Comparación de modelos

Minería de datos

Master Universitario en Tecnologías de Análisis de Datos Masivos
Escola Técnica Superior de Enxeñaría (ETSE)
Universidade de Santiago de Compostela

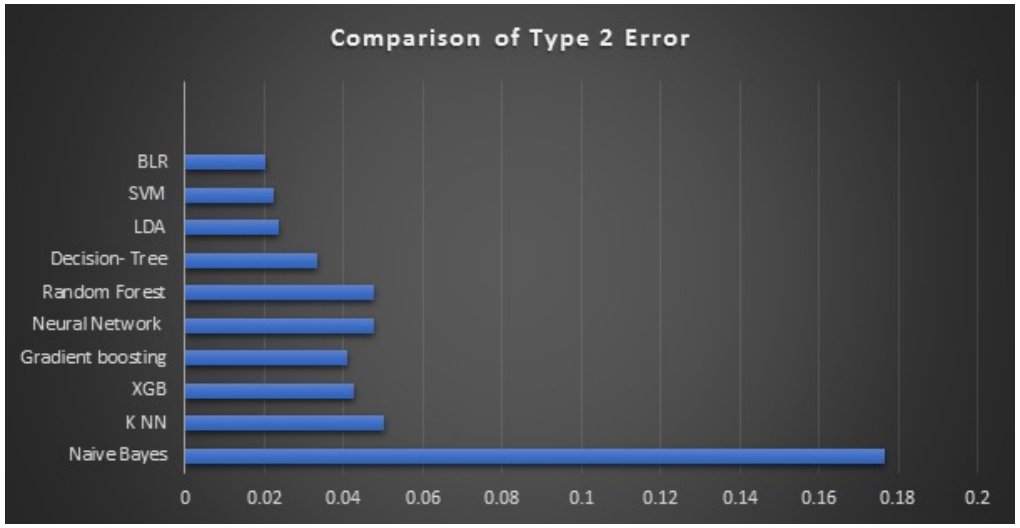
Contenidos de la presentación

- Introducción
- Análisis ROC
- Tests estadísticos
  - Dos clasificadores en un dominio
  - Dos clasificadores en varios dominios
  - Varios clasificadores en varios dominios
  - Varios clasificadores en un dominio
- Conclusiones

Introducción

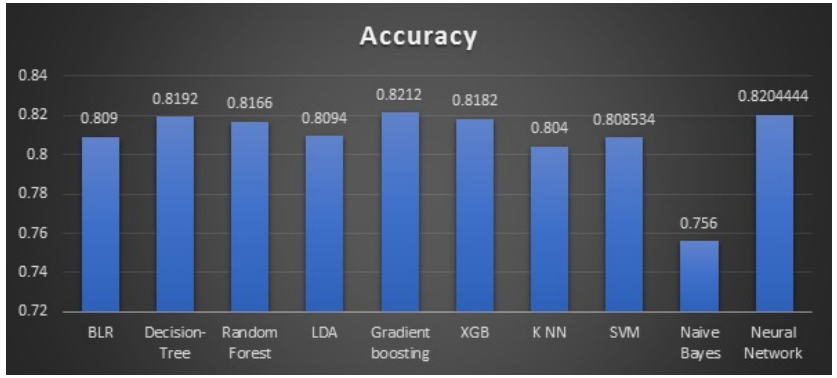
- En el tema anterior, analizamos distintas técnicas para evaluar la calidad de un modelo y de generalizar los resultados de esa evaluación.

Sin embargo, los resultados obtenidos no siempre nos permiten identificar cuál de los modelos es el mejor.



Introducción

- Si se realizan las pruebas en un **único conjunto** de datos, es probable que se puedan clasificar los algoritmos en base a sus resultados.
- Cuando se evalúan en **varios conjuntos** de datos, la comparativa se hace mucho más complicada.



¿Cómo se pueden comparar diferentes modelos entre sí?

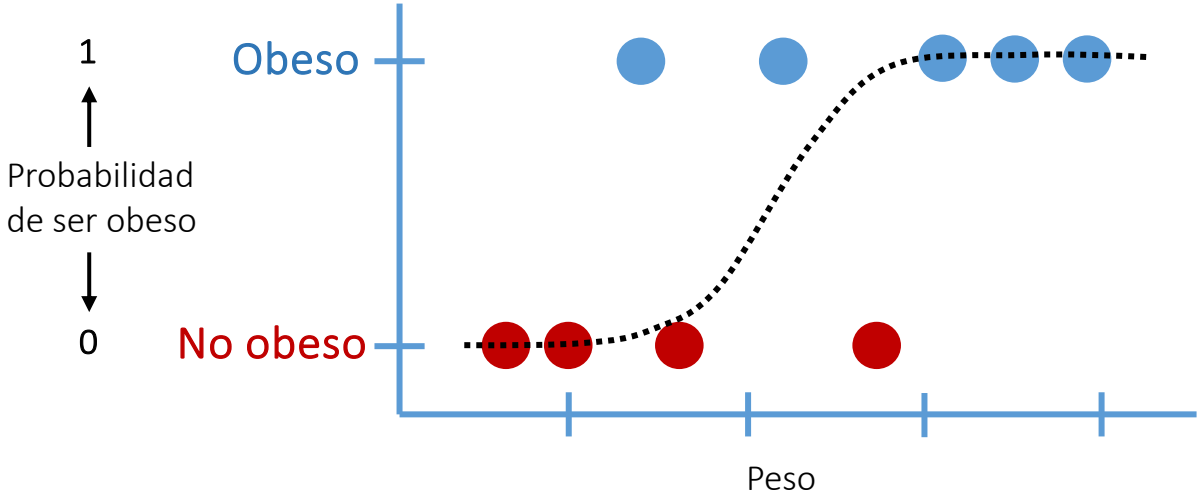
	BPI 2012	BPI 2012 A	BPI 2012 Complete	BPI 2012 O	BPI 2012 W	BPI 2012 W Complete	BPI 2013 Closed Problems	BPI 2013 Incidents	Env Permit	Helpdesk	Nasa	Sepsis
Camargo	83.28	75.98	77.93	81.35	76.4	68.95	54.67	66.68	**85.78**	82.93	-	-
Evermann	59.33	75.82	62.38	79.42	75.37	67.53	58.83	**66.78**	76.19	**83.66**	20.37	40.0
Hinkka	**86.65**	**81.19**	**80.64**	**87.23**	84.78	**70.54**	**63.47**	**74.69**	84.43	83.08	**88.42**	**63.5**
Khan	42.9	74.9	47.37	66.08	60.15	52.22	43.58	51.91	83.59	79.97	12.71	21.01
Mauro	**84.66**	**79.76**	**80.06**	**82.74**	**85.98**	68.64	24.94	36.67	53.59	31.79	21.03	**61.52**
Pasquadibisceglie	83.25	74.12	74.6	78.88	81.19	68.34	47.45	46.03	**86.69**	**83.93**	88.27	56.15
Tax	**85.46**	**79.53**	**80.38**	**82.29**	85.35	69.79	**64.01**	**70.09**	85.71	**84.19**	**89.44**	**64.22**
Theis et al. (w/o attributes)	82.89	65.5	75.26	78.38	**86.22**	**80.06**	**59.48**	59.41	**86.29**	78.77	**88.96**	55.74
Theis et al. (w/ attributes)	80.96	65.67	75.75	76.89	**86.86**	**83.84**	54.65	51.5	85.12	79.69	86.34	58.14

Curvas ROC

Es una representación gráfica de la sensibilidad frente a la especificidad en un clasificador binario según se varía el umbral de discriminación.

Supongamos el siguiente ejemplo, donde se clasifican ratones en base a si son obesos o no.

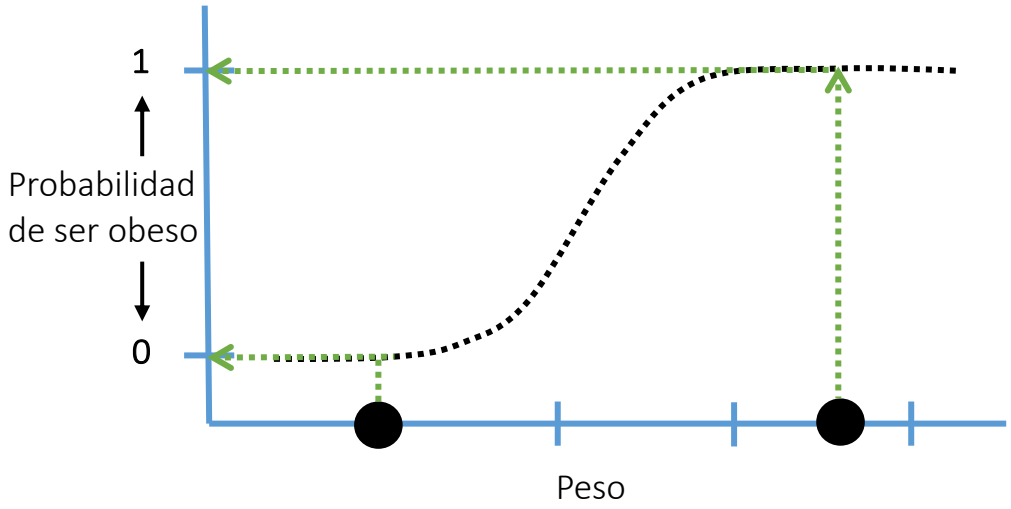
- Los puntos azules representan a ratones obesos y los puntos rojos a ratones no obesos.
- En el eje X, representamos el peso de esos ratones y en el eje Y la probabilidad de ser obeso.



Ajustamos una curva basada en la función de regresión logística para clasificar los puntos:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Ejemplo de cálculo de curva ROC

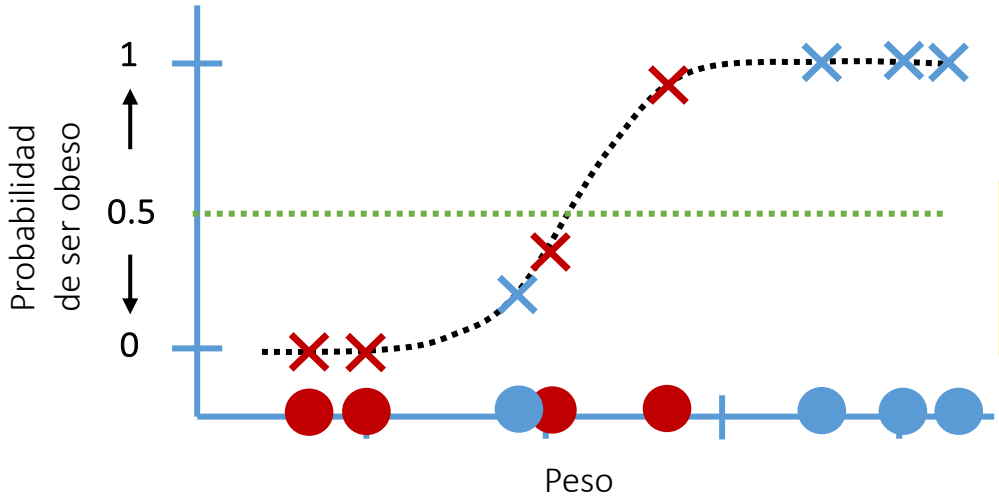


La función logística nos indica la probabilidad de que una rata sea obesa en base a su peso

Si lo que queremos es clasificar los ratones en obesos o no obesos, entonces necesitamos una forma de transformar las probabilidades en clases.

Una forma de hacerlo es definiendo un umbral

Ejemplo de cálculo de curva ROC



Lo que queda por encima del umbral se clasifica como obeso. Lo que queda por debajo del umbral se clasifica como no obeso.

Si se evalúa la efectividad de esta regresión logística, con el umbral de clasificación en 0.5, para un nuevo conjunto de ratones obesos (azules) y no obesos (rojos) obtendríamos:

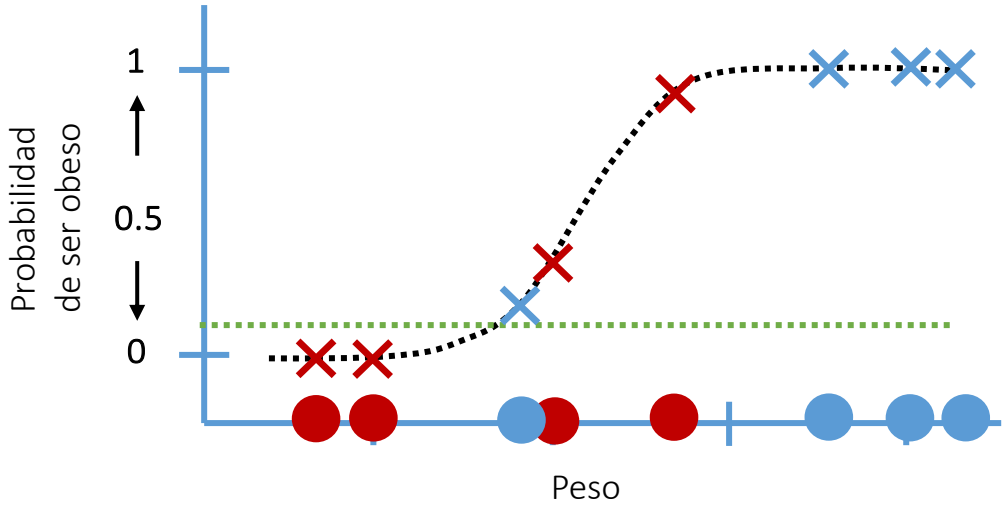
A partir de esta matriz de confusión podríamos calcular la precisión, recall, especificidad, ...

		Predicción: Obeso	Predicción: No obeso
Real	Obeso	3	1
Real	No obeso	1	3

Ejemplo de cálculo de curva ROC

¿Y si se evalúa con un umbral diferente?

Por ejemplo, si fuese muy importante clasificar correctamente los ejemplos obesos, podríamos establecer el umbral en 0.1



Con el umbral a 0.1, todos los ratones obesos se clasifican correctamente.

¡Reducimos los falsos negativos!

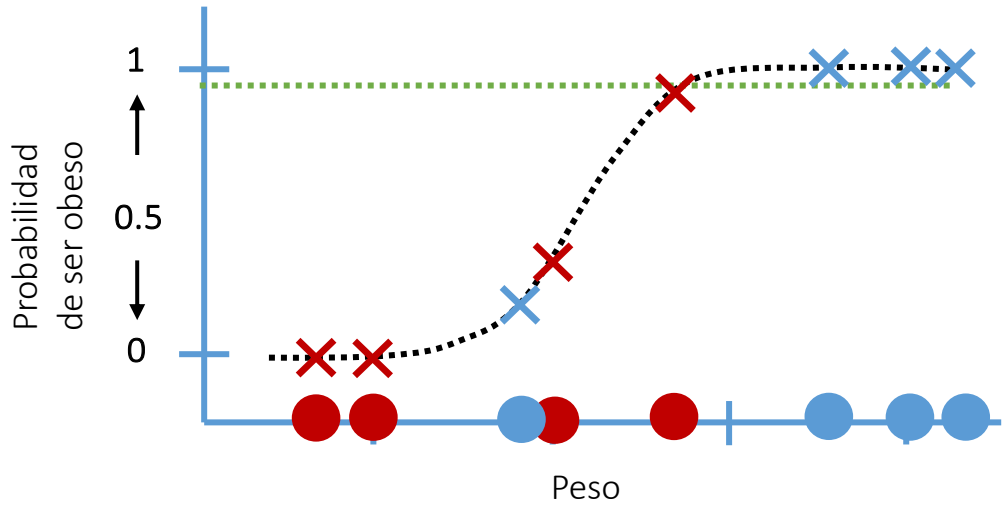
		Predicción: Obeso	Predicción: No obeso
Real	Obeso	4	0
Real	No obeso	2	2

¿Es mejor clasificador?

¿Y si estuviésemos clasificando pacientes con covid o con ébola?
En ese caso tendría sentido bajar el umbral, aún a costa de falsos positivos

Ejemplo de cálculo de curva ROC

¿Qué pasa si fijamos el umbral en 0.9?



Con el umbral a 0.9, todos los ratones no obesos se clasifican correctamente.

¡Ya no hay falsos positivos!
Todos los no obesos se clasifican correctamente.

Predicción

Real	Obeso	No obeso
Obeso	3	1
No obeso	0	4

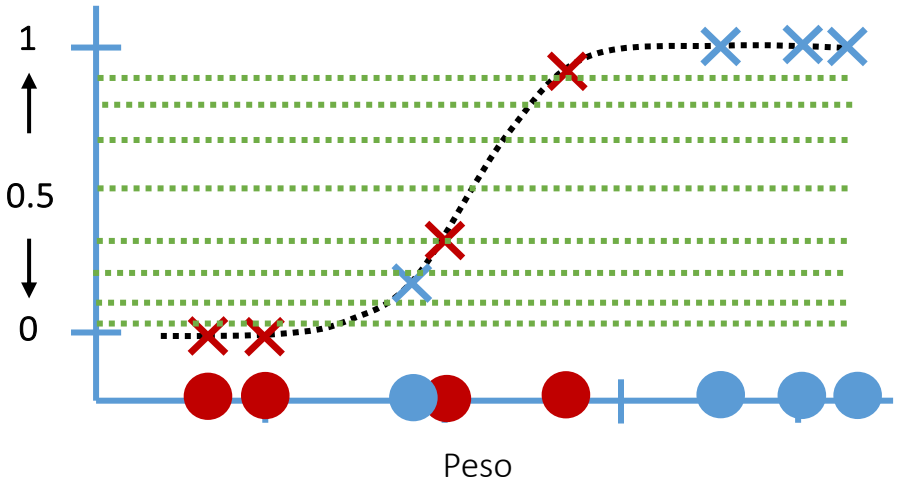
Por ahora, parece el mejor clasificador

Sin embargo, el umbral puede establecerse en cualquier valor entre 0 y 1

¿Cómo determinamos cuál es el mejor umbral de todos?

Ejemplo de cálculo de curva ROC

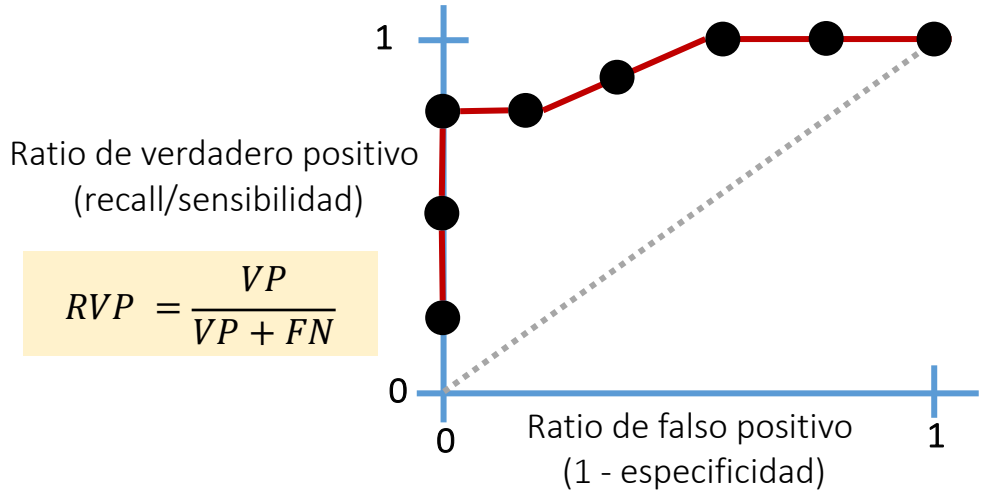
¿Tenemos que calcular las matrices de confusión de todos umbrales?



¡¡¡Hay infinitos umbrales entre 0 y 1!!!

Algunos se podrían no calcular, ya q sabemos que su matriz de confusión es idéntica a otra.

Las curvas ROC (Receiver Operator Characteristics) proporcionan una visualización gráfica de esta comparación, simplificando la visualización de un elevado conjunto de matrices de confusión.



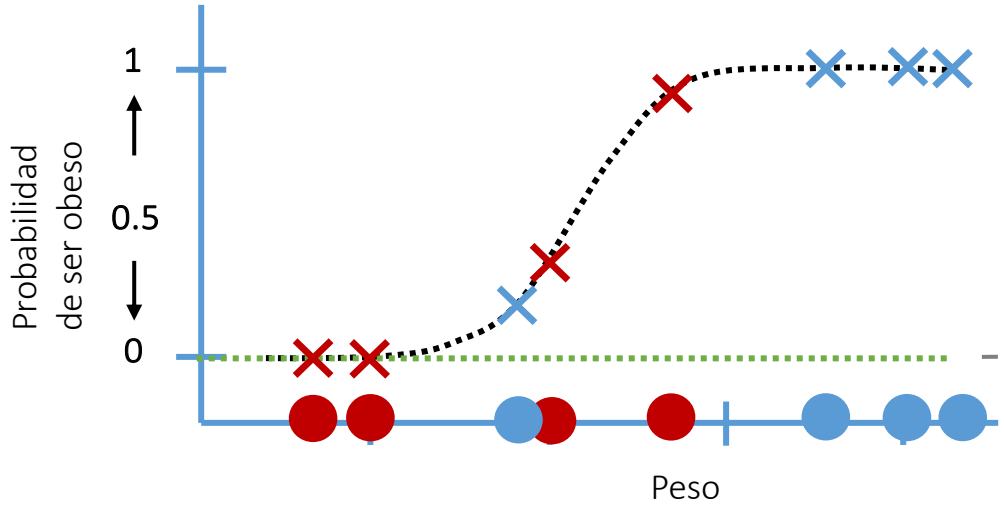
$$RVP = \frac{VP}{VP + FN}$$

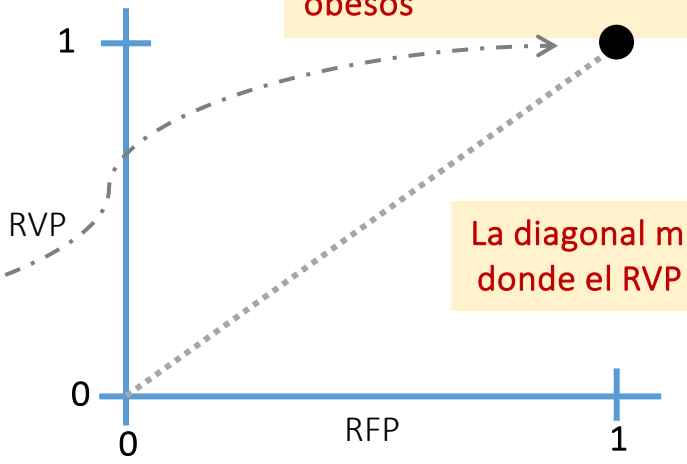
$$RFP = (1 - especificidad) = \frac{FP}{FP + VN}$$

Ejemplo de cálculo de curva ROC

Veamos cómo se crea una curva ROC a partir del ejemplo anterior.

Empecemos por un umbral igual a 0:





Esto significa que se han clasificado correctamente todos los obesos pero se han clasificado incorrectamente todos los no obesos

La diagonal muestra donde el RVP = RFP

Predicción

Real	Obeso	No obeso
Obeso	4	0
No obeso	4	0

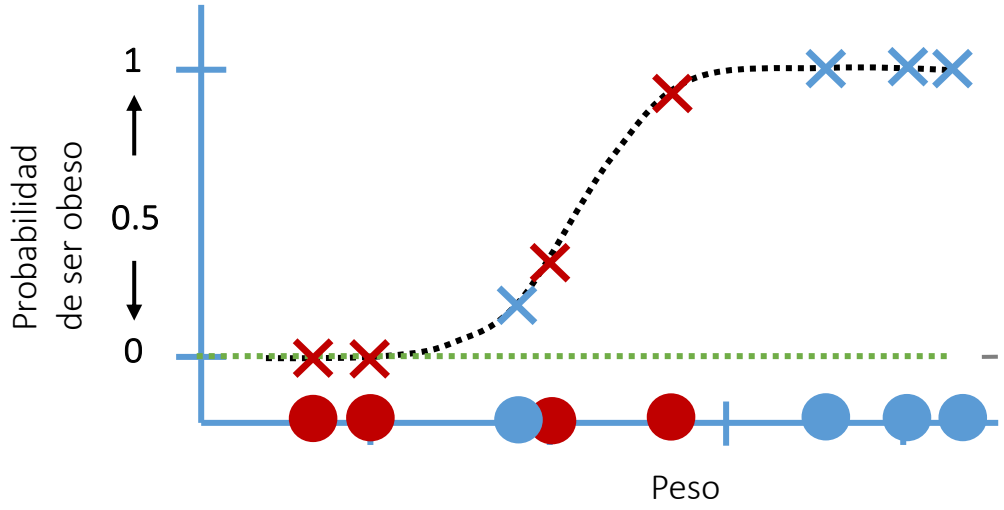
$$RVP = \frac{4}{4+0} = 1$$

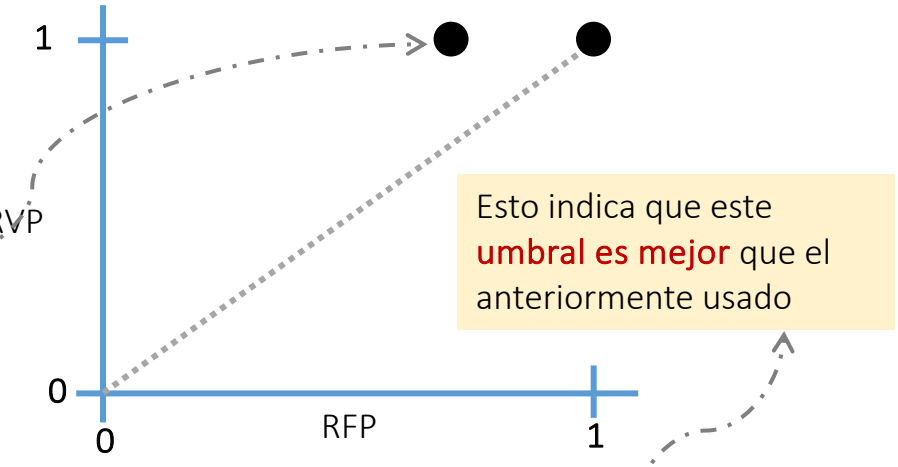
$$RFP = \frac{4}{4+0} = 1$$

Cualquier punto en esa diagonal indica que la proporción de ejemplos clasificados correctamente como obesos es la misma que la proporción de ejemplos clasificados incorrectamente que no son obesos.

Ejemplo de cálculo de curva ROC

Subamos el umbral a 0.05, de forma que todos los puntos menos el primero pertenezcan a la clase obeso.





Esto indica que este **umbral es mejor** que el anteriormente usado

Predicción

Real		Obeso	No obeso
	Obeso	4	0
	No obeso	3	1

$$RVP = \frac{4}{4+0} = 1$$

$$RFP = \frac{3}{3+1} = 0.75$$

Al estar el nuevo punto a la izquierda de la diagonal, sabemos que la proporción de ejemplos correctamente clasificados como obesos (VP) es mayor que la proporción de ejemplos clasificados incorrectamante como obesos (FP).

Ejemplo de cálculo de curva ROC

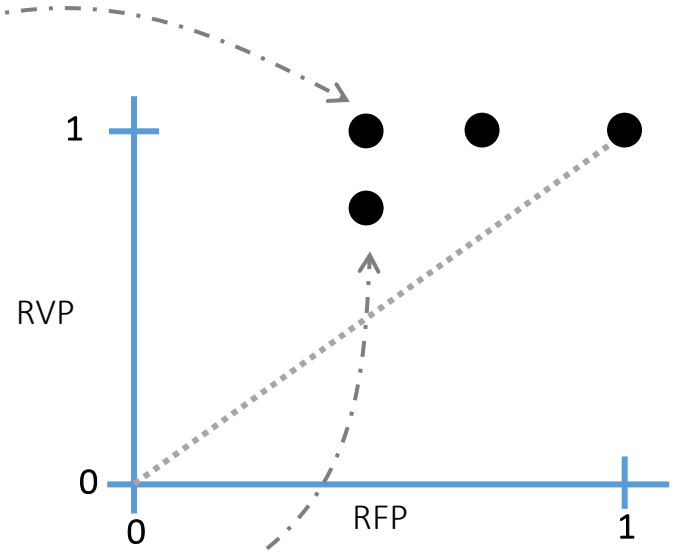
Subamos el umbral a 0.1, de forma que todos los puntos menos los dos primeros pertenezcan a la clase obeso.

Este umbral es nuevamente mejor

Real \ Predicción	Obeso	No obeso
Obeso	4	0
No obeso	2	2

$$RVP = \frac{4}{4+0} = 1$$

$$RFP = \frac{2}{2+2} = 0.5$$



Subamos el umbral a 0.25, de forma que todos los puntos menos los tres primeros pertenezcan a la clase obeso.

Real \ Predicción	Obeso	No obeso
Obeso	3	1
No obeso	2	2

$$RVP = \frac{3}{3+1} = 0.75$$

$$RFP = \frac{2}{2+2} = 0.5$$

Ejemplo de cálculo de curva ROC

Añadamos el umbral igual a 0.5:

Real \ Predicción	Obeso	No obeso
Obeso	3	1
No obeso	1	3

$$RVP = \frac{3}{3+1} = 0.75$$

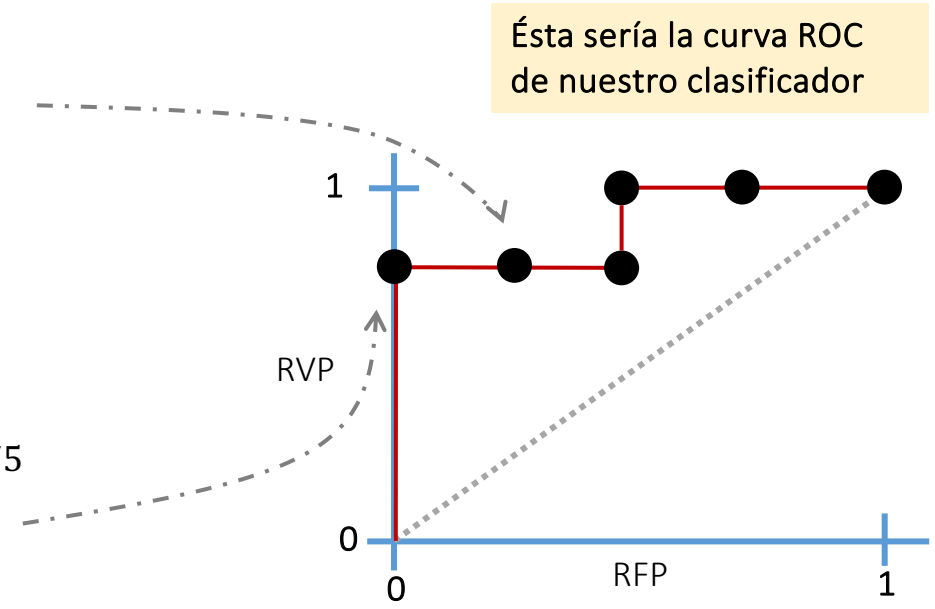
$$RFP = \frac{1}{1+3} = 0.25$$

- Añadamos el umbral igual a 0.9:

Real \ Predicción	Obeso	No obeso
Obeso	3	1
No obeso	0	4

$$RVP = \frac{3}{3+1} = 0.75$$

$$RFP = \frac{0}{0+4} = 0$$



- A partir del umbral 0.9, daría los mismos valores de RVP y RFP.

La selección del umbral dependerá de si queremos aceptar más o menos FP

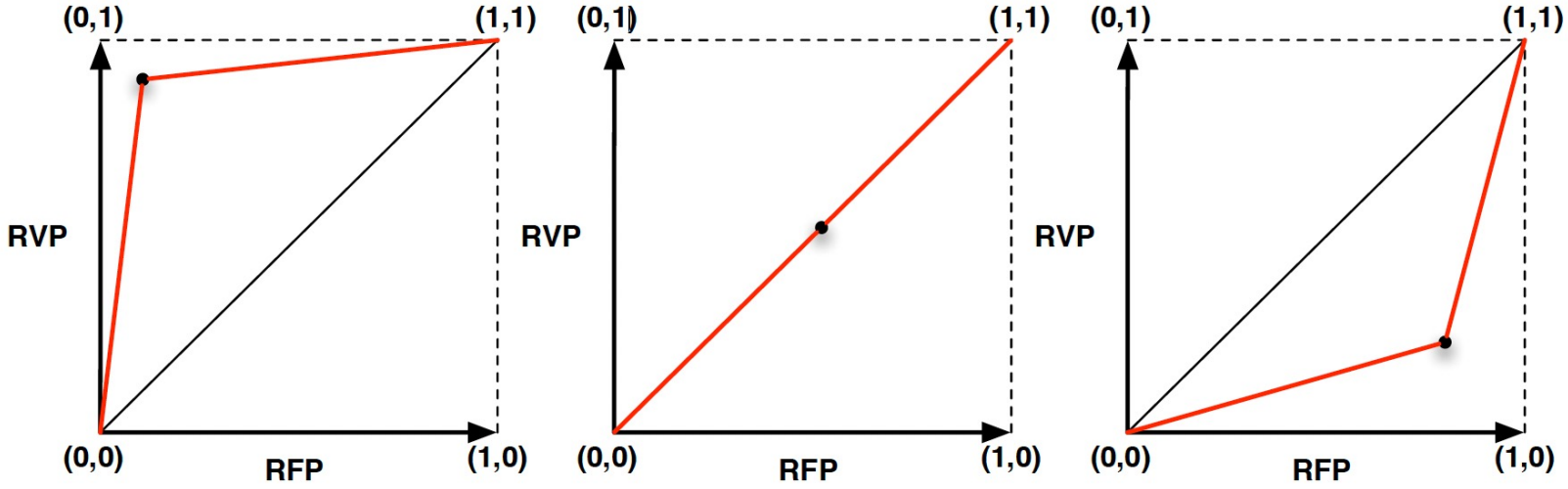
Normalmente nos interesan puntos con el mayor RVP y el menor RFP

Curvas ROC

Resumen:

- (0,1) es el clasificador perfecto
- (0,0) es un clasificador que predice todo como clase negativa
- (1,0) es un clasificador que predice todo como clase positiva
- Recta (0,0) a (1,1) es un clasificador aleatorio (mismo número de VP que de FP).

Por lo tanto, debemos obtener clasificadores por encima de la diagonal



Buen Clasificador:
- Alta tasa de VP
- Baja tasa de FP

Clasificador Aleatorio
- Misma tasa de FP y VP

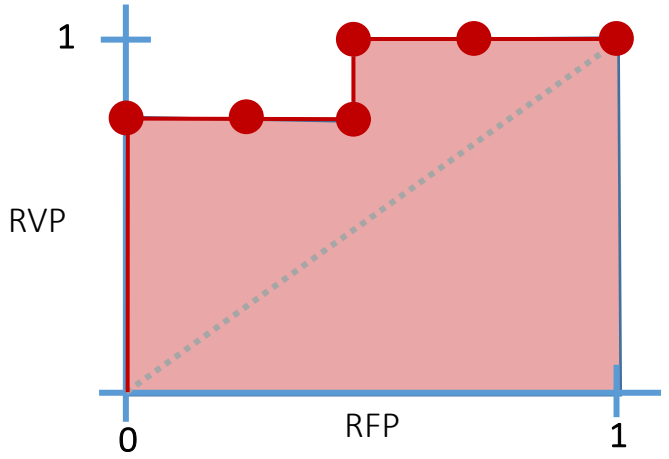
Mal Clasificador:
- Baja tasa de VP
- Alta tasa de FP

Área bajo la curva ROC

El **área bajo la curva** (AUC – *Area Under the Curve*) es el área bajo la curva ROC.

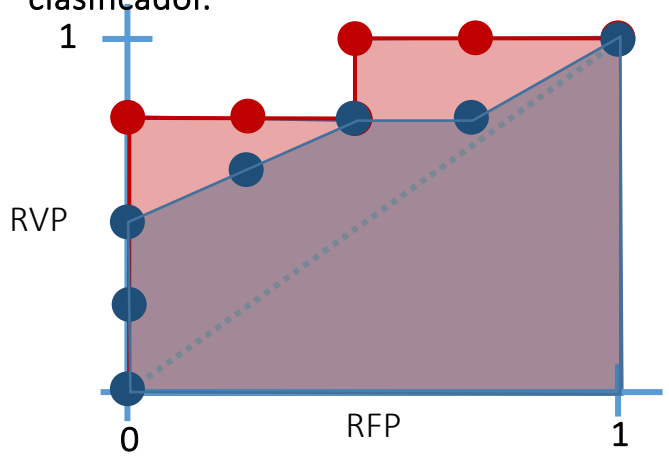
- Se suele calcular a través de integrales, aunque en nuestro caso se puede calcular de forma geométrica como:

$$AUC = 1 - (0.25 \cdot 0.5) = 0.875$$



Supongamos ahora una segunda curva obtenida por otro clasificador, por ejemplo, un SVM o un Random Forest (azul)

La AUC de la regresión logística es mayor, y por lo tanto, es un mejor clasificador.

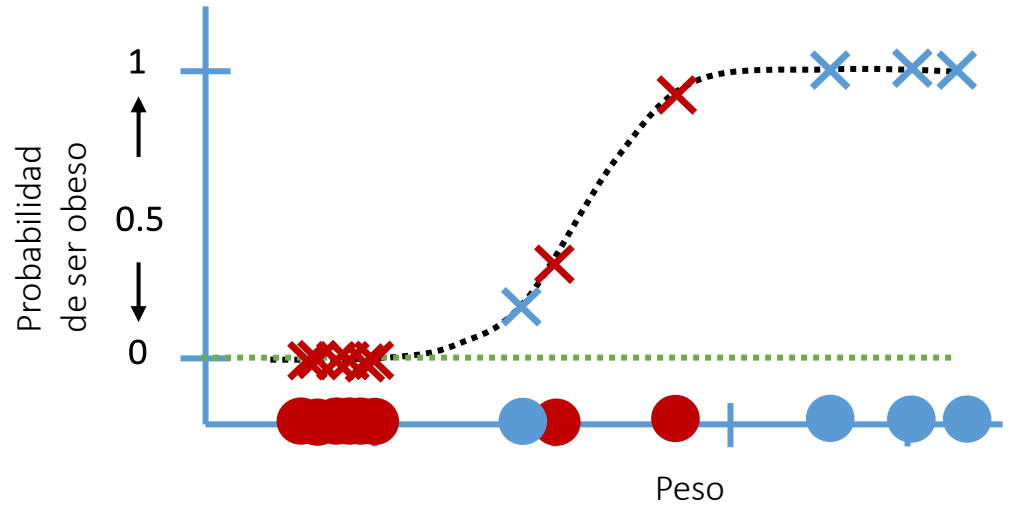


- El AUC y el test de Wilcoxon-Mann-Whitney son equivalentes.
- El estadístico AUC mide la probabilidad de que, si elegimos al azar un ejemplo de la clase positiva y otro de la clase negativa, el clasificador asigne una mayor puntuación al ejemplo positivo.
- No garantiza que los clasifique bien, pero garantiza que existe un umbral que los clasifique bien.

Área bajo la curva ROC

- Aunque las curvas ROC se suelen dibujar usando la RVP y la RFP para resumir la información de las matrices de confusión, se pueden usar otras métricas.
- Por ejemplo, se suele reemplazar la RFP por la precisión, ya que proporciona una mejor representación respecto a la proporción de positivos correctamente clasificados.

Por ejemplo:



Si tuviésemos muchas muestras de no obesos, la precisión sería más adecuada ya q no incluye los VN y no se vería afectada por el desbalanceo

En medicina, y ensayos clínicos, suele ser habitual este tipo de desbalanceo, ya que suele haber muchos más sanos que enfermos.

Ejercicio de cálculo de curva ROC

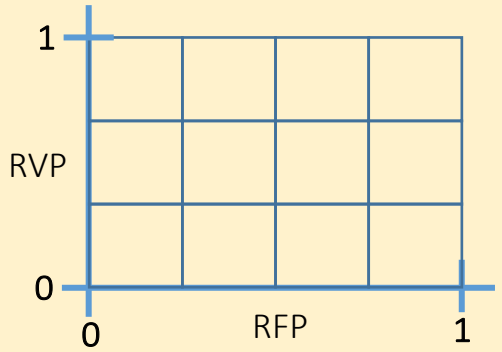
En este caso, se dispone de la estimación del clasificador para cada individuo:

id	estimación	clase
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Paso 1: Ordenar la tabla por el valor estimado de cada objeto.

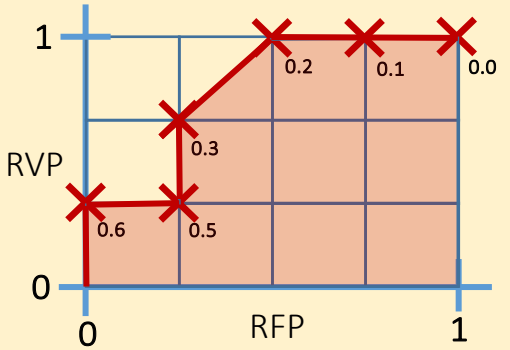
id	estimación	clase
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Paso 2: Dividir el espacio en una rejilla con *m* partes horizontales y *n* verticales, donde *m* es el número objetos de la clase positiva y *n* de la clase negativa.



Paso 3: A partir de la tabla ordenada, dibujar las líneas, empezando en el (0,0). Si el valor de la etiqueta de la fila analizada es 1, subimos un escalón; si es 0, nos deslizamos un escalón a la derecha.

$$AUC = 9.5/12 \sim 0.79$$



NOTA: Si varios objetos tienen el mismo valor, nos deslizamos al punto *i* bloques superior y *j* bloques a la derecha, donde *i* es el número de individuos positivos y *j* negativos.

Ejercicio de cálculo de curva ROC

Faltaría por decidir el umbral, tal y como hicimos en el ejemplo de los ratones:

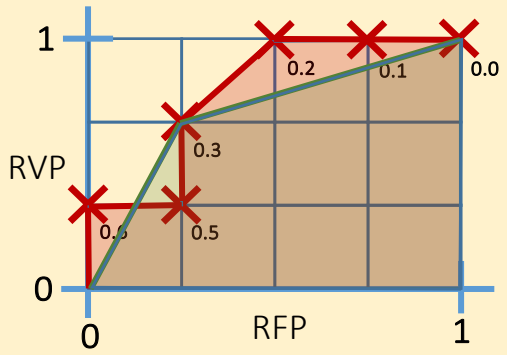
id	estimación	clase
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Paso 4: Decidir la pertenencia del objeto a la clase.

id	> 0.25	clase
4	1	1
1	1	0
6	1	1
3	0	0
5	0	1
2	0	0
7	0	0

Paso 5: Aplicamos el mismo procedimiento que en el paso 3 para dibujar los puntos de la curva ROC.

$$AUC = 8.5/12 \sim 0.71$$



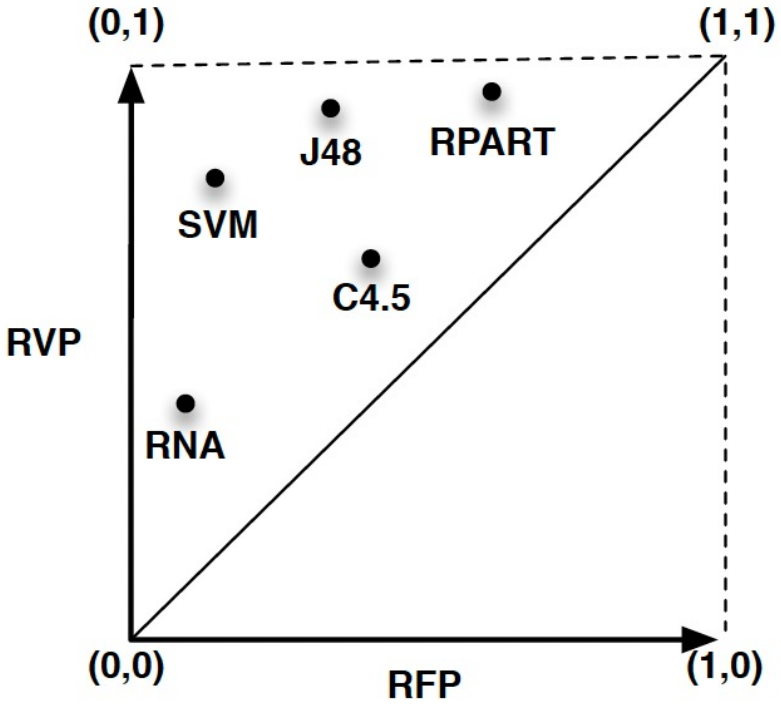
$$RVP = 2/3$$
$$RFP = 1/4$$

Comparación de modelos

Antes vimos que se pueden comparar distintos modelos si disponemos de su curva ROC

- A mayor área, el clasificador distingue mejor entre la clase positiva y la negativa

¿Qué pasa si solo disponemos de una medida para cada uno de los clasificadores?

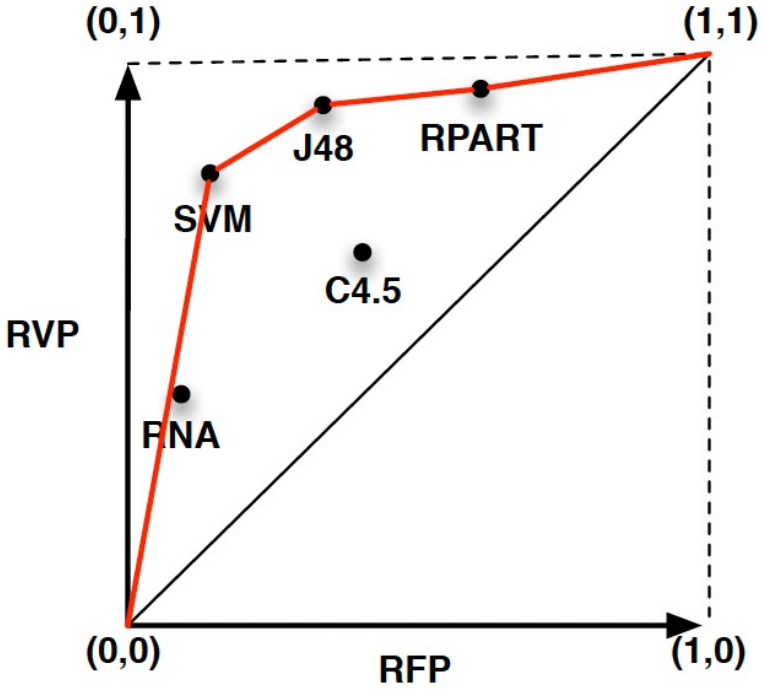


Paso 1: Representar cada modelo en el espacio ROC

Paso 2: Calcular la envolvente convexa teniendo en cuenta los puntos (0,0) y (1,1)

Todo modelo por debajo de la envolvente convexa debe ser descartado

El mejor modelo se calcula en función del coste y el contexto (*skew*).



Análisis ROC de un conjunto de modelos

En una situación normal la eficacia dependerá:

- De la matriz de costes (no todos los errores pesan igual)
- Del contexto (*skew*) definido por la distribución de clases.

Estos dos aspectos se pueden agrupar en una medida basada en el espacio ROC: pendiente (*slope*)

$$slope = \frac{coste(FP)}{coste(FN)} \cdot \frac{N}{P}$$

donde:

- N=número de ejemplos negativos
- P=número de ejemplos positivos

Para determinar el modelo más apropiado a la situación planteada:

- Trazar una recta con pendiente *slope* en el punto (0,1).
- Trasladar dicha recta hasta la curva ROC.
- El primer punto que toque es el mejor modelo.

Si se desconocen dichos datos, se puede suponer slope = 1.

- Se elige el punto más cercano al punto (0,1).

Ejemplo: Vamos a suponer que nuestro conjunto de prueba tiene 300 clases negativas y 150 clases positivas.

- Caso1: Coste(FP) = 2 y Coste(FN) = 4

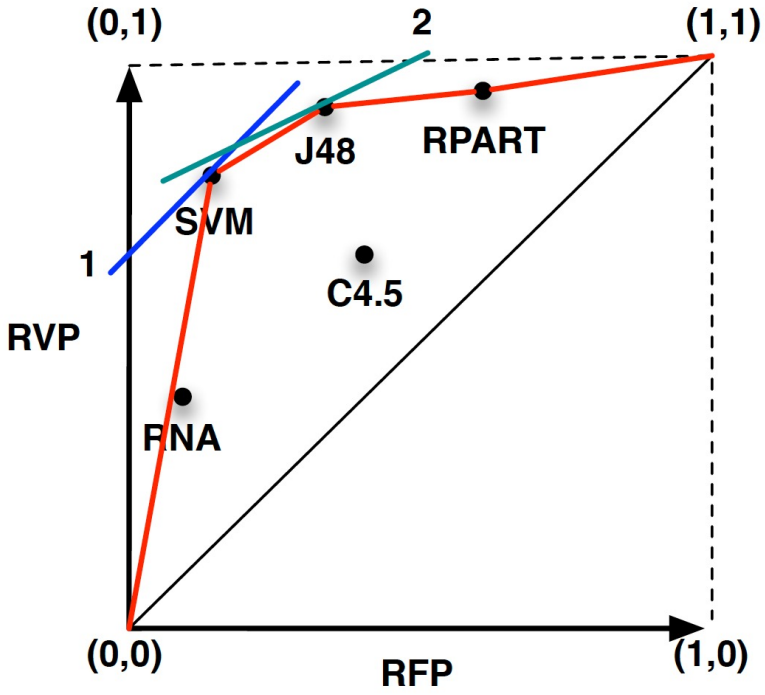
$$slope_1 = \frac{2}{4}\frac{300}{150} = 1$$

- Caso2: Coste(FP) = 1 y Coste(FN) = 4

$$slope_2 = \frac{1}{4}\frac{300}{150} = 0.5$$

Análisis ROC de un conjunto de modelos

- Para una pendiente igual a 1, el SVM es el mejor clasificador.
- En cambio, si la pendiente es 2, el J48 es el mejor clasificador.



Comparación de modelos vía tests estadísticos

Test de Hipótesis:
Procedimiento estadístico mediante el cual se investiga la verdad o falsedad de una hipótesis acerca de una característica de una población o un conjunto de poblaciones

Tests paramétricos:
Conocida una v.a. con una determinada distribución, se establecen afirmaciones sobre los parámetros de dicha distribución

Ejemplo:

- Sea $X_1, X_2, ..., X_n$ una m.a.s. de una v.a. X con distribución normal, $N(\mu, \sigma)$. Establecemos la siguiente afirmación: μ<10

Tests no paramétricos:
Las afirmaciones establecidas no se hacen en base a la distribución de las observaciones, que a priori es desconocida.

Ejemplos:

- Análisis de la aleatoriedad de la muestra
- Una variable aleatoria X tiene una distribución Normal
- Dos variables aleatorias X e Y son independientes
- Dos muestras independientes proceden de la misma población

Hipótesis del test

Hipótesis del test:

- Hipótesis nula (H_0): Hipótesis que se plantea en un problema de contraste
- Hipótesis alternativa (H_1): Hipótesis contraria a la hipótesis nula

Ejemplo para test paramétrico:

H_0: $\mu \leq 10$

H_1: $\mu > 10$

Ejemplo para test no paramétrico:

H_0: La muestra se ha seleccionado aleatoriamente

H_1: La muestra no se ha seleccionado aleatoriamente

Estadístico del test:

Es una variable aleatoria, con distribución de probabilidad conocida, y cuyos valores nos permiten tomar la decisión de aceptar o rechazar la hipótesis nula.

Ejemplo:

$$\left.\begin{matrix} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{matrix}\right\} \qquad \bar{X} \rightarrow N\left(\mu\,;\,\frac{\sigma}{\sqrt{n}}\right)$$

El valor concreto que toma el estadístico del test para la muestra escogida se llama Valor Experimental o Estadístico del contraste:

$$x_1, x_2, \ldots, x_n \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Tipos de errores

Errores asociados al contraste:

- Error tipo I: Error que se comete al rechazar la hipótesis nula, H_0, cuando ésta es cierta.

$$\alpha = P[\text{Error tipo I}] = P[\text{Rechazar } H_0 / H_0 \text{ es verdadera}]$$

- Error tipo II: Error que se cometa al no rechazar la hipótesis nula, H_0, cuando ésta no es cierta.

$$\beta = P[\text{Error tipo II}] = P[\text{Rechazar } H_0 / H_0 \text{ es verdadera}]$$

H_0	Rechazo	No rechazo
Verdadero	Error tipo I (α)	Correcto
Falso	Correcto	Error tipo II (β)

- Potencia del test: Probabilidad que se tiene en el contraste de detectar que H_0 es falsa.

$$1 - \beta = P[\text{Rechazar } H_0 / H_0 \text{ es falsa}]$$

Probabilidad de detectar un efecto real cuando éste exite

Contraste de hipótesis

Ejemplo: contrastar si la media de una población $N(\mu;\sigma)$ con σ conocida, toma un valor $\mu = \mu_0$

1. Planteamiento del test: $H_0: \mu = \mu_0$
   $H_1: \mu \neq \mu_0$

2. Estadístico del test: $\bar{X} \rightarrow N\left(\mu\,;\,\frac{\sigma}{\sqrt{n}}\right)$

   Bajo la hipótesis nula: $\bar{X} \rightarrow N\left(\mu_0;\,\frac{\sigma}{\sqrt{n}}\right)$

   Se toma un m.a.s. concreta $x_1, x_2, \ldots, x_n$ cuya media valdrá $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

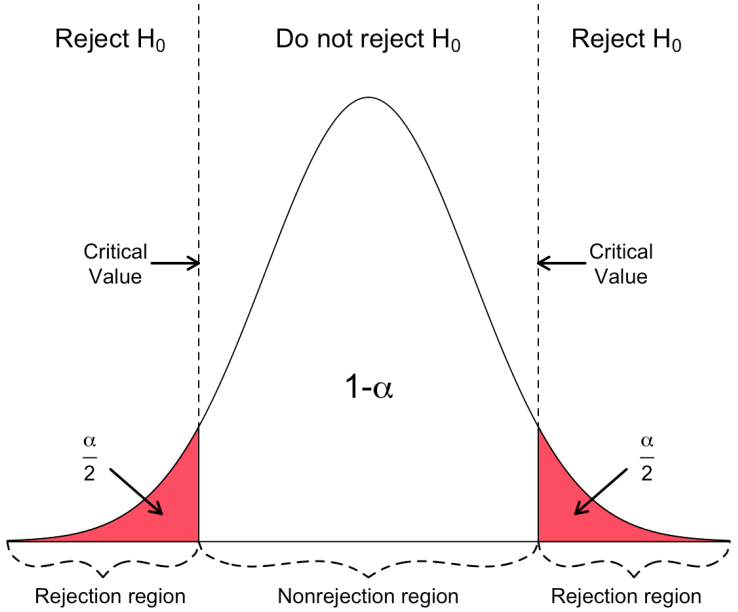
   Si H_0 es cierta, la mayoría de los valores de la media muestral están próximos al valor μ_0.

3. Criterio de decisión: Comprobar si el valor de la media muestral calculada está o no muy alejado de μ_0.
   - Rechazamos H_0 si la media muestral no está "próxima" a μ_0.
   - No rechazamos H_0 si la media muestral está "próxima" a μ_0.

Contraste de hipótesis

Ejemplo: contrastar si la media de una población $N(\mu;\sigma)$ con σ conocida, toma un valor $\mu = \mu_0$

4. Determinación de las zonas de rechazo y no rechazo
   - Zona de rechazo: 100α % de los valores restantes.
   - Zona de no rechazo: $100(1-\alpha)$ % de los valores más cercanos a μ_0.



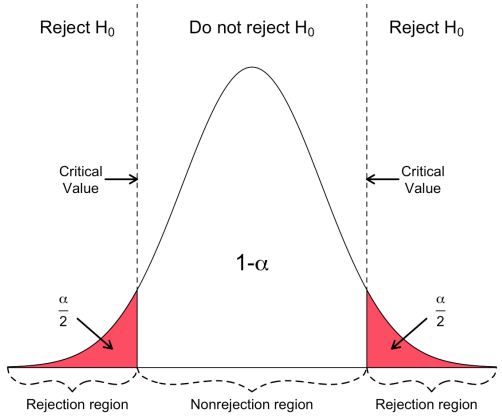
Contraste de hipótesis

Ejemplo: contrastar si la media de una población $N(\mu;\sigma)$ con σ conocida, toma un valor $\mu = \mu_0$

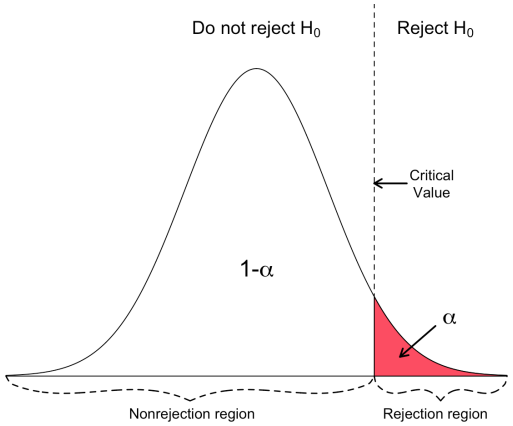
Tipos de hipótesis

- **Hipótesis simples**: La hipótesis asigna un único valor al parámetro desconocido, $H_0: \theta = \theta_0$.
- **Hipótesis compuestas**: La hipótesis asigna varios valores al parámetro desconocido, $H_0: \theta \in (\theta_1, \theta_2)$.

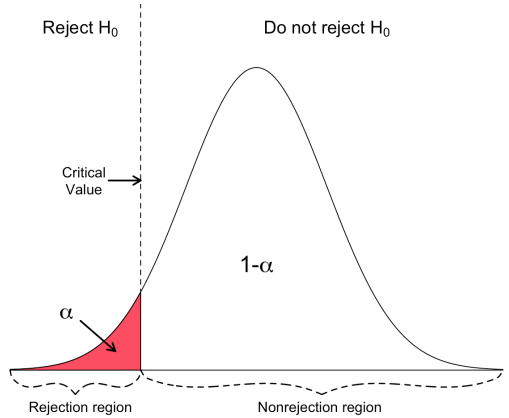
$H_0: \mu = \mu_0 \mid H_1: \mu \neq \mu_0$
Simple – Compuesta
Contraste bilateral



$H_0: \mu \leq \mu_0 \mid H_1: \mu > \mu_0$
Compuesta – Compuesta
Contraste unilateral derecho



$H_0: \mu \geq \mu_0 \mid H_1: \mu < \mu_0$
Compuesta – Compuesta
Contraste unilateral izquierdo



Contraste de hipótesis

Ejemplo: contrastar si la media de una población $N(\mu;\sigma)$ con σ conocida, toma un valor $\mu = \mu_0$

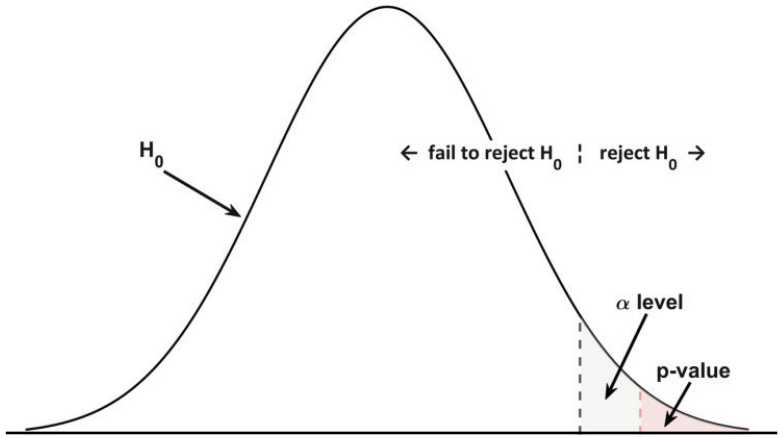
Al aplicar un contraste de hipótesis, clasificamos los puntos del espacio muestral en dos regiones excluyentes y complementarias:

- Región de rechazo o región crítica: La formada por el conjunto de los valores del estadístico de contraste que nos llevan a rechazar la hipótesis nula H_0 (los puntos que delimitan la región crítica se llaman puntos críticos).
- Región de no rechazo o región de aceptación: La formada por el conjunto de los valores del estadístico de contraste que nos lleva a aceptar la hipótesis nula H_0

- p-valor o nivel de significación observado: Probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula.

Un p-valor bajo indica que los datos observados serían muy improbables bajo la suposición de que la hipótesis nula es cierta.

Un p-valor alto sugiere que los datos observados son consistentes con la hipótesis nula, indicando que no hay suficiente evidencia para rechazarla.



Elegido un nivel de significación α, se rechazará H_0 si $p < \alpha$.

Tipos de análisis

	ANÁLISIS PARAMÉTRICO	ANÁLISIS NO PARAMÉTRICO
Base de prueba estadística	Distribución.	Arbitrario.
Nivel de medición	Datos en intervalo o razón.	Datos ordinales o nominales.
Tipo y tamaño de muestra	Aleatoria >30 sujetos	No aleatoria <30 sujetos
Variables	Aplicable en variables nominales.	Aplicable en variables categóricas.
Valores perdidos	No se consideran una fuente de información.	Se consideran una fuente de información.
Consideraciones	Debe contar con normalidad y homocedasticidad.	Menor presunción y alcance más amplio.
Generalidades	-Se conoce el modelo de distribución de la población. -Mientras más grande sea la muestra más exacta será la estimación, mientras más pequeña, más distorsionada será la media de las muestras. -Las hipótesis se basan en valores numéricos, especialmente en promedios.	-Se desconoce cómo están distribuidos los datos. -Se puede utilizar, aunque se desconozca los parámetros de la población en estudio. -Es utilizada para contrastar con la hipótesis. -Las hipótesis se redactan sobre rangos, mediana o frecuencia d ellos datos.
Ventajas	-Más eficiencia. -Poca probabilidad de errores. -Sus estimaciones son exactas. -Presentan sensibilidad a los rasgos de los datos recogidos. -Muestras grandes.	-Empleada en diferentes situaciones porque no cumple con parámetros estrictos. -Sus métodos son más afables. -Se aplica en datos no numéricos. -Muestras pequeñas.
Desventajas	-Complejos de calcular. -Presentan una limitación en los datos.	-No son sistemáticas. -Complica seleccionar la elección correcta. -Provoca confusión. -Requiere fuentes y respaldo. -Probabilidad de errores. -No hay exactitud.

Tipos de análisisStudent

	ANÁLISIS PARAMÉTRICO	ANÁLISIS NO PARAMÉTRICO
	Pruebas estadísticas más usadas	
	Coeficiente de correlación de Pearson T de Student ANOVA	Chi cuadrada
	Pruebas estadísticas de acuerdo con sus datos	
Única muestra	Medias	Ji cuadrada Binominal Rachas Kolmogorov-Sminov
De 2 muestras independientes	T de Student para grupos independientes Levene para igualdad de varianzas T de igualdad de Medias	U de Mann-Whitney Reacciones Extremas de Moses
De 2 muestras relacionadas	Correlación Pearson	Wilcoxon de los Signos Mc Nemar
De una muestra medida en 2 momentos diferentes	T de Student para una muestra relacionada	W de Wilcoxon
De 3 o más muestras independientes	Análisis de Varianza (ANOVA)	Prueba de Análisis de Varianza de Kruskal-Wallis
De una muestra medida en 3 o más momentos diferentes	ANOVA para una muestra relacionada	Análisis de Varianza por Rangos Señalados de Friedman
De varias muestras relacionadas	ANOVA de factor	Friedman Coeficiente de Concordancia W de Kendall Cochran

Estadístico Z

Contraste sobre los parámetros de una **distribución normal**:

$$X_1, X_2, \ldots, X_n \text{ una m.a.s. de } X \rightarrow N(\mu, \sigma)$$

Contrastes sobre la **media** con la **varianza conocida**		
Hipótesis del test	Estadístico de contraste	Criterio de rechazo
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \rightarrow N(0;1)$	$z_{exp} \leq -z_{\alpha/2}$ $z_{exp} \geq z_{\alpha/2}$
$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$		$z_{exp} \geq z_{\alpha}$
$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$		$z_{exp} \leq -z_{\alpha}$

Test t de Student

Contraste sobre los parámetros de una **distribución normal**:

$$X_1, X_2, \ldots, X_n \text{ una m.a.s. de } X \rightarrow N(\mu, \sigma)$$

Contrastes sobre la **media** con **varianza desconocida**		
Hipótesis del test	Estadístico de contraste	Criterio de rechazo
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \rightarrow t_{n-1}$	$t_{exp} \leq -t_{\alpha/2;n-1}$ $t_{exp} \geq t_{\alpha/2;n-1}$
$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$		$t_{exp} \geq t_{\alpha;n-1}$
$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$		$t_{exp} \leq -t_{\alpha;n-1}$

Ejercicio

En un preparado alimenticio infantil se especifica que el contenido medio de proteínas es al menos del 42%. Tratamos de comprobar esta especificación y para ello tomamos 10 preparados que analizamos para determinar su contenido en proteínas, obteniendo una media del 40% y una cuasidesviación típica del 3.5%. **¿Es correcta la especificación citada para un nivel de significación del 0.05, suponiendo normal la distribución de la variable contenido proteico?**

X: contenido proteico

$X \to N(\mu, \sigma)$

$n = 10;\ \bar{x} = 40; s = 3.5$

Estadístico del contraste:

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \to t_{n-1}$$

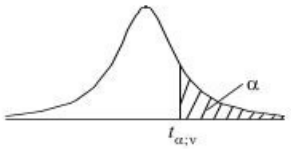
Contraste de hipótesis:

H_0: $\mu \geq 42$ $\quad \alpha = 0.05$

H_1: $\mu < 42$ $\quad t_{0.95;9} = -t_{0.05;9} = -1.833$

Table of the Student's *t*-distribution

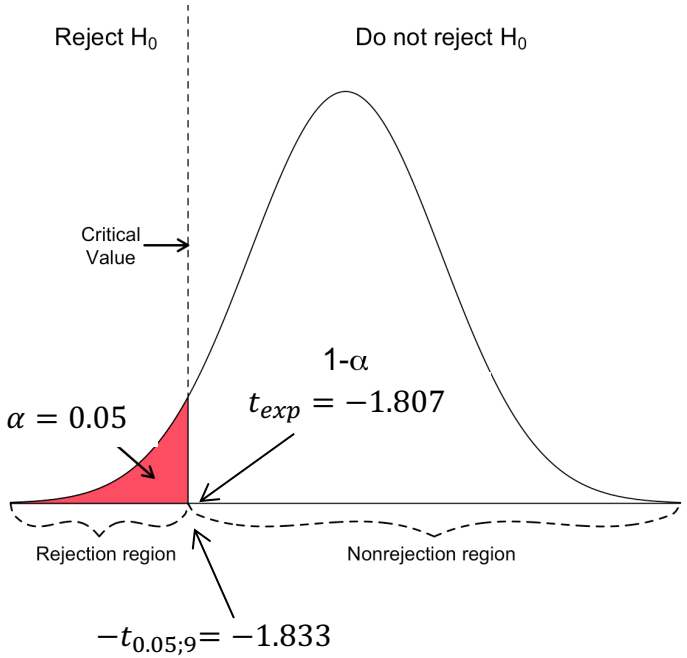
The table gives the values of $t_{\alpha;\nu}$ where $\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Ejercicio

En un preparado alimenticio infantil se especifica que el contenido medio de proteínas es al menos del 42%. Tratamos de comprobar esta especificación y para ello tomamos 10 preparados que analizamos para determinar su contenido en proteínas, obteniendo una media del 40% y una cuasidesviación típica del 3.5%. **¿Es correcta la especificación citada para un nivel de significación del 0.05, suponiendo normal la distribución de la variable contenido proteico?**



$$t_{exp} = \frac{40 - 42}{3.5/\sqrt{10}} = -1.807 \Longrightarrow \text{No rechazamos } H_0$$

Admitimos como correcta la especificación del preparado acerca del contenido proteico.

Test paramétrico - Test de chi-cuadrado χ^2

Contrate sobre los parámetros de una distribución normal:

$$X_1, X_2, \ldots, X_n \text{ una m.a.s. de } X \rightarrow N(\mu, \sigma)$$

Contrastes sobre la **varianza** con **media desconocida**		
Hipótesis del test	Estadístico de contraste	Criterio de rechazo
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \rightarrow {\chi^2}_{n-1}$	${\chi^2}_{exp} \leq {\chi^2}_{1-\alpha/2;n-1}$ ${\chi^2}_{exp} \geq {\chi^2}_{\alpha/2;n-1}$
$H_0: \sigma^2 \leq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$		${\chi^2}_{exp} \geq {\chi^2}_{\alpha;n-1}$
$H_0: \sigma^2 \geq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$		${\chi^2}_{exp} \leq -{\chi^2}_{1-\alpha;n-1}$

Ejercicio

La varianza habitual para la altura de los machos de Lhasa Apso es de 0.25. Un criador está intentando reducir esta cifra. Después de un período de crianza selectiva, se selecciona una muestra de 15 perros a los que se mide, obteniendo una cuasivarianza muestral de 0.21. **¿Tenemos evidencias que nos permitan afirmar que ha disminuido la variabilidad en la altura de esta raza de perros?**

X: altura de los machos L.A.

$X \rightarrow N(\mu, \sigma)$

$n = 15; s^2 = 0.21$

Estadístico del contraste:

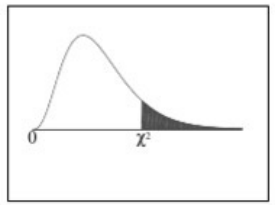
$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \rightarrow {\chi^2}_{n-1}$$

Contraste de hipótesis:

$H_0: \sigma^2 \geq 0.25$ $\quad \alpha = 0.05$

$H_1: \sigma^2 < 0.25$ $\quad {\chi^2}_{0.95;14} = 6.571$

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

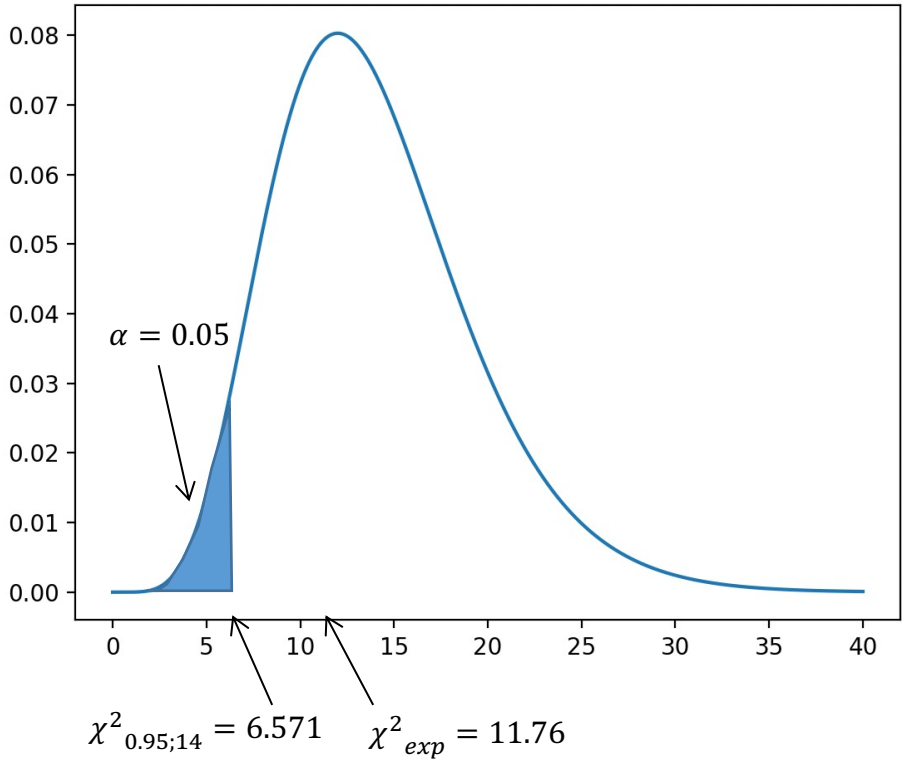
df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Ejercicio

La varianza habitual para la altura de los machos de Lhasa Apso es de 0.25. Un criador está intentando reducir esta cifra. Después de un período de crianza selectiva, se selecciona una muestra de 15 perros a los que se mide, obteniendo una cuasivarianza muestral de 0.21. **¿Tenemos evidencias que nos permitan afirmar que ha disminuido la variabilidad en la altura de esta raza de perros?**

$$\chi^2_{exp} = \frac{14 \cdot 0.21}{0.25} = 11.76 \Longrightarrow \text{No rechazamos } H_0$$

No tenemos suficientes pruebas para sostener la información de que la crianza selectiva haya reducido la variabilidad en las alturas de los machos de Lhasa Apso.



Test paramétrico – Varias distribuciones

Contrate sobre los parámetros de dos distribuciones normales independientes:

$X_1, X_2, \ldots, X_{nx}$ una m.a.s. de $X \rightarrow N(\mu_{nx}, \sigma)$ $\quad$ $Y_1, Y_2, \ldots, Y_{ny}$ una m.a.s. de $Y \rightarrow N(\mu_{ny}, \sigma)$

Contrastes sobre la diferencia de **medias** con **varianzas conocidas**		
Hipótesis del test	Estadístico de contraste	Criterio de rechazo
$H_0: \mu_X - \mu_Y = \mu_0$ $H_1: \mu_X - \mu_Y \neq \mu_0$	$Z = \dfrac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}} \rightarrow N(0;1)$	$Z_{exp} \leq Z_{\alpha/2}$ $Z_{exp} \geq Z_{\alpha/2}$
$H_0: \mu_X - \mu_Y \leq \mu_0$ $H_1: \mu_X - \mu_Y > \mu_0$		$Z_{exp} \geq Z_{\alpha}$
$H_0: \mu_X - \mu_Y \geq \mu_0$ $H_1: \mu_X - \mu_Y < \mu_0$		$Z_{exp} \leq -Z_{\alpha}$

Test paramétrico – Varias distribuciones

Contrate sobre los parámetros de dos distribuciones normales independientes:

$X_1, X_2, \ldots, X_{nx}$ una m.a.s. de $X \rightarrow N(\mu_{nx}, \sigma)$ $\quad$ $Y_1, Y_2, \ldots, Y_{ny}$ una m.a.s. de $Y \rightarrow N(\mu_{ny}, \sigma)$

Contrastes sobre la **diferencia de medias** con **varianzas desconocidas, pero iguales**		
Hipótesis del test	Estadístico de contraste	Criterio de rechazo
$H_0: \mu_X - \mu_Y = \mu_0$ $H_1: \mu_X - \mu_Y \neq \mu_0$	$T = \frac{(\bar{X} - \bar{Y}) - \mu_0}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \rightarrow t_{n_X+n_Y-2}$	$Z_{exp} \leq -t_{\alpha/2;n_X+n_Y-2}$ $Z_{exp} \geq t_{\alpha/2;n_X+n_Y-2}$
$H_0: \mu_X - \mu_Y \leq \mu_0$ $H_1: \mu_X - \mu_Y > \mu_0$		$Z_{exp} \geq t_{\alpha/2;n_X+n_Y-2}$
$H_0: \mu_X - \mu_Y \geq \mu_0$ $H_1: \mu_X - \mu_Y < \mu_0$		$Z_{exp} \leq -t_{\alpha/2;n_X+n_Y-2}$

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

Ejercicio

En un estudio sobre la angina de pecho en ratas, se dividió aleatoriamente a 18 animales afectados en dos grupos de 9 individuos cada uno. A un grupo se le suministró un placebo y al otro un fármaco experimental FL113. Después de un ejercicio controlado sobre una "cinta sin fin", se determinó el tiempo de recuperación de cada rata, obteniéndose los siguientes resultados:

Placebo	**FL113**
$n_X = 9$	$n_Y = 9$
$\bar{x} = 339$ seg.	$\bar{y} = 283$ seg.
$S_X = 45$ seg.	$S_Y = 43$ seg.

¿Se puede concluir que el fármaco experimental tiende a reducir el tiempo de recuperación? (Se supone igualdad en las varianzas poblacionales)

X: Tiempo de recuperación ratas con placebo
Y: Tiempo de recuperación ratas con FL113

$X \to N(\mu_X, \sigma_X)$
$Y \to N(\mu_Y, \sigma_Y)$ independientes

Contraste de hipótesis:

$$H_0: \mu_X - \mu_Y \leq 0$$
$$H_1: \mu_X - \mu_Y > 0$$

Estadístico del contraste:

$$T = \frac{(\bar{X} - \bar{Y}) - \mu_0}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \to t_{n_X+n_Y-2}$$

Ejercicio

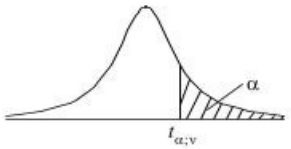
$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2} = \frac{8 \cdot 45^2 + 8 \cdot 43^2}{9 + 9 - 2} = 1937$$

$$\left.\begin{array}{c} t_{exp} = \dfrac{(329 - 283)}{\sqrt{1937} \cdot \sqrt{2/9}} = 2.223 \\ t_{0.05;16} = 1.796 \end{array}\right\} \text{Rechazamos la hipótesis}$$

El fármaco experimental es eficaz en la reducción del tiempo de recuperación en ratas con angina de pecho

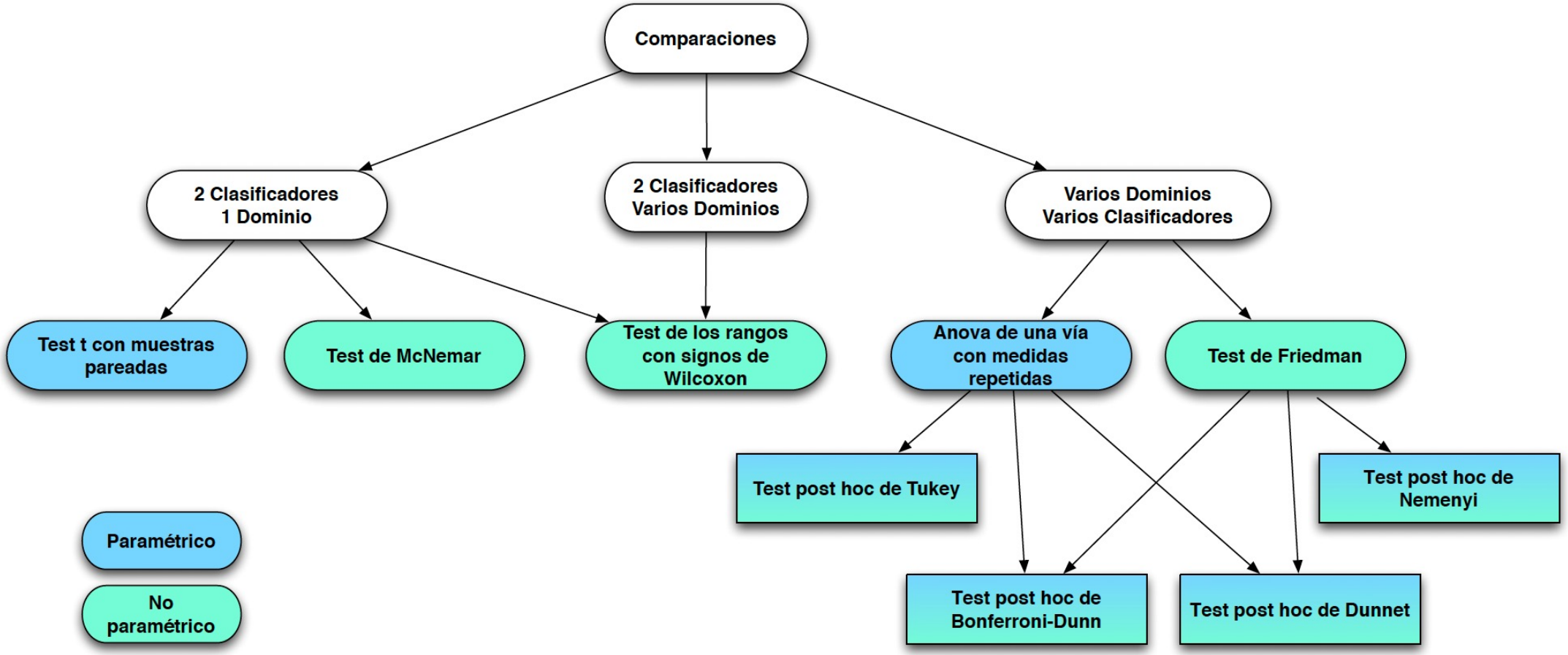
Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Comparación de modelos vía tests estadísticos



Comparación de modelos

Objetivo:
Dadas dos técnicas de clasificación. A y B, y un conjunto de datos S, **¿qué técnica producirá el clasificador más preciso a partir de conjuntos del mismo tamaño?**

Sean $\widehat{f_A}$ y $\widehat{f_B}$ los clasificadores generados por las técnicas A y B respectivamente a partir del conjunto de entrenamiento R.

Hipótesis nula:
Para un conjunto de entrenamiento R seleccionado de forma aleatoria del conjunto de datos S, las dos técnicas producirán clasificadores con la misma tasa de error/acierto.

La selección del test estadístico a utilizar depende de la situación en la que nos encontremos:

- Comparar dos algoritmos en un mismo dominio (data set).
- Comparar varios algoritmos en un mismo domino.
- Comparar varios algoritmos en varios dominios.

Dos clasificadores en un dominio (paramétrico)

Test t de Student por pares con m-1 grados de libertad permite determinar si la diferencia entre las medias de dos medidas pareadas es significativa.

1. Se realizan m particiones (entrenamiento y test) del conjunto inicial S: R_1, ..., R_m y T_1, ..., T_m
2. Se calculan las diferencias de las medidas de error $p^i = p_A^i - p_B^i$
3. Se calcula el estadístico t
4. No se rechaza la hipótesis nula si $|t| \leq t_{1-\alpha/2;m-1}$ con una significancia $\alpha/2$

El estadístico se calcula de la siguiente forma:

$$t = \frac{\bar{p}\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{m}(p^i - \bar{p})^2}{n-1}}}$$

donde $p^i = p_A^i - p_B^i$ son las m diferencias entre los clasificadores A y B, y $\bar{p} = \frac{1}{n}\sum_{i=1}^{m} p^i$.

Test t de Student por pares

El test t de Student nos indica si la diferencia entre las medidas de rendimiento es significativa, pero no nos dice cuán importante es dicha diferencia.

Para ello es necesario calcular el estadístico $\boldsymbol{d}$ de Cohen:

$$d_{cohen} = \frac{\overline{p_A} - \overline{p_B}}{\sigma_p} \text{ donde } \sigma_p = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

La interpretación es la siguiente:

- d_{cohen} sobre 0.3 indica que el efecto es pequeño, pero probablemente significativo.
- d_{cohen} sobre 0.5 indica que el efecto es medio pero apreciable.
- d_{cohen} sobre 0.8 indica que el efecto es grande.

Condiciones de aplicabilidad:

- Normalidad (test de Kolmogorov-Smirnov, Shapiro-Wilk o Anderson-Darling). El t de Student es bastante robusto si no se cumple. Sería suficiente disponer de conjuntos de test con más de 30 muestras.
- Aleatoriedad de las muestras. Difícil de comprobar.
- Homocedasticidad (tests de Finger, Barlett, Levene o Brown-Forsythe). La **igualdad de varianzas en las poblaciones** se puede comprobar visualmente mediante un gráfico de cajas.

Dos clasificadores en un dominio (no paramétrico)

El test de McNemar's es la alternativa no paramétrica al test t de Student.

1. Se divide el conjunto inicial S en conjunto de entrenamiento R y prueba T
2. Se generan los clasificadores $\widehat{f_A}$ y $\widehat{f_B}$
3. Se calcula la siguiente tabla de contingencia:

n_{00} = num. de casos mal clasificados por $\widehat{f_A}$ y $\widehat{f_B}$	n_{01} = num. de casos mal clasificados por *bien por* $\widehat{f_A}$ y bien por $\widehat{f_B}$
n_{10} = num. de casos bien clasificados por $\widehat{f_A}$ y mal por $\widehat{f_B}$	n_{11} = num. de casos bien clasificados por $\widehat{f_A}$ y $\widehat{f_B}$

donde $|T| = n_{00} + n_{01} + n_{10} + n_{11}$

4. No se rechaza la hipótesis nula si el estadístico es menor que 3.85 con el 95% de confianza. Si no, el mejor clasificador es aquel que presenta menor error.

Está basado en el siguiente estadístico que se ajusta a una distribución χ_1^2:

$$M = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Test de McNemar's

Hipótesis nula: $n_{01} = n_{10}$ (ambos clasificadores tienen la misma ratio de error)

Condiciones de aplicabilidad: $n_{01} + n_{10} > 20$

Desventajas:

- No tiene en cuenta la aleatoriedad intrínseca de la técnica y de la partición de S. Por lo tanto, solo es aplicable si creemos que la aleatoriedad es pequeña.
- Las técnicas solo se comparan usando un único conjunto de entrenamiento.
- Se debe asumir que la diferencia observada en R se mantiene en S.
- En el caso de problemas multiclase no se puede aplicar.

Dos clasificadores en varios dominios

Estos tests permiten comparar de forma genérica las diferencias entre los clasificadores.

Se podría pensar en extender los tests para dos clasificadores en un dominio:

- El test t de Student asumiendo que las medidas de rendimiento en diferentes dominios tienen que ser comparables.
- El test de McNemar no está pensado para más de dos clasificadores.

Recomendación: Test de los rangos con signo de Wilcoxon.

Test de los rangos con signo de Wilcoxon para muestras pareadas

Es un test no paramétrico para comparar las medianas

Supongamos que tenemos dos clasificadores $\widehat{f_A}$ y $\widehat{f_B}$ evaluados sobre n dominios distintos:

1. Sean p_A^i y p_B^i las medidas de rendimiento de cada clasificador en el dominio i.
2. Se calculan las diferencias entre las medidas para cada dominio $d_i = p_A^i - p_B^i$.
3. Se ordena el valor absoluto de d_i de menor a mayor y se les asigna un rango. En caso de empate se asignan la media de los rangos empatados.
4. Se calculan los siguientes valores (las diferencias iguales a 0 se eliminan):
   W_{s1} = Suma en valor absoluto de los rangos positivos
   W_{s2} = Suma en valor absoluto de los rangos negativos
5. Se calcula el estadístico $T_{Wilcox} = \min(W_{s1}; W_{s2})$ que sigue una distribución T de Wilcoxon.
6. En el caso de que $n > 25$ el estadístico T_{Wilcox} puede ser aproximado por una distribución normal.
7. Se rechaza la hipótesis nula (no existen diferencias significativas) si el estadístico es menor que el valor crítico, para unos grados de libertad y significancia concretos.

Test de los rangos con signo de Wilcoxon para muestras pareadas

Se calcula el siguiente estadístico:

$$Z_{Wilcox} = \frac{T_{Wilcox} - \mu_{T_{Wilcox}}}{\sigma_{T_{Wilcox}}}$$

Donde $\mu_{T_{Wilcox}}$ y $\sigma_{T_{Wilcox}}$ son la media y la desviación estándar de la aproximación a la distribución T_{Wilcox} en el caso de que la hipótesis nula sea cierta.

$$\mu_{T_{Wilcox}} = \frac{n(n+1)}{4} \qquad \sigma_{T_{Wilcox}} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Varios clasificadores en varios dominios

Estos tests nos permiten evaluar varias estrategias de aprendizaje:

- En varios conjuntos de datos para analizar las características generales de los algoritmos.
- En varios conjuntos del mismo problema para ver cuál es la mejor aproximación.

Se podría pensar en hacer comparaciones dos a dos mediante el test t de Student: muchas comparaciones.

Tests omnibus:
Son tests (paramétricos y no paramétricos) permiten realizar contrastes de varias hipótesis al mismo tiempo.

Procedimiento:

1. Aplicar el test omnibus apropiado:
   - Paramétrico: Anova de una vía con medidas repetidas.
   - No paramétrico: Test de Friedman.
2. En el caso de que existan diferencias significativas, aplicar un test post hoc para determinar dónde se encuentran dichas diferencias.
3. Puede darse el caso de que un test omnibus detecte diferencias significativas pero los tests post hoc no.
   - Si esto ocurre, entonces existen diferencias, pero no se pueden identificar debido a la escasa potencia de los tests post hoc.

ANOVA de una vía con medidas repetidas

Al igual que el test t de Student compara las diferencias observadas entre las medias.

Permite determinar si las diferencias observadas entre cualquier número de medias son estadísticamente significativas:

$$H_0: \mu_0 = \mu_1 = \cdots = \mu_n$$
$$H_1: \text{al menos dos medias son distintas}$$

Nos permite descubrir si las diferencias entre las medias (medidas de rendimiento) entre los diferentes grupos (datasets) son estadísticamente significativas.

Idea general:

- Se divide la varianza total en:
  - Varianza causada por el error aleatorio (varianza dentro de los grupos).
  - Varianza causada por las diferencias observadas entre las medias (varianza entre grupos).
- Si se cumple la hipótesis nula, la suma de los cuadrados dentro de los grupos debe ser más o menos igual a la suma de los cuadrados entre números.
- Esto último se puede comprobar utilizando un test F, que determina si la ratio de dos varianzas, medidas como la suma de cuadrados, es significativamente mayor que 1.

ANOVA de una vía con medidas repetidas

Condiciones de aplicabilidad:

- Normalidad
- Homogeneidad entre varianzas (esfericidad). La varianza en cada grupo debe ser similar (test de Mauchly's).
- Las medidas de rendimiento deben tener la misma escala.
- Los conjuntos de datos deben tener aproximadamente el mismo tamaño.

Problemas:

- El test ANOVA para medidas repetidas es robusto (dentro de unos ciertos límites) a la violación de la condición de normalidad.
- La dificultad de comprobar la esfericidad ha llevado a muchos autores a desaconsejar la utilización de este test para
- comparar clasificadores.
- En muchos casos, tenemos medidas de rendimiento categóricas o no monótonas que incumplen la condición de la escala.

Alternativa no paramétrica: Test de Friedman

Test de Friedman

El test de Friedman es la alternativa no paramétrica al test ANOVA con medidas repetidas.

Se comparan las medianas en vez de las medias:

H_0: todas las medianas son iguales

H_1: al menos dos medianas son distintas

Al igual que en el test de Wilcoxon, el test de Friedman basa su análisis en los rangos de cada clasificador más que en sus medidas de rendimiento.

ANOVA vs Friedman:

- Si se cumplen las condiciones de aplicabilidad, ANOVA es más potente.
- En caso de no cumplirse, Friedman es más potente.

Tests post hoc

Los tests omnibus anteriormente comentados solo nos dicen si hay diferencias significativas entre los clasificadores.

Si existen diferencia (se rechaza la hipótesis nula) habrá que localizar dónde están dichas diferencias. Para ello hay que aplicar los test post hoc.

Post hoc paramétricos:
Estos test se aplicarán en el caso de que el test ANOVA de medidas repetidas indique que hay diferencias significativas.

Test de Turkey: Intenta detectar la variación aleatoria entre todos los pares de medias.

- Dichas variaciones aleatorias se comparan con las diferencias reales.
- El estadístico calculado nos indica cuan grande es dicha deferencia comparada con la variación general aleatoria entre medias.

Test de Dunnet: Se puede utilizar cuando las comparaciones no son dos a dos, sino de todos los clasificadores con uno de control.

Tests post hoc

Test de Bonferroni: Equivalente al anterior, solo que se utiliza la corrección de Bonferroni para todas las comparaciones.

- Funciona bien cuando el número de comparaciones es pequeño.
- Cuando el número de comparaciones es grande tiende a ser conservador.

Test de Bonferroni-Dunn: Intenta corregir el conservadurismo del anterior test.

- Divide el nivel de significancia α por el número de comparaciones a realizar.
- También conocido como el test de Dunn.

Post hoc no paramétricos:
Estos test se aplicarán en el caso de que el test Friedman indique que hay diferencias significativas.

Test de Nemenyi: Se basa en un estadístico que mide la diferencia promedio entro los rangos de los clasificadores.

Otros métodos: Se basan en escalar los niveles de significancia.

- Test de Hommel, Test de Holm y Test de Hochberg.

Varios clasificadores en un dominio

Se pueden adaptar los tests anteriores para un solo dominio.

La idea básica consiste en generar varios conjuntos de datos a partir del disponible.

- Generando varios conjuntos de datos con los ejemplos permutados o reordenados.
- Utilizando alguna técnica de muestreo: bootstraping, hold-out, validación cruzada, ...

Sin embargo, corremos el riesgo de que un clasificador siempre dé mejores resultados que otro si éste es robusto respecto al orden o permutaciones de los ejemplos.

- Para evitar esto se recomienda utilizar la validación cruzada sin repetición.

Se pueden aplicar los test post hoc directamente.

Conclusiones

- En este capítulo hemos analizado como diferentes aproximaciones para comparar clasificadores.
- Se ha empezado por analizar cómo utilizar las curvas ROC para la comparación de clasificadores.
- También se han presentado diferentes tests estadísticos que pueden ser utilizados en diferentes circunstancias:
  - Dos clasificadores en un dominio.
  - Dos clasificadores en varios dominios.
  - Varios clasificadores en varios dominios.
- Por último, han analizado las características de dichos tests, sus condiciones de aplicabilidad y su interpretación.

Bibliografía:

- Ethem Alpaydin. Introduction to Machine Learning. MIT Press 2004.
- Demsar, Janez. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research. vol. 7, pp 1{30 (2006).
- Tom Fawcett. An introduction to ROC analysis. Pattern Recognition Letters 27 (2006) 861-874.
- Dietterich, Thomas G. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation. 10(7) pp 1895-1923. (1998).