Comparación de Modelos

Minería de Datos

José T. Palma

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

DIIC, UMU, 2025

Contenidos de la presentación

Introducción

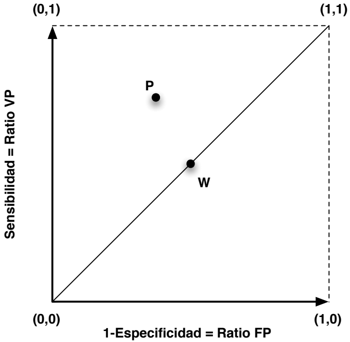
- Hemos analizado técnicas que nos permiten obtener diferentes modelos de clasificación.
- Esto nos permitirá, para un mismo problema:
  - Construir varios modelos.
  - Construir distintas versiones de un mismo modelo.
- Pero ¿Cómo podemos comparar los diferentes modelos entre sí?
- Responder a esta pregunta es clave si queremos proporcionar el mejor modelo posible.

Curva ROC

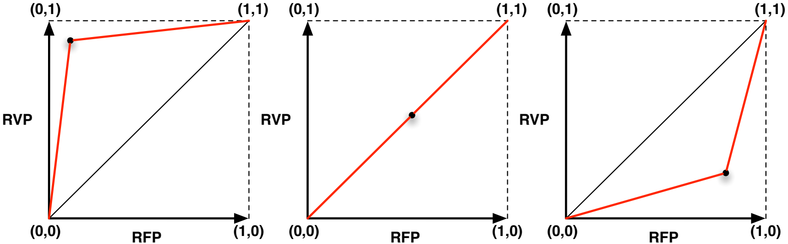
- En principio, el análisis de curvas ROC está ideado para problemas de dos clases. Aunque existen extensiones para multiclases.
- El análisis de curvas ROC (Receiver Operating Characteristics Analysis) nos permite comparar diferentes modelos en un espacio bidimensional.
- El espacio ROC tiene dos coordanadas:
  - En el eje Y se representa la *Sensibilidad* o el *Ratio de Verdaderos positivos*,
  - En el eje X se representa el *Ratio de Falsos Positivos* o *1-Especificidad*
- Por lo tanto, cada modelo quedará representado en dicho espacio mediante un punto.

Curva ROC

- (0,1) $\rightarrow$ clasificador perfecto.
- (0,0) $\rightarrow$ clasificador que predice todo como clase positiva.
- (1,0) $\rightarrow$ clasificador que predice todo como clase negativa.
- Recta (0,0) (1,1) $\rightarrow$ clasificador aleatorio (mismo número de FP y VP)
- Por lo tanto, siempre debemos obtener clasificadores que operen por encima de la diagonal.



Curva ROC



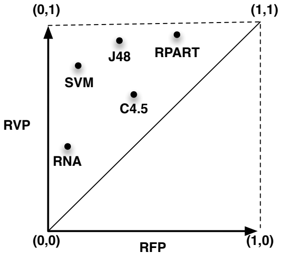
Buen Clasificador:
- Alta tasa de VP
- Baja tasa de FP

Clasificador Aleatorio
- Misma tasa de FP y VP

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

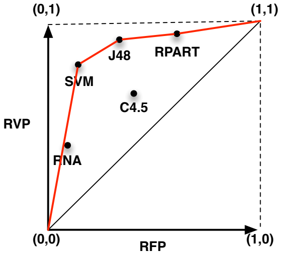
Análisis ROC de un conjunto de modelos

- Representar cada modelo en el espacio ROC.



Análisis ROC de un conjunto de clasificadores

- Representar cada clasificador en el espacio ROC.
- 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).



Evaluación sensible al contexto (skew)

- En una situación normal la eficacia dependerá:
  - de la matriz de costes (no todos los errores pesan igual).
  - el contexto (skew) definido por la distribución de clases.
- Estos dos aspectos se pueden agrupar en una medida basada en el espacio ROC: **pendiente** (slope):

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

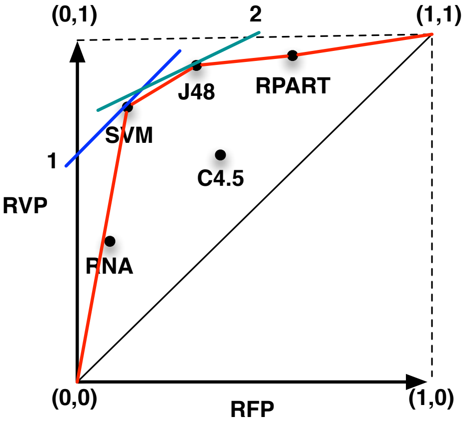
- donde $N =$ número de ejemplos negativos y $P =$ números de ejemplos positivos.

Evaluación sensible al contexto (skew)

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

$$slope_1 = \frac{2}{4}\frac{300}{150} = 1 \; ; \; slope_2 = \frac{1}{4}\frac{300}{150} = 0{,}5$$

Evalaución sensible al contexto (skew)



Cálculo de curvas ROC

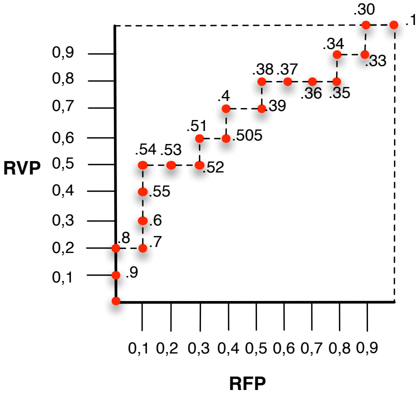
- Para calcular una curva ROC para un modelo de clasificación hay que tener en cuenta el tipo de clasificador.
  - **Clasificador Discreto (Crisp):** predice la clase a partir de un conjunto de clases predefinidas.
  - **Clasificador Probabilístico (Soft):** a parte de la clases da información sobre el grado de credibilidad sobre la pertenencia a dicha clase:
    - La mayor parte de los clasificadores se pueden convertir en probabilísticos $\rightarrow$ bastaría con calcular las probabilidades de clase.
    - Escalado o calibración de Platt $\rightarrow$ ajuste a una función sigmoide.

Cálculo de curvas ROC

- Todos los clasificadores probabilísticos (para problemas de dos clases) nos permiten definir alguna medida que permita indicar el grado con el que la instancia pertenece a una clase.
    - Con Naïve Bayes podemos utilizar la probabilidad.
- Dicha medida hace posible una ordenación de las instancias.
- Para conseguir un clasificador discreto sólo hace falta definir un umbral a partir del cual se considera que la instancia pertenece a la clase positiva.
- Distintos valores del umbral nos permiten obtener distintos clasificadores con valores de sensibilidad y especificidad distintos $\rightarrow$ Curva ROC.

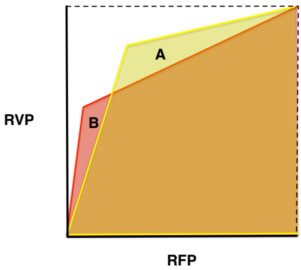
Cálculo de curvas ROC:ejemplo

Nº	Clase	Punt.
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.37
13	p	.38
14	n	.36
15	n	.35
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1



Al lado de cada punto figura el umbral que lo genera.

Comparación de curvas ROC



- ¿Qué clasificador es mejor? ¿El A o el B?
- El clasificador con mayor área bajo la curva (AUC) mayor (en nuestro caso el A).
- La medida AUC es una alternativa al error de clasificación.
- Se prefieren aquellas técnicas que generen clasificadores con mayor AUC.
- Para el caso de curvas ROC de un sólo punto:

$$AUC = \frac{Sensibilidad + Epecificidad}{2}$$

Área bajo la curva ROC

- El test de Wilconxon-Mann-Whitney y la medida AUC 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.

- Sin embargo esto no garantiza que los clasifique bien.
  - Pero garantiza que existe un umbral que los clasifique bien.

Área bajo la curva ROC.

- Para un clasificador probabilístico, AUC evalúa la capacidad del clasificador para ordenar sus predicciones de acuerdo con la medida de confidencia utilizada.
- Obviamente, el error de clasificación y el AUC están relacionados
  - Si el AUC está cercano a 1 el error estará cercano a 0.
  - Pero si el error está cercano a 0 puede que el AUC no esté cercano a 1.

Tests estadísticos I

- Llegados a este punto podemos calcular la eficiencia de un clasificador/predictor con bastante precicisón.
  - Tenemos diferentes técnicas para calcularla.
  - Disponemos, además, de diferentes medidas.
- Pero estas medidas no son suficientes para determinar qué clasificador es mejor sobre un conjunto de datos.
  - ¿Son significativas las diferencias entre las medidas?

Tests estadísticos II

- **Objetivo:** Dados dos técnicas de clasificación. A y B, y un conjunto de datos S ¿qué técnica producira el clasificador más preciso a paritr de conjuntos del mismo tamaño?
- Sea $\hat{f}_A$ y el $\hat{f}_B$ los clasificadores generados por las técnica 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.

Tests estadísticos III

- 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), para cuando queramos comprobar el rendimiento de un algoritmo concreto con el rendimiento de otros algoritmos en un problema concreto.
    - Un algoritmo particular o nuevo en un dominio concreto.
  - Comparar varios algoritmos en un mismo domino. El mismo caso que el anterior, pero la comparación de rendimiento de realiza frente a un conjunto de algoritmos de referencia.
    - Un algoritmo particular frente a otro de forma genérica.
  - Comparar varios algoritmos en varios dominios. Un análisis más amplio de diferentes algoritmos en un conjunto de dominios de referencia o en un problema concreto.

Dos clasificadores en un dominio: Test t de Student por pares I

- Este test nos va a permitir determinar si las diferencias entre las medias de dos medidas pareadas es significativa
  - Es decir, si proceden de la misma población.
- Evaluar el rendimiento de los clasificadores un determinado número de veces m en el mismo dominio.
- Cada vez se obtiene un conjunto de entrenamiento R y de prueba T distintos.
- Este proceso nos lleva a que al final vamos a obtener m medidas de error distintas, p_A^i y p_B^i con $i = 1, \cdots, m$.

Dos clasificadores en un dominio: Test t de Student por pares II

- Calculamos m diferencias $p^i = p^i_A - p^i_B$ y calculamos el estadístico:

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

- donde $\overline{p} = \frac{1}{n}\sum_{i=1}^{m} p^i$
- El estadístico t sigue una distribución t de Student con $m-1$ grados de libertad.
- No rechazamos la hipótesis nula si $|t| \leq t_{m-1,1-\alpha/2}$ con una significancia de α.

Dos clasificadores en un dominio: Test t de Student por pares III

- Aplicación:
  1. Realizamos las m particiones del conjunto S: $R_1, \cdots, R_m$ y $T_1, \cdots, T_m$
  2. Calculamos las diferencias de las medidas de error $p^i = p^i_A - p^i_B$
  3. Calculamos el estadístico t.
  4. No rechazamos la hipótesis nula si $|t| \leq t_{m-1,1-\alpha/2}$ con una significancia $\alpha/2$.
     - Con $m = 30$ y $\alpha/2 = 0,05$, $t_{29,0,975} = 2{,}04523$.

Tamaño del efecto

- El test t de Student nos indica si la diferencia entre las medidas de rendimiento son significativas
  - Pero no nos dice cuán importante es dicha diferencia.
- Para ello debemos calcular el estadístico d de Cohen.

$$d_{cohen} = \frac{\overline{p_A} - \overline{p_b}}{\sigma_p} \quad , \quad \sigma_p = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

- Interpretación:
  - d_{cohen} sobre 0.2 o 0.3 indica que el tamaño del efecto es pequeño pero probablemente significativo.
  - d_{cohen} sobre 0.5 indica un efecto medio pero apreciable.
  - d_{cohen} sobre 0.8 indica un efecto grade.

Condiciones de aplicabilidad del test t de Student I

- **Normalidad**. Las muestras deben proceder de poblaciones normalmente distribuidas.
  - El test t de Student es bastante robusto si no se cumple esta condición.
  - Sería suficiente con que unos conjuntos de test con más de 30 muestras.
  - Con validación cruzada necesitamos un data set con más de $10 \times 30 = 300$ instancias.
  - Alternativamente se pueden utilizar test estadísticos para su comprobación: Kolmogorov-Smirnov, Shapiro-Wilk or Anderson-Darling.

Condiciones de aplicabilidad del test *t* de Student II

- **Alatoriedad de la muestra.** Se asume que las muestras utilizadas para calcular las medias son representativas.
  - Las muestras deben haber sido seleccionadas de forma independiente e idénticamente distribuidas a partir de una distribución normal.
  - Difícil de comprobar, debemos confiar en las personas que construyeron el conjunto de datos.
- **Homocedasticidad: Igualdad de varianzas en las poblaciones**. Las muestras deben proceder de poblaciones con la misma varianza
  - Se puede comprobar de forma visual mediante un gráfico de cajas
  - Se pueden utilizar los tests: Finger, Barlett, Levene o Brown-Forsythe.

Técnicas de muestreo para el test t de Student I

- **Hold-out con repetición**, lo más habitual es 30 veces.
  - Los conjuntos de entrenamiento se solapan $\rightarrow$ no se cumple la condición de normalidad.
  - Los conjuntos de entrenamiento se solapan $\rightarrow$ las muestras no son independientes.
  - Alta probabilidad de un error de tipo I (no se acepta H_0 siendo cierta).
    - Aumenta con las repeticiones.
  - Se puede utilizar la corrección de Nadeu y Bengio:

$$t = \frac{\overline{p}\sqrt{n}}{\sqrt{(\frac{1}{n} + \frac{|Train|}{|Test|})\sum_{i=1}^{m}(p^i - \overline{p})^2}}$$

Técnicas de muestreo para el test t de Student II

- **Validación cruzada con k-pliegues**. En este caso se realiza una validación cruzada con k-pliegues.
  - Tiene la ventaja de que los conjuntos de prueba son independientes.
  - Sin embargo, los conjuntos de entrenamiento se solapan.
    - En una validación cruzada con 10 pliegues los conjuntos de entrenamiento comparten el 80 % de los casos.
  - Para favorecer la replicabilidad se suele repetir el proceso unas 10 veces.

Técnicas de muestreo para el test t de Student III

- **Test t de Student por pares en validación cruzada 5x2**. Se realizan 5 repeticiones de una validación cruzada con dos pliegues. Esto da lugar 5 particiones del conjunto de datos en dos conjuntos de igual tamaño.

$$\{R_1^1, R_2^1, R_3^1, R_4^1, R_5^1\} \text{ y } \{R_1^2, R_2^2, R_3^2, R_4^2, 5_5^2\}$$

  - En cada iteración se generan dos modelos, uno entrenado con R_i^1 y validado con R_i^2, y el otro entrenado sobre R_i^2 y validado sobre R_i^1.
  - Es decir, cinco repeticiones de una validación cruzada con 2 pliegues.

Técnicas de muestreo para el test *t* de Student IV

- El hecho de que los conjuntos de entrenamiento sean disjuntos en cada iteración, hace que las medidas sean más independientes que en el caso de una validación cruzada con 10 pliegues.
- Sin embargo, presenta el problema de que los conjuntos de entrenamiento son del mismo tamaño que los de tests.

Test de McNemar's I

- Es la alternativa no paramétrica al test t de Student.
- Se divide el conjunto de datos S en conjunto de entrenamiento R y de prueba T.
- Se generan los dos clasificadores $\hat{f}_A$ y $\hat{f}_B$.
- Se genera la siguiente tabla de contingencia:

n_{00} = nº de casos mal clasificados por $\hat{f}_A$ y $\hat{f}_B$	n_{01} = nº de casos mal clasificados por $\hat{f}_A$ y bien por $\hat{f}_B$
n_{10}= nº de casos bien clasificados por $\hat{f}_A$ y mal $\hat{f}_B$	n_{11} = nº de casos bien clasificados por $\hat{f}_A$ y $\hat{f}_B$

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

Test de McNemar's II

- **Hipótesis nula**: $n_{01} = n_{10}$
- El test de McNemar esta basado en el siguiente estadístico que se ajusta a una distribución χ_1^2:

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

- No se rechaza la hipótesis nula de que ambos clasificadores tiene el mismo error con una significancia α si $M \leq \chi^2_{\alpha,1}$.

Test de McNemar's III

- Para aplicar el test de McNemar para comparar dos clasificadores:
    1. Comprobar que $n_{01} + n_{10} > 20$
    2. Se genera la tabla de contingencia anteriormente descrita quedándonos con los valores n_{01} y n_{10}.
    3. Calculamos el estadístico.
    4. La hipótesis nula es: "ambos clasificadores tienen el mismo ratio de error".
    5. No se rechaza la hipótesis nula si el estadístico es menor que 3.85 con el 95 % de confianza.
    6. Si no, el mejor clasificador es aquél que presenta menor error.

Test de McNemar's IV

- **Desventajas:**
  - No tiene en cuenta la aleatoriedad intrínseca de la técnica y de la partición de S.
  - Las técnicas sólo se comparan usando un único conjunto de entrenamiento.
  - Sólo aplicable si creemos que dicha aleatoriedad es pequeña.
  - Se debe asumir que la diferencia observada en R se mantiene en S.
  - En el caso de problemas multiclase no se puede aplicar $\rightarrow$ test de la homogeneidad marginal.

Dos clasificadores en un dominio conclusiones I

- Experimentos sugieren que el test basado en la validación cruzada 5x2 es el más potente y satisfactorio al utilizar diferentes conjuntos de entrenamiento y prueba.
    - Tiene un error de tipo I aceptable (fallo en determinar que los dos modelos producen resultados similares)
    - Puede fallar cuando los ratios de error en cada iteración varían mucho.
    - Esto nos lleva a una mala estimación de la varianza.
- A pesar de que el test de McNemar no tiene en cuenta el efecto de utilizar diferentes conjuntos de entrenamientos, también genera buenos resultados.

Dos clasificadores en un dominio conclusiones II

- El test basado en la validación cruzada con 10 pliegues también es potente, pero presenta un error de Tipo I alto.
  - Por lo que es recomendable en los casos en los que el error de tipo II (fallo en la detección de una diferencia real entre los modelos) sea más importante.

Dos Clasificadores en varios dominios

- Una situación más usual que el caso anterior es la comparación de dos clasificadores en varios dominios.
    - En este caso estaríamos comparando de forma genérica las diferencias entre los clasificadores.
- Primera opción: Extender los test anteriores a varios dominios.
    - El test t de Student asume 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 signos de Wilcoxon para muestras pareadas I

- Es un test no paramétrico para comparar las medianas.
- Se utiliza como alternativa al test t de Student (también es conocido como test t de Wilcoxon).
- Comprueba si hay diferencias entre las medianas.
  - En caso de que se cumplan las condiciones de aplicabilidad, se podría aplicar el test t de Student.
  - Sirve para comparar diferentes pruebas realizadas (datasets) sobre la misma población bajo dos circunstancias distintas (clasificadores).
- Supongamos que tenemos dos clasificadores $\hat{f}_A$ y $\hat{f}_B$, evaluados sobre n dominios distintos.

Test de los rangos con signos de Wilcoxon para muestras pareadas II

1. Sea p_A^i y p_B^i las medidas de rendimiento de cada clasificador en el domino i.
2. Se calculan las diferencias entre las medidas para cada dominio $d_i = p_A^i - p_B^i$.
3. Se ordena d_i de menor a mayor de su valor absoluto 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):

$$W_{s_1} = \text{Suma en valor absoluto de los rangos positivos}$$
$$W_{s_2} = \text{Suma en valor absoluto de los rangos negativos}$$

Test de los rangos con signos de Wilcoxon para muestras pareadas III

5. Se calcula el estadístico $T_{Wilcox} = min(W_{s_1}, W_{s_2})$ que sigue una distribución T de Wilcoxon.
6. En el caso de que $n > 25$ es estadístico T_{Wilcox} puede ser aproximado por una distribución normal.
   - 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 normal de la distribución T_{Wilcox} en el caso de que la hipótesis nula sea cierta.

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

Test de los rangos con signos de Wilcoxon para muestras pareadas IV

- El valor del estadístico z_{Wilcox} se puede comparar con el valor crítico en la tablas de la distribución normal.

7. En ambos casos, 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.

Adaptación del test de Wilcoxon para un sólo 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 de mejores resultados que otro.
  - Sobre todo si el clasificador es robusto respecto al orden o permutaciones de los ejemplos.
- Para evitar esto se recomienda utilizar la validación cruzada sin repetición.

Varios clasificadores en varios dominios I

- Nos permite evaluar varias estrategias de aprendizaje:
  - en varios conjuntos de referencia para analizar las características generales de los algoritmos.
  - en varios conjuntos del mismo problema para ver cuál es la mejor aproximación.
- Podemos pensar en hacer comparaciones dos a dos mediante el test t de Student.
  - Muchos test para poder realizar todas las comparaciones.
  - A medida que aumenta el número de test aumenta la probabilidad de cometer un error de tipo I $\rightarrow$ ajuste del valor p.

Varios clasificadores en varios dominios II

- En estadística podemos encontrar tests que evitan realizar todas las comparaciones dos a dos.
- Estos tests (paramétricos o no paramétricos) nos permiten realizar contrastes de varias hipótesis al mismo tiempo $\rightarrow$ tests omnibus.
  - Al menos existen dos diferencias que son significativas.
- 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.

ANOVA de una vía con medidas repetidas I

- Al igual que el test t de Student compara las diferencias observadas entre las medias.
- Sin embargo permite determinar si las diferencias observadas entre cualquier número de medias es estadísticamente significativa.
  - $H_0 : \mu_0 = \mu_1 = \cdots = \mu_n$
  - H_1 : 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.

ANOVA de una vía con medidas repetidas II

- 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 los 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 cuadrados entre los grupos.
    - Esto se pude comprobar con un test F, que determina si el ratio de dos varianzas, medidas como suma de cuadrados, es significativamente mayor que 1.
- Para medir el tamaño del efecto se utiliza el estadístico η^2 o la f de Cohen.

Condiciones de aplicabilidad del test ANOVA

- **Normalidad**: Las muestras debe ser extraídas de forma independiente y estar igualmente distribuidas a partir de una distribución normal.
- **Homogeneidad de las varianzas (Esfericidad)**: La varianza en cada grupo debe ser similar.
  - Test: test de Mauchly's.
- Las medidas de rendimiento deben tener la misma escala.
- Los conjuntos de datos deben tener aproximadamente el mismo tamaño.

La necesidad de una alternativa paramétrica

- 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 además tenemos medidas de rendimiento categóricas o no monótonas que incumplen la condición de la escala.
- Aternativa: Test de Friedman.

Test de Friedman

- El test de Friedman es la alternativa no paramétrica al test ANOVA con medidas repetidas.
- En este caso se comparan las medianas en vez de las medias.
  - H_0 : todas las medianas son iguales.
  - H_1 : al menos dos medianas difieren.
- 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 medidas repetidas vs. Friedman

- El test ANOVA es relativamente robusto a la condición de normalidad.
  - Sin embargo, en el caso de comparación de clasificadores la condición de esfericidad es muy difícil de comprobar.
  - Si se cumplen las condiciones de aplicabilidad en más potente.
- El test de Friedman es más potente en el caso de que no se cumplan las condiciones.
  - Incluso en el caso de que se cumplan las condiciones no suelen existir muchas diferencias entre los tests.
- Puede darse el caso de que un test omnibus detecte diferencias significativas pero los tests post hoc no.
  - Si esto ocurre $\rightarrow$ existen diferencias pero no se pueden identificar debido a la escasa potencia de los tests post hoc.

Tests Post hoc

- Los tests omnibus anteriormente comentados sólo nos dicen si hay diferencias significativas entre los clasificadores.
- Si existen diferencia (se rechaza la hipótesis nula) habría que localizar dónde están dichas diferencias.
- Para ello hay que aplicar los test Post Hoc.
- Al igual que para el caso de los test omnibus existen version paramétricas y no paramétricas.

Tests post hoc paramétricos I

- Estos test se aplicarían en el caso de que el test ANOVA de medidas repetidas indique que hay diferencias significativas.
- **Test de Tukey**. 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.
  - A diferencia del Test t, este test se utiliza una especie de error de estándar de propósito general para comparar cualquier par de medias.
  - Tiene menos probabilidad de cometer un error de tipo I que el test t.

Tests post hoc paramétricos II

- **Test de Dunnet**. Se puede utilizar cuando las comparaciones no son dos a dos, sino de todos los clasificadores con uno de control.
- **Test de Bonferroni**. Equivalente al anterior, sólo 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.

Tests post hoc no paramétricos

- Estos test se aplicarían 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 hacer consideraciones similares a las que se hicieron cuando se adaptó el test de Wilcoxon para un sólo dominio.
- Este caso también se pueden generar varios conjuntos de datos a partir del disponible utilizando alguna técnica de muestreo.
- Pero hay que tener cuidado ya que un clasificador puede predominar sobre los otros.
- Se pueden aplicar los test post hoc directamente.

Test estadísticos analizados

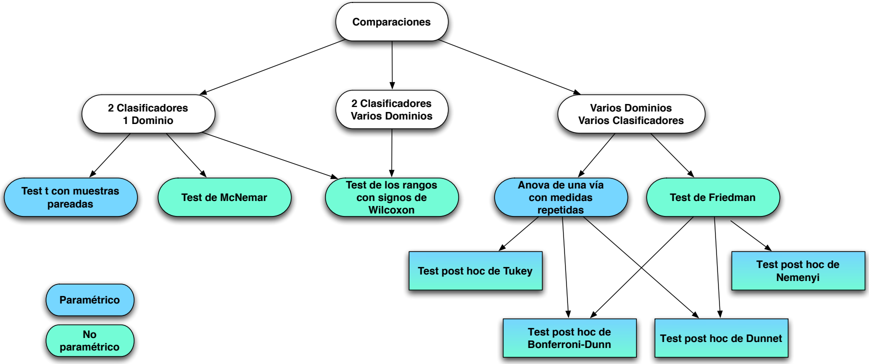


Figura: Test estadísticos para comparar clasificadores.

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.
- 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 relacionada I

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

Bibliografía relacionada II

- Guerrero Vázquez, Elisa, Yañez Escolano, Andrés and Galindo Riaño, Pedro and Pizarro Junquera, Joaquín. *Repeated measures multiple comparison procedures applied to model selection in neural networks*. Bio-Inspired Applications of Connectionism. pp 88-95. Springer-Verlag.
- José Hernández Orallo, Mª José Ramírez Quintana and César Ferri Ramirez. *Introducción a la Minería de Datos*. Pearson-Prentice-Hall. 2004
- José Hernández Orallo. Classifier Evalaution in Data Mining: ROC Analysis. http://users.dsic.upv.es/ jorallo/Albacete/
- C. Nadeau and Y Bengio. Inference for generalization error. *Machine Learning*, 52:239-281, 2003.

Bibliografía relacionada III

- Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.