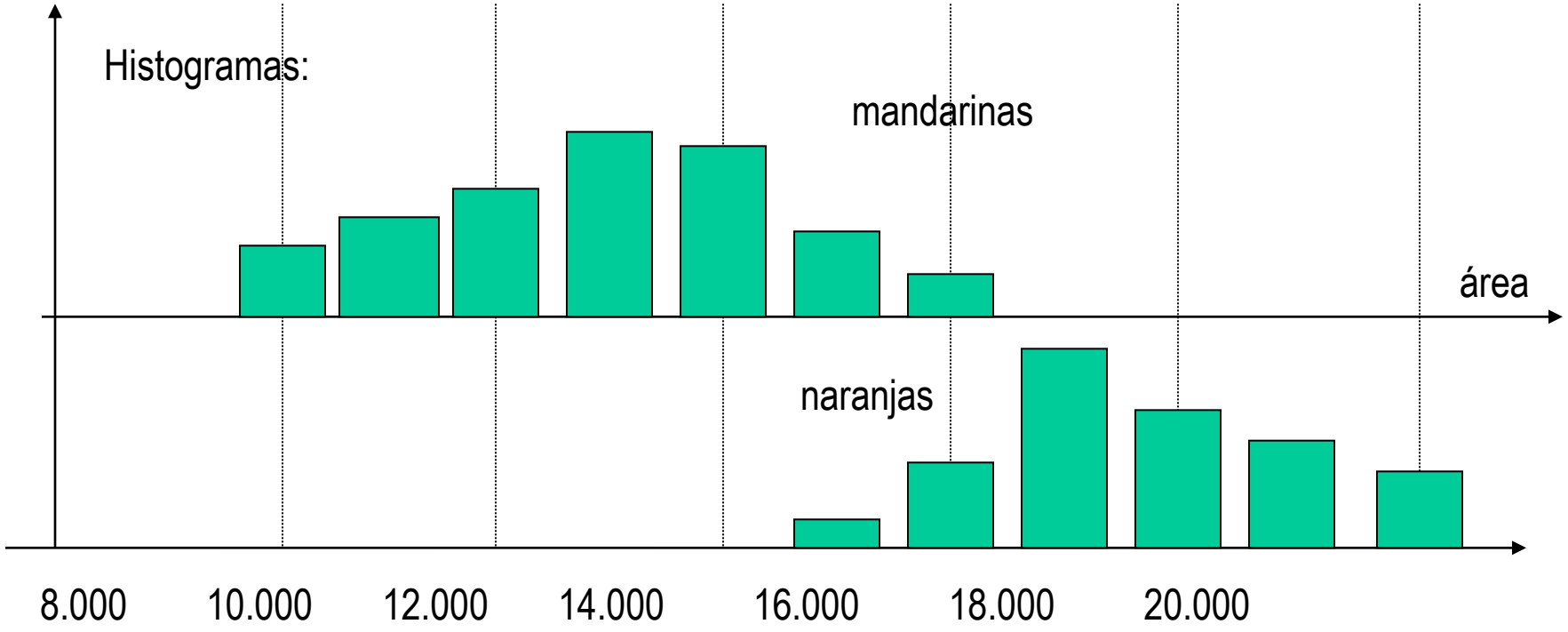

$$\epsilon(\mathrm{x}) = 1 - P(\omega_j|\mathrm{x})\,.$$

# Clasificación Estadística Paramétrica: un ejemplo

Determinar si es naranja o mandarina de acuerdo con el tamaño (área)

| | | | |
|---|---|---|---|
| Naranja-01 | 19.327 | Mandarina-01 | 13.221 |
| Naranja-02 | 18.265 | Mandarina-02 | 14.987 |
| Naranja-03 | 17.456 | Mandarina-03 | 15.321 |
| Naranja-04 | 19.341 | Mandarina-04 | 15.987 |
| Naranja-05 | 16.342 | Mandarina-05 | 16.345 |
| Naranja-06 | 16.987 | Mandarina-06 | 15.965 |
| Naranja-07 | 17.001 | Mandarina-07 | 16.341 |
| : | 19.056 | : | |
| Naranja-75 | 15.900 | Mandarina-50 | 13.439 |

# Clasificación Estadística Paramétrica: un ejemplo



# Medidas de Evaluación de Modelos

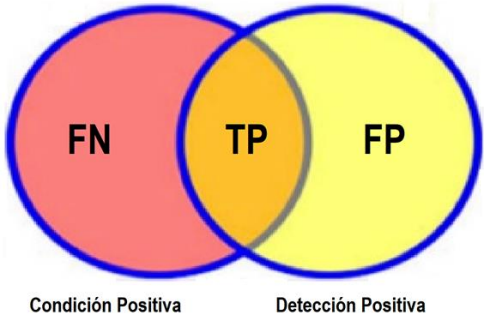
| | | predicted condition | | | |
|---|---|---|---|---|---|
| | total population | prediction positive | prediction negative | Prevalence $= \frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$ | |
| true condition | condition positive | **True Positive (TP)** | **False Negative (FN)** (type II error) | True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection $= \frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$ | False Negative Rate (FNR), Miss Rate $= \frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$ |
| | condition negative | **False Positive (FP)** (Type I error) | **True Negative (TN)** | False Positive Rate (FPR), Fall-out, Probability of False Alarm $= \frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$ | True Negative Rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$ |
| | Accuracy $= \frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$ | Positive Predictive Value (PPV), Precision $= \frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$ | False Omission Rate (FOR) $= \frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$ | Positive Likelihood Ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$ | Diagnostic Odds Ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ |
| | | False Discovery Rate (FDR) $= \frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$ | Negative Predictive Value (NPV) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$ | Negative Likelihood Ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$ | |

# Medidas de Evaluación de Modelos

No todos estos parámetros son importantes (y hay además otros parámetros importantes más). Los más utilizados son TPR (sensitivity, recall), precisión (poder predictivo positivo), y exactitud (accuracy). En el contexto de la detección, o cuando las clases no son simétricas o balanceadas, los TN son muy frecuentes y por lo tanto tenerlos en cuenta no aporta. Por eso se utilizan

f-measure = (2*precisión*recall) / (precisión + recall)

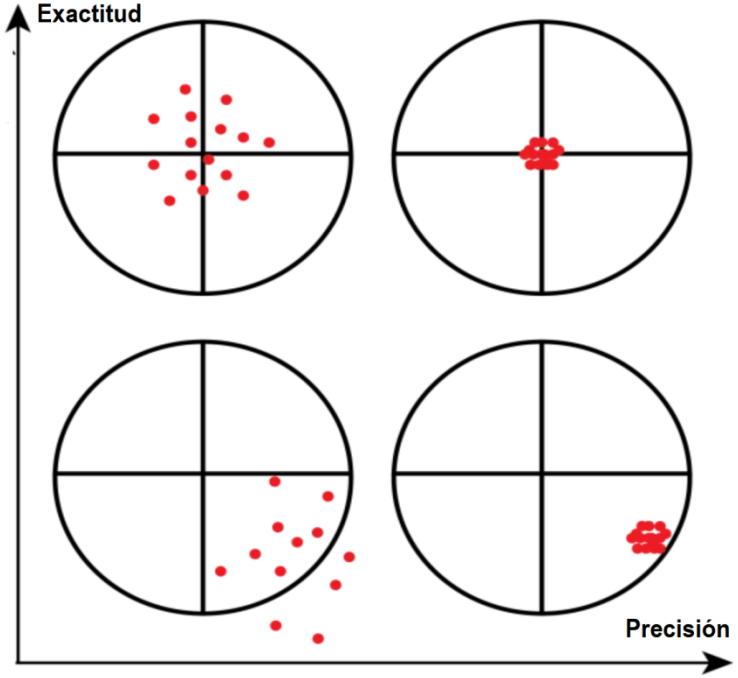
Índice Jaccard, o IoU = TP / (TP+FP+FN)



# Medidas de Evaluación de Modelos

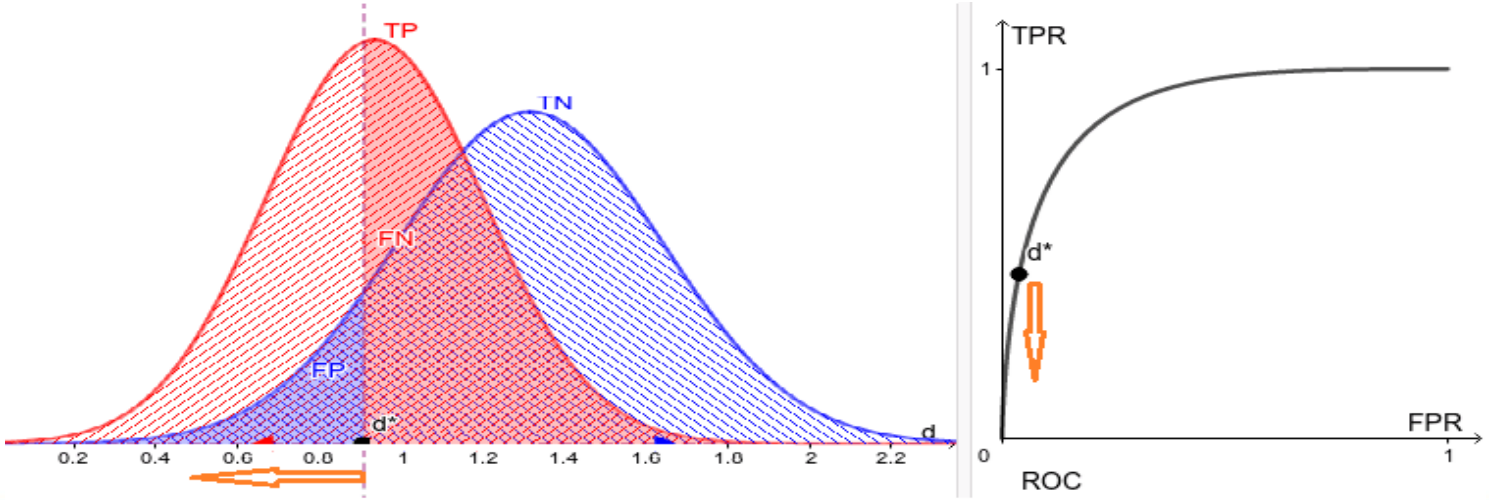
Es importante distinguir exactitud (accuracy) y precisión, y ponerlas en términos de sus «opuestas» paramétricas (sesgo y variancia).

También, en general el sesgo en un modelo (o en un proceso) es más fácil y rápido de corregir que la variancia.



# Medidas de Evaluación de Modelos: Curva ROC

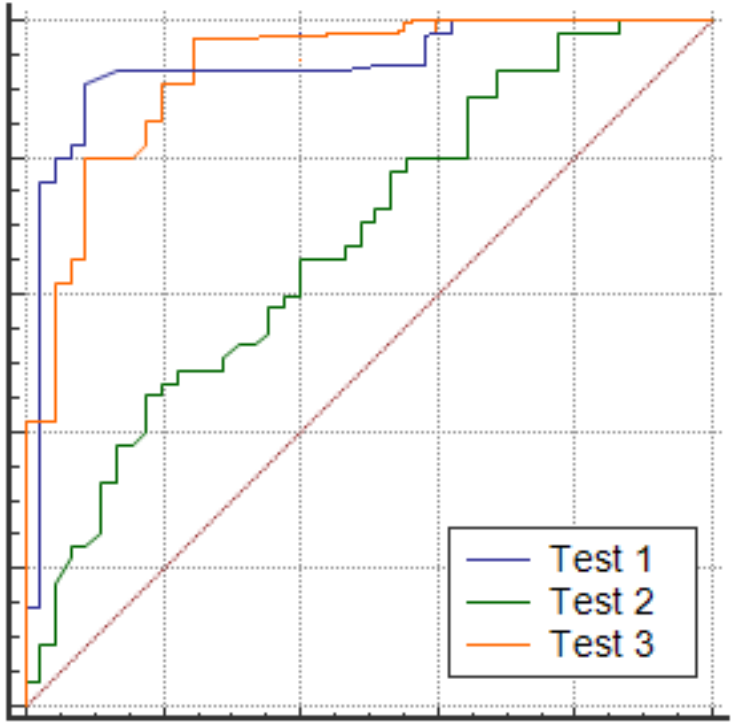
En varios contextos es necesario no utilizar el valor óptimo de la regla del mínimo error, sino que movemos el valor «umbral» de la variable en función de otros aspectos (fundamentalmente la *prevalencia* y el *costo* del error).



# Medidas de Evaluación de Modelos: Curva ROC

La recta identidad se corresponde con la clasificación puramente aleatoria. Cualquier clasificador que se aparte de esta recta aporta información (aún los que estén debajo).

La curva ROC nos permite comparar varios posibles clasificadores, y eventualmente combinarlos entre si.

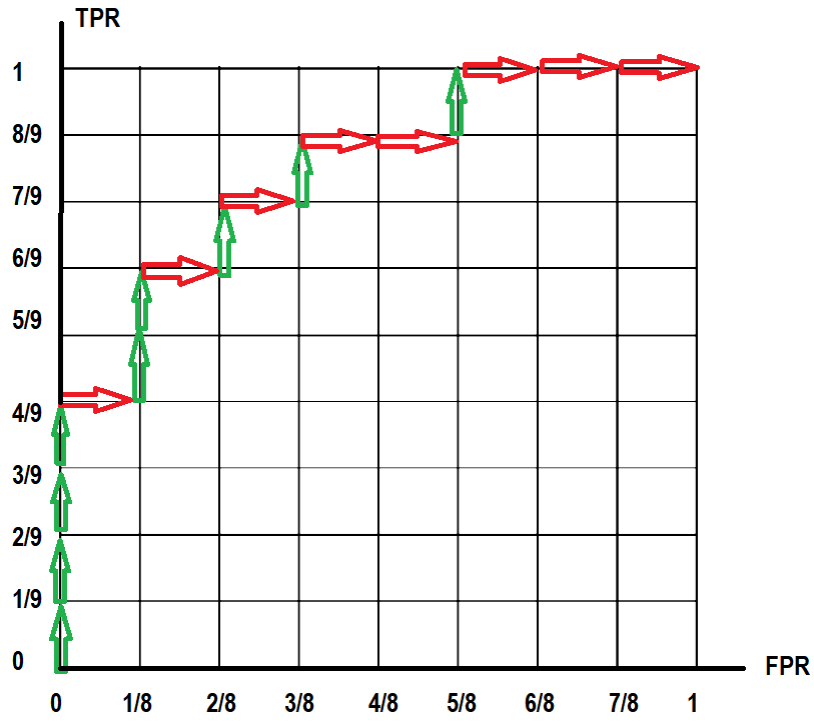


# Medidas de Evaluación de Modelos: Curva ROC

En la práctica, con un dataset de T casos positivos más F casos negativos, la curva se computa fácilmente. Se divide el eje X (FPR) en T intervalos iguales, el eje Y (TPR) en F intervalos iguales.

Comenzando desde abajo en la tabla y en el gráfico nos desplazamos una celda hacia arriba por cada V y una celda hacia la derecha por cada F.

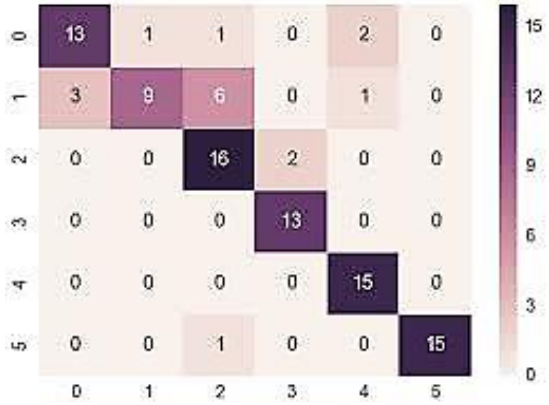
| Valor | Clasif. |
|---|---|
| 2 | F |
| 5 | F |
| 8 | F |
| 11 | V |
| 14 | F |
| 18 | F |
| 22 | V |
| 29 | F |
| 33 | V |
| 39 | F |
| 47 | V |
| 56 | V |
| 68 | F |
| 82 | V |
| 89 | V |
| 91 | V |
| 99 | V |



# Medidas de Evaluación de Modelos

En clasificadores de más de dos clases, la evaluación de modelos se generaliza a la matriz de contingencia. Se pueden analizar los valores de TP, TN, FP y FN para cada clase y encontrar valores medios para éstos y para los demás parámetros derivados.

También podemos tener medidas globales, por ejemplo la exactitud global es la suma de la diagonal principal sobre la suma de toda la matriz.



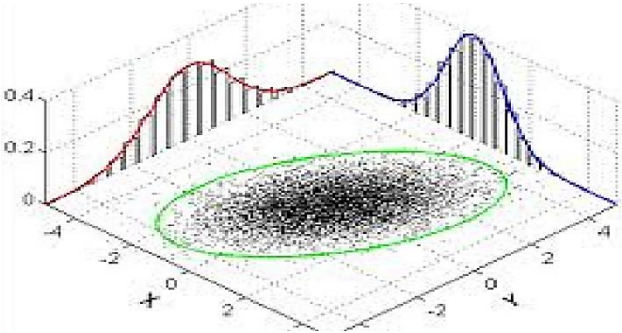
# Clasificación Estadística Paramétrica Multivariada

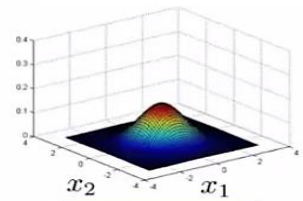
Supongamos que las clases responden a distribuciones normales, o sea que las densidades de probabilidad condicionales de cada clase tienen la forma

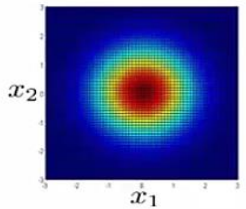
$$p(\mathrm{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathrm{x}-\mu_i)\Sigma_i^{-1}(\mathrm{x}-\mu_i)^T\right) \quad \forall\, i = 1, \cdots, m\,.$$

$$\mu_i = E\left[\mathrm{x}|\omega_i\right]$$

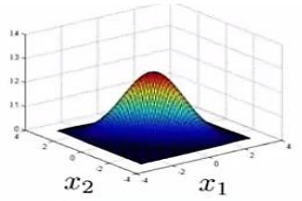
$$\Sigma_i = \mathrm{Cov}[\mathrm{x}|\omega_i] = E\left[(\mathrm{x}-\mu_i)(\mathrm{x}-\mu_i)^T|\omega_i\right]$$

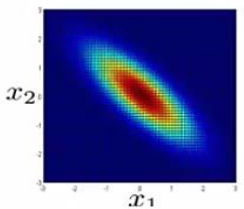



$$\mu = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$



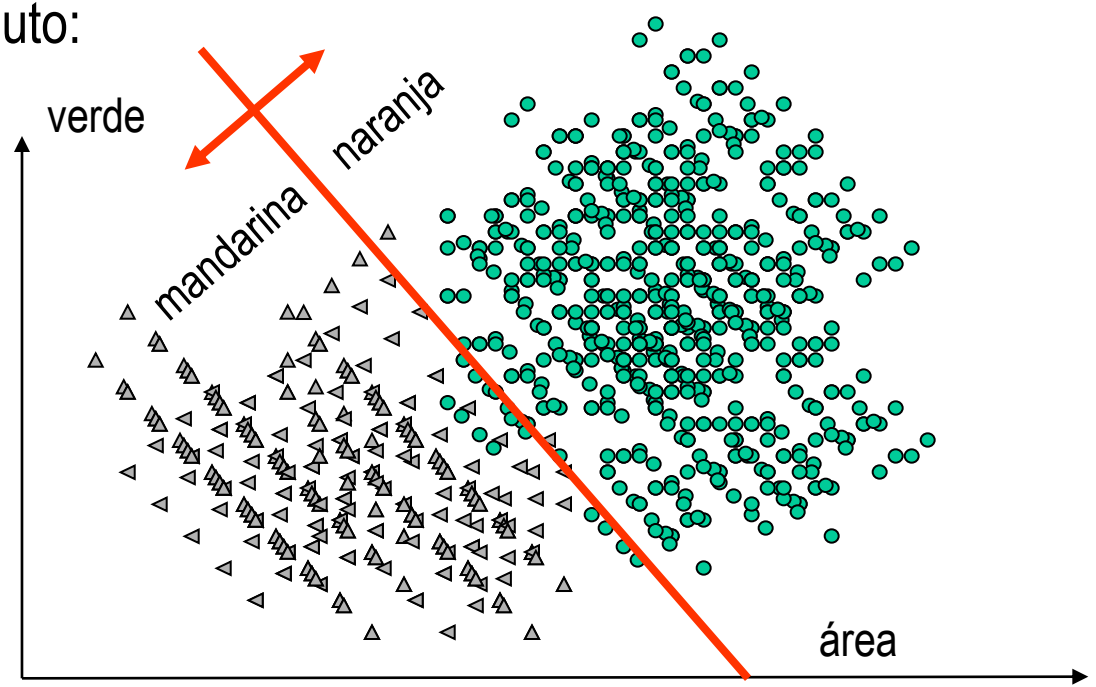

$$\mu = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}$$

# Clasificación Estadística Paramétrica Multivariada

Podemos tener más de un atributo:

Verde = 23.6%

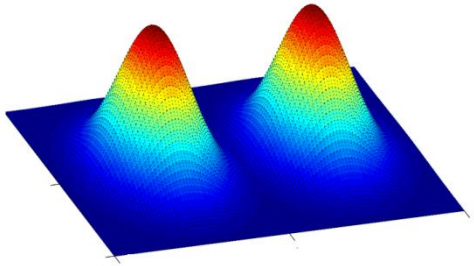
Verde = 46%



# Clasificación Estadística Paramétrica Multivariada

Supongamos que las probabilidades a priori y las covarianzas son constantes en las clases

$$\Sigma_i = \Sigma_j = \Sigma \quad \text{y} \quad P(\omega_i) = P(\omega_j) \quad \forall 1 \leq i \neq j \leq m$$



Definimos la norma y distancias asociadas a $\Sigma$ como antes:

$$\|\mathbf{x}\|^2 = \mathbf{x}\Sigma^{-1}\mathbf{x}^T \quad \Rightarrow \quad d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\| \quad \forall\, \mathbf{x}, \mathbf{x}' \in \Omega$$

Esta distancia se conoce como *distancia de Mahalanobis* asociada a la covarianza $\Sigma$. Notar que en el caso particular que $\Sigma = \mathbf{I}$ está distancia coincide con la distancia cuadrática Euclidiana.

Por lo tanto la regla del mínimo error se reduce a asignar el patrón $\mathbf{x}$ a la clase cuyo vector medio sea el más cercano según la distancia de Mahalanobis (*Regla de decisión por media más cercana*):

$$\text{Asignar } \mathbf{x} \rightarrow \omega_j \quad \Longleftrightarrow \quad d(\mathbf{x}, \boldsymbol{\mu}_j) = \min_{1 \leq k \leq m} d(\mathbf{x} - \boldsymbol{\mu}_k)$$

# Clasificación Estadística Paramétrica Multivariada

Ahora analizaremos el problema general de decisión entre dos clases

$$\Rightarrow \quad m = 2, \quad \Sigma_1 \neq \Sigma_2$$

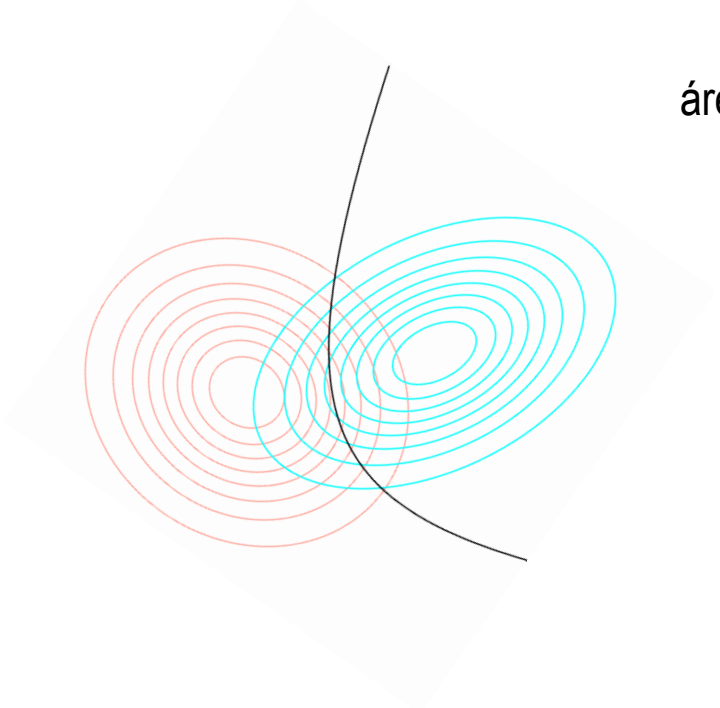
la ecuación

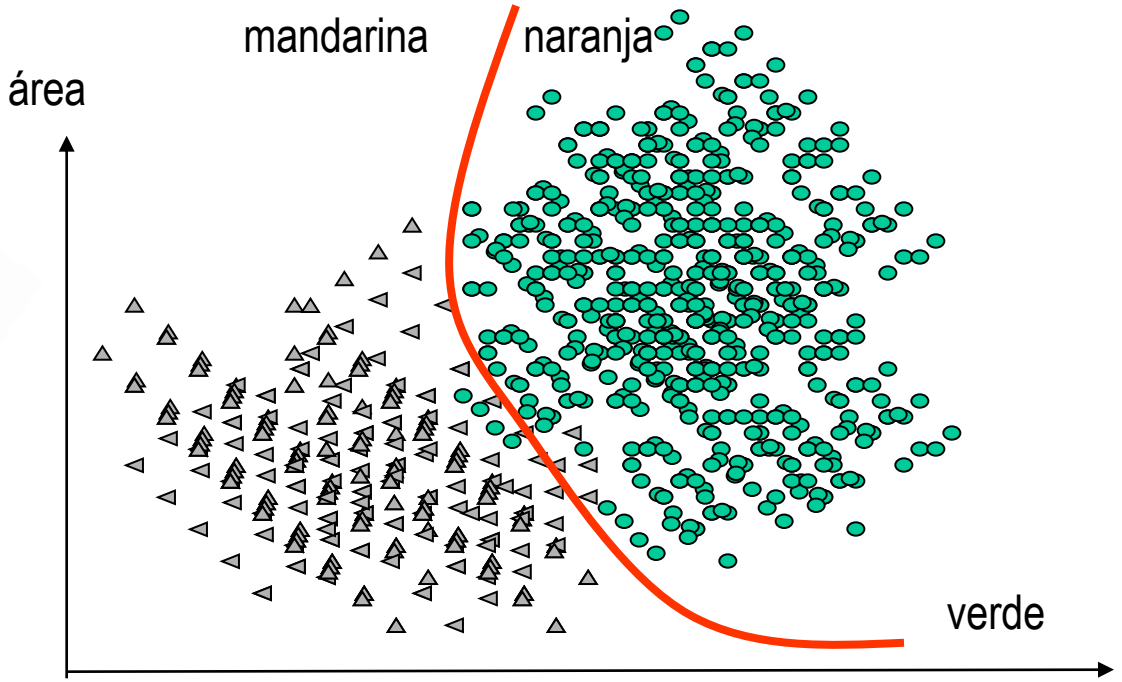
$$f(\mathrm{x}) = \|\mathrm{x} - \boldsymbol{\mu}_1\|_1 - \|\mathrm{x} - \boldsymbol{\mu}_2\|_2 + C_2 - C_1 = 0$$

define la superficie de separación, conocida como *superficie discriminante*, entre las regiones asociadas a cada una de las clases $\omega_1$ y $\omega_2$. En general esta superficie es cuadrática ya que su ecuación resulta:

$$\underbrace{\mathrm{x}^T\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)\mathrm{x}}_{\text{cuadrático}} - \underbrace{2\mathrm{x}^T\left(\Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_2^{-1}\boldsymbol{\mu}_2\right)}_{\text{lineal}} + \underbrace{\left(\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2 + C_2 - C_1\right)}_{\text{constante}} = 0\,.$$

# Clasificación Estadística Paramétrica Multivariada





# Prueba de hipótesis

Una hipótesis en estadística es una afirmación acerca de un valor, de un intervalo, de la forma de una distribución, etc. Las hipótesis no se *prueban*, sino que se les asigna una probabilidad, y se contrasta esa probabilidad respecto de la *hipótesis nula* $H_0$(básicamente, que la afirmación es falsa).

La «prueba» de la hipótesis consiste en demostrar que la probabilidad de la hipótesis nula es inferior a un determinado porcentaje (la *significatividad* $\alpha$, también llamada *valor-p*).

Se interpreta siempre que rechazar una hipótesis nula verdadera (error tipo 1) es más grave que aceptar una hipótesis nula falsa (error tipo 2).

# Prueba de hipótesis

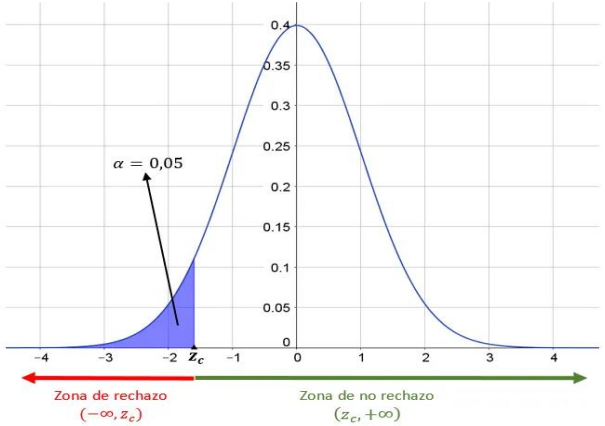
Pasos involucrados en la prueba de hipótesis:

1. Reconocer y definir la o las variables.
2. Formular las hipótesis nula y alternativa.
3. Establecer un estadístico de prueba adecuado.
4. Seleccionar un nivel $\alpha$ de significatividad.
5. Determinar la zona de rechazo y establecer la regla de decisión.
6. Calcular el valor observado del estadístico de prueba.

# Prueba de hipótesis

Por ejemplo, una bodega proclama vende botellas de vino de 1 litro, con un DS de 10 cc (hipótesis nula). La hipótesis alternativa es que la media del contenido es menor. Queremos estar seguros con un 95% de probabilidad de que $H_0$ es falsa (valor p=0.05) evaluando el contenido de N=100 botellas.

Si asumimos que la distribución es normal, la zona de rechazo es que la media de la muestra esté aproximadamente a 1.6 DS por debajo de la media esperada si $H_0$ fuese verdadera.



# Ejercicios

**Ejercicio 1**: En https://github.com/manlio99/Materia-de-aprendizaje/blob/master/4_DataWrangling/data/casos_covid_bahia.csv

y https://github.com/manlio99/Materia-de-aprendizaje/blob/master/4_DataWrangling/data/camas_covid_bahia.csv

están datasets de la situación epidemiológica y sanitaria relacionada con el covid en Bahía Blanca en los últimos meses (fuente https://datos.bahia.gob.ar/dataset?groups=covid19).

Las variables son discretas (personas) pero se puede aplicar razonablemente el análisis estadístico como si fuesen valores continuos.

Hay variables que puedan ser razonablemente consideradas *normales*? Hay variables que tienen un comportamiento sospechoso?

# Ejercicios

**Ejercicio 2**: Generar un dataset similar al de la pág. 12 de este apunte (dos conjuntos  Gaussianos con diferente media y DS, N=50 c/u, uno con etiqueta A y otro con etiqueta B).

Utilizar la curva ROC para proponer un umbral para un clasificador por mínimo error. Evaluar algunos de los parámetros de calidad (exactitud, precisión, f-measure).

Recalcular para algunas variantes (por ejemplo, acercando las medias de los grupos A y B, cambiando el valor umbral, etc.).

Cómo podrían hacer no supervisado este proceso?

| Etiq. | Val. |
|---|---|
| A | 2 |
| A | 3 |
| A | 5 |
| B | 6 |
| A | 8 |
| A | 9 |
| B | 10 |
| A | 11 |
| A | 13 |
| B | 15 |
| A | 15 |
| A | 16 |
| B | 16 |
| A | 17 |
| A | 19 |
| ... | ... |

## Ejercicio 3 (optativo):

Cómo serían los pasos si el dataset fuese 2D?
Puedo utilizar la curva ROC sin ninguna estimación previa?
Caso negativo, qué estimación sería útil (y cómo obtenerla)?

| Etiq. | X | Y |
|---|---|---|
| A | 12 | 1 |
| A | 13 | 1 |
| B | 17 | 1 |
| B | 22 | 2 |
| A | 24 | 2 |
| A | 29 | 2 |
| B | 30 | 2 |
| B | 31 | 3 |
| A | 33 | 3 |
| B | 35 | 3 |
| A | 36 | 3 |
| A | 36 | 4 |
| B | 36 | 4 |
| B | 37 | 4 |
| A | 39 | 4 |
| ... | ... | ... |