Minería de Datos

Tema 4 - Comparación de Modelos

Nombre: Alejandro pérez Belando

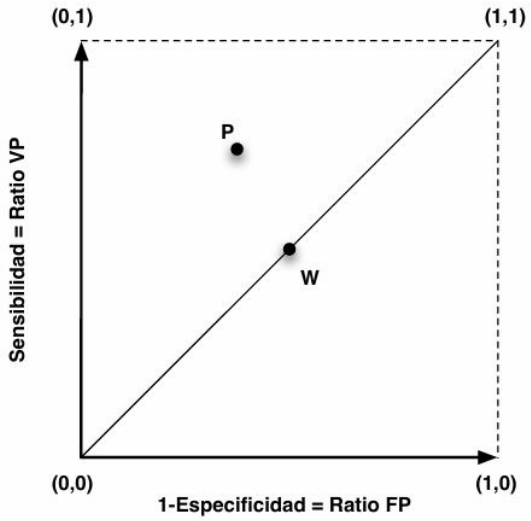
---

1. Introducción

Las técnicas desarrolladas en el tema anterior permite construir o bien varios modelos, o bien distintas versiones de un mismo modelo. En este tema nos centramos en cómo comparar los diferentes modelos entre sí.

2. Curvas ROC (Receiver Operating Characteristics)

Inicialmente fue ideado para problemas de dos clases (aunque tiene extensión multiclase). Nos permite comparar distintos modelos en un espacio bidimensional:

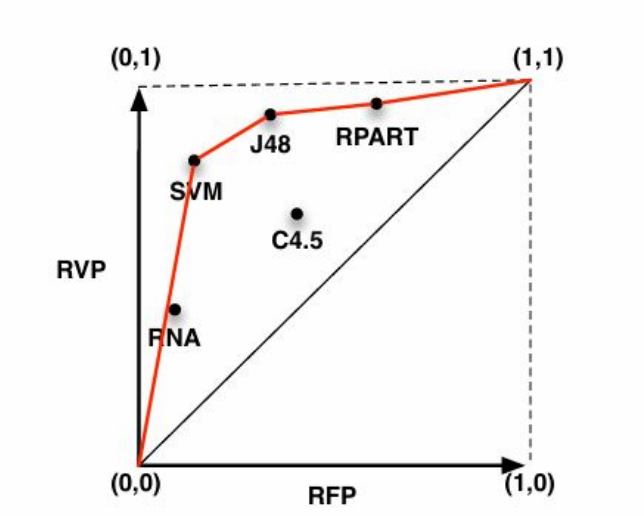


- **Eje x**: ratio de falsos positivos (también se le conoce como 1-especificidad).
- **Eje y**: ratio de verdaderos positivos (sensibilidad).

- **Punto (0,1)**: clasificador perfecto. Predice correctamente todos los positivos y no da falsos negativos.
- **Punto (0,0)**: lo predice todo como clase positiva.
- **Punto (1,0)**: lo predice todo como clase negativa.
- **Recta (diagonal) que une (0,0) y (1,1)**: clasificador aleatorio. Da el mismo número de verdaderos positivos que de falsos positivos.

Por lo tanto, lo más interesante es obtener clasificadores que se acerquen lo máximo posible al punto (0,1), siendo un buen clasificador aquel que está bastante por encima de la diagonal y uno malo aquel que está por debajo.

Si tenemos varios clasificadores, se representa cada uno en el espacio ROC y se calcula la **envolvente convexa**:



Todo modelo por debajo de la envolvente convexa se descarta.

2.1. Evaluación sensible al contexto (skew)

Se usa para calcular el mejor modelo. La eficacia depende de la matriz de costes (no todos los errores pesan igual) y del contexto de la distribución de clases. Estos dos conceptos se agrupan en la **pendiente (slope)**:

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

Donde $N \equiv$ número de ejemplos negativos y $P \equiv$ número de ejemplos positivos.

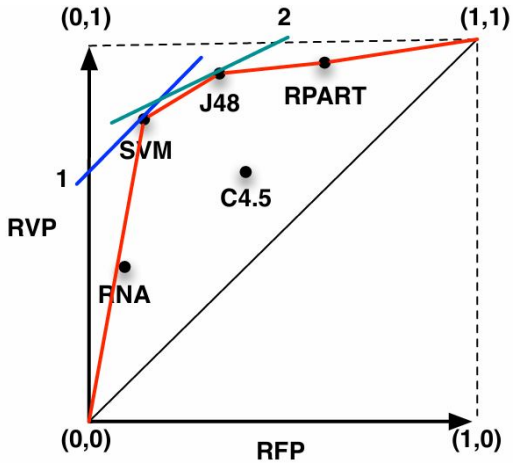
Para determinar el modelo más apropiado, se debe:

1. Trazar una recta cuya pendiente sea la *slope* calculada en el punto (0,1).
2. Trasladar la recta hasta la curva ROC.
3. El primer punto que toque la recta es el menor modelo.

Ejemplo
Nuestro conjunto de prueba de ejemplo tiene 300 clases negativas y 150 clases positvas con dos casos:

- Caso 1: Coste(FP) = 2 ; Coste(FN) = 4. Entonces: $slope_1 = \frac{2}{4}\frac{300}{150} = 1$
- Caso 2: Coste(FP) = 1 ; Coste(FN) = 4. Entonces: $slope_2 = \frac{1}{4}\frac{300}{150} = 0{,}5$

Es decir, tenemos dos pendientes, una de 1 y otra de 0.5 lo siguiente es llevarnos la recta hasta el punto (0,1) e ir bajándola hasta que toque uno de los puntos de la figura anterior. El resultado es el siguiente:



La recta del caso 1 (señalada en azul) toca primero el punto que representa las máquinas de vectores soporte, mientras que la del caso 2 (señalada en verde) representa otro modelo.

Nota: si se desconocen estos datos, basta con suponer *slope* = 1 y se selecciona el punto más cercano a (0,1).

2.2. Cálculo de curvas ROC

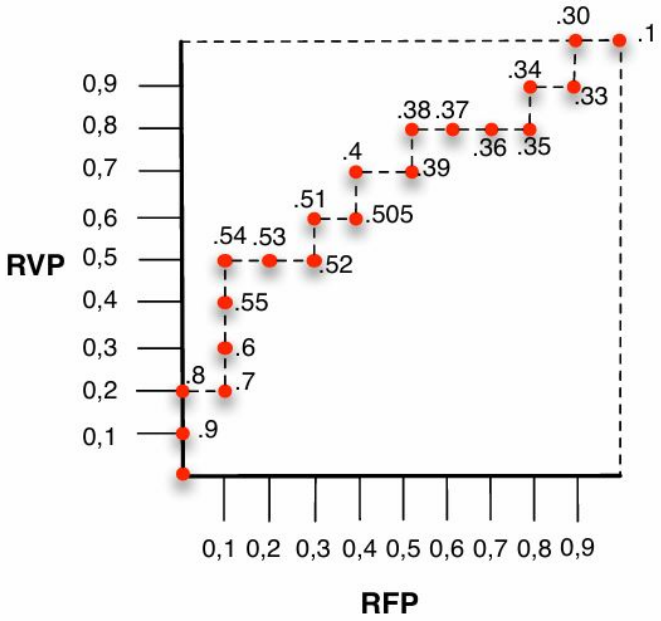
Para ello, hay que tener en cuenta el tipo del clasificador, que puede ser:

- **Discreto (*crisp*)**: predice la clase a partir de un conjunto de clases predefinidas. Para conseguir un clasificador discreto solo hace falta definir un umbral a partir del cual se considera que la instancia pertenece a la clase positiva.

- **Probabilístico (*soft*)**: además de las clases, da información sobre cómo de creíble es que pertenezca a dicha clase; esto hace posible la ordenación de instancias. La mayoría de los clasificadores pueden convertirse en probabilísticos.

Ejemplo
En este caso tenemos una tabla con 20 ejemplos y clasificados en positivos (p) y negativos (n). Para calcular la curva ROC vamos a comenzar en el punto (0,0) e iremos moviendo una casilla hacia arriba (sumamos al ratio de verdaderos positivos) cuando la clasificación sea positiva y una hacia la derecha (sumamos al ratio de falsos positivos) cuando sea negativa. Al lado de cada punto en la gráfica se muestra la puntuación (el umbral que lo genera).

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



2.3. Área bajo la curva ROC (AUC)

El mejor clasificador es aquel que tiene mayor área bajo la curva (AUC).

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.

Esto no garantiza que se clasifique bien, pero sí que hay un umbral que los clasifique bien. Para un **clasificador probabilístico**, AUC evalúa la capacidad del clasificador para ordenar sus predicciones de acuerdo con la medida de confianza usada.

Relación del AUC con el error de clasificación:

- Si el AUC es cercano a 1, el error de clasificación es cercano a 0.
- Sin embargo, que el error de clasificación sea cercano a cero, no quiere decir que el AUC sea cercano a 1.

3. Test estadísticos

Sea un conjunto de datos (S) y dos técnicas de clasificación (A y B) con clasificadores respectivos (f_A y f_B) respectivamente a partir del conjunto de entrenamiento (R), la **hipótesis nula (H_0)** para R es que las dos técnicas producirán clasificadores con la misma tasa de error/acierto.

La elección del test estadístico depende de la situación en la que estemos:

- Comparar dos algoritmos en un mismo dominio (dataset).
- Comparar varios algoritmos en un mismo dominio.
- Comparar varios algoritmos en varios dominios.

4. Dos clasificadores en un dominio

4.1. Test *t* de Student (por pares)

Este test permite, por un lado, determinar si las diferencias entre las medias de dos medidas pareadas es significativa (si pertenecen a la misma población) y por el otro, evaluar el rendimiento de los clasificadores m veces en el mismo dominio.

Loa pasos para la aplicación son:

1. Realizar m particiones del conjunto de datos S en datos de entrenamiento $(R_1, \ldots, R_m)$ y prueba $(T_1, \ldots, T_m)$, de forma que se obtienen distintas medidas de error (p_A^i y p_B^i; donde $i = 1, \ldots, m$).
2. Calcular las m diferencias de las medidas de error ($p^i = p_A^i - p_B^i$).
3. Calcular el estadístico t:

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

4. No rechazar la hipótesis nula si $|t| \leq t_{m-1,1-\alpha/2}$, donde $\alpha \equiv$ significancia

   Por ejemplo, cuando $m = 30$ y $\alpha = 0{,}05$, entonces $t_{29,0,975=2,04523}$

   A efectos prácticos, si $|t| < 0{,}05$, rechazamos la hipótesis nula.

4.1.1. Estadístico d de Cohen

como el estadístico t de Student nos dice si la diferencia entre las medidas de rendimiento son significativas, pero no nos dice cómo de significativas son, usaremos le estadístico d de Cohen:

$$d_{Cohen} = \frac{\bar{P}_A - \bar{P}_B}{\sigma_p}; \qquad \sigma_p = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

Interpretación este estadístico:

- $d_{Cohen} \approx 0{,}2$ o $d_{Cohen} \approx 0{,}3$: el tamaño del efecto es pequeño pero probablemente significativo.
- $d_{Cohen} \approx 0{,}5$: efecto medio pero apreciable.
- $d_{Cohen} \approx 0{,}8$: efecto grande.

4.1.2. Condiciones de aplicabilidad del *t* de Student

- **Normalidad**: las muestras deben proceder de poblaciones con distribución normal. Si no se cumple esta condición, el test t de Student es bastante robusto. Sería suficiente con un dataset con 300 instancias.
- **Aleatoriedad de la muestra**: se asume que las muestras usadas para calcular las medias son representativas.
- **Igualdad de varianzas en las poblaciones (homocedasticidad)**: se puede comprobar con un gráfico de cajas.

4.1.3. Técnicas de muestreo para el *t* de Student

- **Hold-out con repetición**: lo más habitual es repetir 30 veces. Sin embargo, los conjuntos de entrenamiento se solapan, por lo que no se cumple la condición de normalidad y por lo tanto las muestras no son independientes; esto lleva a una alta probabilidad de error de tipo 1 (No se acepta H_0 aún siendo cierta).

  Se puede emplear la corrección de Nadeu y Bengio:

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

- **Validación cruzada con k-pliegues**: tiene la ventaja de que los conjuntos de prueba son independientes, pero los de entrenamiento se solapan (en una 10-fold CV los conjuntos de entrenamiento comparten el 80 % de los casos). Presenta error de tipo I, por lo que podría ser interesante aplicarla cuando el error de tipo II sea más importante.

- **Test t de Student por pares en validación cruzada 5x2**: 5 particiones del cnjunto de datos en dos conjuntos de igual tamaño (5 repeticiones de una CV de 2 pliegues):

$$\left\{R_1^1, R_2^1, R_3^1, R_4^1, R_5^1\right\} \ y \ \left\{R_1^2, R_2^2, R_3^2, R_4^2, R_5^2\right\}$$

  En cada iteración se generan dos modelos:

  - Uno entrenado con R_i^1 y validado por R_i^2
  - Y otro entrenado con R_i^2 y validado por R_i^1

  En este caso las medidas son más independientes que en el de 10-fold CV, pero los conjuntos de entrenamiento son del mismo tamaño que los de test.

4.2. Test de McNemar's

Es la alternativa **no paramétrica** al test t de Student. Se divide el conjunto de datos S en entrenamiento R y prueba T y se generan los dos clasificadores f_A y f_B. Se genera la siguiente tabla de contingencia:

$n_{00} = n^\circ$ de casos mal clasificados por $\hat{f}_A$ y $\hat{f}_B$	$n_{01} = n^\circ$ de casos mal clasificados por $\hat{f}_A$ y bien por $\hat{f}_B$
$n_{10} = n^\circ$ de casos bien clasificados por $\hat{f}_A$ y mal $\hat{f}_B$	$n_{11} = n^\circ$ de casos bien clasificados por $\hat{f}_A$ y $\hat{f}_B$

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

En este caso, la **hipótesis nula (H_0)**: $n_{01} = n_{10}$

El **test de McNemar's** se basa 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 rechaza la hipótesis nula si $M \leq \chi^2_{\alpha,1}$, donde $\alpha \equiv significancia$; a efectos prácticos,si el estadístico es menor que 3.85 con el 95 % de confianza.

Desventajas del test:

- No tiene en cuenta la **aleatoriedad intrínseca** de la técnica de la partición de S. Solo se puede aplicar si creemos que esa aleatoriedad es pequeña.
- Las técnicas solo se comparan usando **un único conjunto de entrenamiento**.
- **No** se puede aplicar para **problemas multiclase**, hay que usar el test de la homogeneidad marginal.

5. Dos clasificadores en varios dominios

La comparación de dos clasificadores en varios dominios es más usual que la de dos clasificadores en un dominio. Una opción sería extender los test anteriores a varios dominios, pero el test de McNemar no está pensado para más de dos clasificadores, por lo que se recomienda usar el **test de los rangos con signo de Wilcoxon** (test t de Wilcoxon)

5.1. Test de rangos con signo de Wilcoxon para muestras pareadas

Se trata de un test no paramétrico para comprobar si hay diferencias en las medianas, por lo que puede considerarse una alternativa al test t de Student

Sean dos clasificadores $\hat{f}_A$ y $\hat{f}_B$, evaluados sobre n dominios distintos, y p_A^i y p_B^i las medidas de tendimiento de cada clasificador en el dominio i. Los pasos para la aplicación son:

1. **Calcular las diferencias** entre las medidas: $d_i = p_A^i - p_B^i$
2. **Ordenar** $|d_i|$ de mayor a menor y asignarles un rango. En caso de empate se asignan la media de los rangos empatados.
3. **Calcular**:
   - W_{S1} = suma en valor absoluto de los rangos positivos
   - W_{S2} = suma en valor absoluto de los rangos negativos
4. Calcular el **estadístico**: $T_{Wilcoxon} = min(W_{S1}, W_{S2})$

   En el caso de que $n > 25$, el estadístico $T_{Wilcoxon}$ puede ser aproximado a una normal. El estadístico en ese caso es:

   $$z_{Wilcoxon} = \frac{T_{Wilcoxon} - \mu_{T_{Wilcoxon}}}{\sigma_{T_{Wilcoxon}}} \begin{cases} \mu_{T_{Wilcoxon}} = \dfrac{n(n+1)}{4} \\ \sigma_{T_{Wilcoxon}} = \sqrt{\dfrac{n(n+1)(2n+1)}{24}} \end{cases}$$

   Donde $\mu_{T_{Wilcoxon}}$ y $\sigma_{T_{Wilcoxon}}$ son la media y desviación estándar (en el caso de que la hipótesis nula sea cierta).
5. se rechaza la hipótesis nula si el estadístico es menor que el valor crítico (dados unos grados de libertar y un nivel de significancia)

5.2. Adaptación del test de Wilcoxon para un solo dominio

Consiste en generar varios conjuntos de datos a partir del disponible, bien permutándolo o reordenándolo, o bien usando alguna técnica de muestreo como es el bootstrappingng, hold-out o CV.

Sin embargo, corremos el riesgo de que un clasificador dé siempre mejores resultados que el otro, por lo que para evitarlo se recomienda usar la validación cruzada sin repetición.

6. Varios clasificadores en varios dominios

Permite evaluar varias estrategias de aprendizaje, o bien en varios conjuntos de referencia (para analizar las características generales de los algoritmos), o bien en varios conjuntos del mismo problema (pera ver cuál es la mejor aproximación).

Podríamos pensar en hacer comparaciones dos a dos mediante el test t de Student, pero podemos encontrar tests (paramétricos y no paramétricos) que nos permitan realizar contrastes de varias hipótesis al mismo tiempo: tests omnibus.

6.1. ANOVA de una vía para maedidas repetidas

Es un test ombibus que compara las diferencias observadas entre las medias (al igual que el t de Student).

- Hipótesis nula: igualdad de las medias $H_0 = \mu_0 = \mu_1 = \cdots = \mu_n$
- Hipótesis alternativa: al menos dos medias son distintas.

Como idea general, se divide la **varianza total** en:

- Varianza causada por el **error aleatorio** (dentro de los datasets)
- varianza causada por las **diferencias** observadas entre las **medias** (entre los grupos)

Si se cumple la hipótesis nula suma de cuadrados dentro de los grupos $\approx$ suma de cuadrados entre los grupos. Esto se puede comprobar con un test f de Cohen.

6.1.1. Condiciones de aplicabilidad del test ANOVA

- Normalidad: las muestras deben extraerse 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.
- Las medidas del rendimiento deben tener la misma escala.
- Los conjuntos de datos deben tener el mismo tamaño.

Sin embargo, se desaconseja el uso de este test debido a la dificultad de comprobar la esfericidad y a que comúnmente hay medidas categóricas que incumplen la condición de escala.

6.2. Test de Friedman

Otro test omnibus. Alternativa no paramétrica al test ANOVA con medidas repetidas. En este caso se comparan las **medianas** en lugar de las medias:

- Hipótesis nula (H_0): igualdad de las medianas
- Hipótesis alternativa (H_1): al menos dos medianas son distintas.

El test de Friedman es más potente en caso de que no se cumplan las condiciones de esfericidad y aplicabilidad requeridas para el test ANOVA. Incluso en el caso de que se cumplan las condiciones, no suele haber muchas diferencias entre los test.

6.3. tests Post Hoc

Los dos tests omnibus presentados anteriormente, solo nos dicen si hay diferencias significativas entre los clasificadores.

En el caso de que existieran, es decir, se **rechaza la hipótesis nula**, habría que localizar dónde están las diferencias mediante los tests **Post Hoc**.

6.3.1. tests Post Hoc Paramétricos

Se aplican en el caso de que el test ANOVA de medias repetidas indique que hay diferencias significativas:

- **Test de Tukey**: intenta detectar variación aleatoria entre todos los pares de medias y se comparan con las diferencias reales. El estadístico nos dice cuan grande es la diferencia comparada con la variación general aleatoria entre medias. Tiene menos probabilidad de cometer error de tipo I.
- **Test de Dunnet**: se puede usar cuando las comparaciones no son dos a dos, sino de todos los clasificadores con uno de control.
- **Test de Bonferroni**: Equivalente al anterior, pero se usa la corrección de Bonferroni para las comparaciones.
  - N° de comparaciones pequeño: funciona bien.
  - N° de comparaciones grande: tiende a ser conservador.
- **Test de Bonferroni-Dunn (test de Dunn)**: Intenta corregir el conservacionismo del test de Bonferroni dividiendo la significancia (α) por el N° de comparaciones a realizar.

6.3.2. tests Post Hoc no Paramétricos

Se aplican en el caso de que el test ANOVA de medias repetidas indique que hay diferencias significativas:

- **Test de Nemenyi**: se basa en un estadístico que mide la diferencia promedio entre los rangos de los clasificadores.
- **Otros métodos**: se basan en escalar los nieveles de significancia (Test de Homel, Test de Holm y Test de Hochberg).

7. Varios clasificadores en un dominio

Se pueden hacer consideraciones similares a las hechas al adaptar el test de Wilcoxon a un solo dominio.

También se pueden generar nuevos conjuntos de datos a partir del original contécnicas de remestreo.

Hay que tener cuidado porque un clasificador puede predominar sobre los otros

Se pueden aplicar los test Post Hoc directamente.

8. Esquema resumen de los tests estadísticos