UNIVERSIDAD D CÓRDOBA

Métricas de rendimiento en regresión y clasificación

Análisis automático de datos para ciencias biomédicas (Transversal Másteres Universitarios)

Juan Carlos Fernández Caballero
Departamento de Informática y Análisis Numérico
Universidad de Córdoba
curso 2021-2022

www.uco.es/ayrna

Índice

Índice

¿Para qué sirven las métricas de evaluación?

- Miden el **rendimiento y calidad** de un modelo, su **error cometido**.
- Existen multitud de métricas de evaluación.
  - **Clasificación:** Intentan minimizar el número de patrones mal clasificados (aumentar el número de correctos).
  - **Regresión:** Intentan minimizar la suma de los errores cometidos entre la predicción y el valor real.
- Un **buen valor en una métrica** no significa necesariamente **buenos valores en las demás**.
- También permiten **comparar diferentes hipótesis o modelos**.
- Se pueden calcular sobre el conjunto de datos de **entrenamiento** y sobre los de **test**, pero ya sabemos que el rendimiento real nos lo dan los **patrones de test, que son con los que no ha aprendido el modelo.**
  - En **Weka** aparecerán las mediciones **sobre el conjunto de test.**

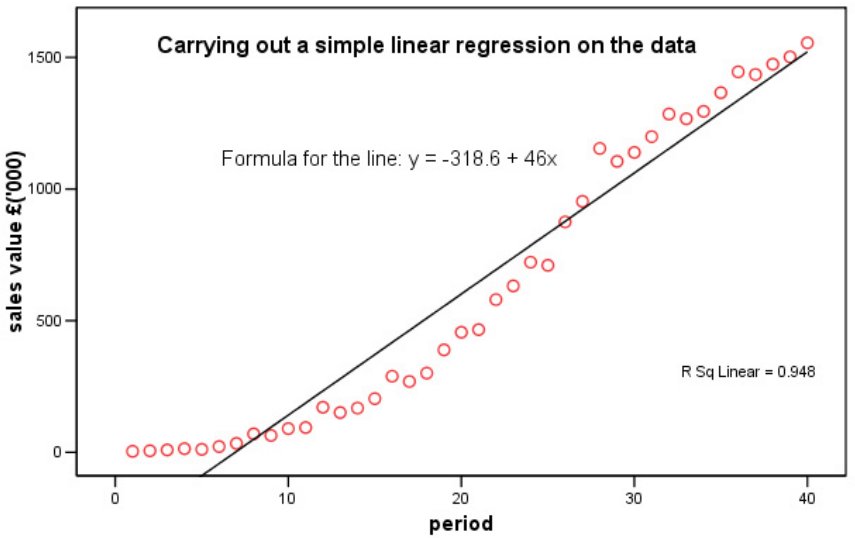
Índice

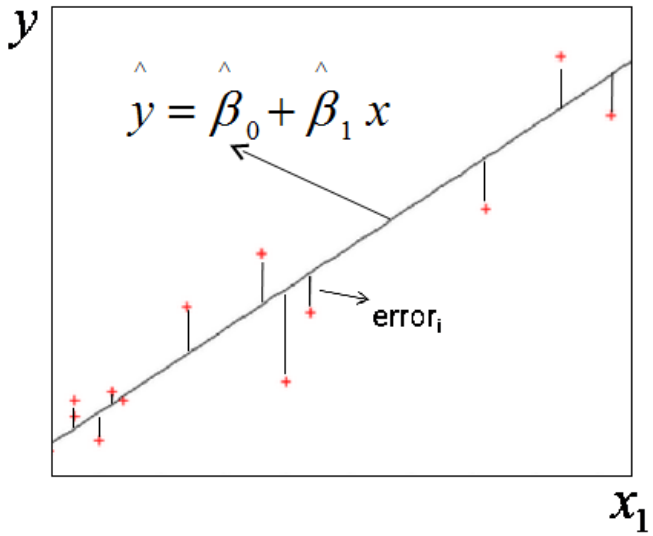
Objetivo de la regresión

Objetivo: Minimizar la suma de errores cuadráticos

- Sea r_i (también denotado como y_i): **Valor real** para el patrón i de un total de m **patrones**.
- Sea p_i (también denotado como $\hat{y}_i$): **Valor predicho** para el patrón i de un total de m patrones. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Minimizar la **suma de errores cuadráticos** (*squared sum of errors* - ***SSE***):

$$SSE = \sum_{i=1}^{m} (r_i - p_i)^2 = \sum_{i=1}^{m} \left(r_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$





Coeficiente de correlación R^2

- p_i: Valor predicho para el patrón i de un total de m patrones.
- r_i: Valor real para el patrón i de un total de m patrones.
- $\bar{r}$: Media de los valores reales.

A maximizar sobre el conjunto de test: Valor entre $[0, 1]$

- *Correlation coefficient* (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(p_i - r_i)^2}{\sum_{i=1}^{m}(r_i - \bar{r})^2}$$

  - Cuanto más cercano a 1 mejor.
  - Mide qué ratio de la varianza de la variable de salida es explicada por el modelo (**cómo de bien se ajustan los datos a la recta de regresión**).
  - Se basa en el coeficiente de correlación de Pearson.
  - Puede darse el caso de que sea **negativo en regresores muy malos o triviales** (se comentan más adelante).

Recordatorio: Correlación

El coeficiente de correlación de *Pearson* devuelve un valor entre **-1 y 1**:

En **Weka**, la **matriz de correlaciones** entre las variables independientes se puede obtener mediante
Select attributes $\longrightarrow$ **(PrincipalComponents, Ranker).**

- **-1**: Correlación negativa completa. Cuando una variable aumenta la otra decrementa en proporción constante.
- **1**: Correlación positiva completa. Cuando una variable aumenta la otra lo hace en proporción constante.
- **0**: No hay correlación.
- Valores $<$**(-0.6)** o valores $>$**(0.6)**: Indica correlación notable.
- Recuerde tambien que en Weka, para problemas de regresión, ***Select attributes*** $\longrightarrow$**(CorrelationAttributeEval,Ranker)** también calcula correlaciones, pero en este caso, de cada **variable independiente respecto a la salida** o variable dependiente de salida, de forma que se puedan visualizar posteriormente las variables de entrada en función de un ***ranking***.

Mean Squared Error (MSE)

A minimizar sobre el conjunto de test: Valor entre $[0, \infty]$

- *Mean Squared Error* (**MSE**) (0-∞): Mide el promedio de los errores cometidos en un conjunto de predicciones. Los **errores grandes aumentan mucho el valor de esta métrica**. Cuanto más cercano a 0 mejor.

$$MSE = \frac{\sum_{i=1}^{m}(p_i - r_i)^2}{m}$$

Esta medida **por si sola es compleja de interpretar**, sino se **compara** con el valor de **MSE que haya obtenido otro modelo** con alguna otra técnica de regresión.

Cabe la posibilidad de comparar con un **regresor trivial** o ingenuo, que sería aquel que predice siempre la **media de los valores reales de salida** (los de entrenamiento).

En clasificación, un **clasificador ingenuo** sería aquel que **clasifica** todos los patrones como pertenecientes a la **clase mayoritaria** del conjunto de datos total.

En Weka, se puede usar **ZeroR** como **clasificador y regresor ingenuo**. Pestaña *Classify*, *classifiers* $\longrightarrow$ *rules* $\longrightarrow$ *ZeroR*

Root Mean Squared Error (RMSE)

A minimizar sobre el conjunto de test: Valor entre $[0, \infty]$

- *Root Mean Squared Error* (**RMSE**) (0-∞): Mide el promedio cuadrático de los errores cometidos en un conjunto de predicciones. **Sensible a errores grandes, como el MSE**. Cuanto más cercano a 0 mejor.

  Más interpretable que MSE. Al hacer la raíz cuadrada el error queda a la **misma escala que los errores de predición**, es decir, las unidades que expresa el RMSE son las mismas que las unidades originales del valor objetivo que se predice.

  Ej: si la variable objetivo tiene las unidades en "dólares", entonces el RMSE también está expresado en unidades de "dólares" y no "dólares al cuadrado" como el *MSE*.

  Ej: un $RMSE = 2{,}3$ en un problema en el que la variable está expresada en metros, te dice que el modelo se equivoca en media 2.3 metros con respecto a los valores reales.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(p_i - r_i)^2}{m}}$$

Mean Absolute Error (MAE)

A minimizar sobre el conjunto de test: Valor entre $[0, \infty]$

- *Mean Absolute Error* (**MAE**) (0-∞): Mide el promedio de los errores cometidos en un conjunto de predicciones. Cuanto más cercano a 0 mejor.
- **No es tan sensible a los valores atípicos** o a grande errores como los puede ser MSE y RMSE.
- Al igual que el RMSE, las unidades que expresa el MAE son las mismas que las unidades originales del valor objetivo que se predice.

$$MAE = \frac{\sum_{i=1}^{m} |p_i - r_i|}{m}$$

Relative Absolute Error (RAE) y *Root Relative Squared Error* (RRSE)

A minimizar sobre el conjunto de test: Valor entre $[0, \infty]$

- *Relative Absolute Error* (**RAE**) (0-∞) en %: Error con respecto al error que se cometería al predecir la media de los valores reales de salida (regresor trivial).

$$RAE = \frac{\sum_{i=1}^{m} |p_i - r_i|}{\sum_{i=1}^{m} |r_i - \bar{r}_i|}$$

  Más dificil de interpretar. Weka lo multiplica por 100.

  Ej: Si se obtiene un 1 (100 %) quiere decir que se tiene el mismo error que un modelo que predice la media (que ya sería un modelo malo).

  Ej: Si se obtiene un 0.5 (50 %) sería la mitad del error que produce un modelo trivial que predeciría la media. Mientras más cercano a 0 mejor.

  Ej: Más de un 1 ($>$ 100 %) significa que el predictor es peor que un predictor básico o trivial.

- *Root Relative Squared Error* (**RRSE**) (0-∞) en %: Igual que el RAE, pero al ser cuadrático **exagera los errores más grandes** mientras que dan **menos importancia a los errores pequeños**.

$$RRSE = \sqrt{\frac{\sum_{i=1}^{m} (p_i - r_i)^2}{\sum_{i=1}^{m} (r_i - \bar{r}_i)^2}}$$

Índice

Métricas de evaluación en clasificación

La matriz de confusión

En **clasificación** la mayoría de las métricas surgen de lo que se llama matriz de confusión.

¿Qué es una matriz de confusión?

- Es una tabla de errores que permite la visualización del desempeño de un modelo supervisado.
- Se obtiene a partir del **conjunto de generalización o *testing*** aplicado al modelo supervisado construido sobre el **conjunto de *training***.
- Cada **fila** representa a las instancias en la **clase real**.
- Cada **columna** representa la **clase inferida o predicha** por el modelo, que puede ser igual o no a la real.

La matriz de confusión en problemas multiclase

- Problema de clasificación con J **clases** ($J > 2$) y n **patrones** de entrenamiento o test (el test nos da el rendimiento más realista).
- El rendimiento de un clasificador g se puede obtener a partir de su matriz confusión definida en la forma:

Clase predicha

$$\begin{array}{c|cccccc} & C_1 & \dots & C_j & \dots & C_J & \\ C_1 & n_{11} & \dots & n_{1j} & \dots & n_{1J} & n_{1\cdot} \\ \dots & \dots & & \dots & & \dots & \dots \\ C_i & n_{i1} & \dots & n_{ij} & \dots & n_{iJ} & n_{i\cdot} = \sum_{j=1}^{J} n_{ij} \\ \dots & \dots & & \dots & & \dots & \dots \\ C_J & n_{J1} & \dots & n_{Jj} & \dots & n_{JJ} & n_{J\cdot} \\ & n_{\cdot 1} & \dots & n_{\cdot j} = \sum_{i=1}^{J} n_{ij} & \dots & n_{\cdot J} & n = \sum_{i=1}^{J}\sum_{j=1}^{J} n_{ij} \end{array}$$

Clase real

$$M(g) = \left\{ n_{ij} \,/\, \sum_{i=1}^{J}\sum_{j=1}^{J} n_{ij} = n \right\}$$

donde n_{ij} representa el número de veces un **patrón de la clase** i se ha predicho como **perteneciente a la clase** j.

La matriz de confusión en problemas binarios (bi-clase)

- 2 clases, **positiva** y **negativa**.
- Lo usual es utilizar como **clase positiva** a la **clase minoritaria** en cuanto a número de patrones, que es lo mas usual en problemas de medicina, pero no tiene porque ser así.

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

Accuracy o Correct Classification Rate (CCR)

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

Accuracy o Correct Classification Rate (CCR). Valor entre $[0, 1]$

- Muestra el porcentaje de patrones correctamente clasificado (a maximizar).
- Se puede expresar también en tanto por ciento, valor entre $[0, 100]$ (ej: 95 %).
- Sirve para problemas **bi-clase** o **multi-clase**.

$$CCR = \frac{TP + TN}{TP + TN + FP + FN}$$

- En un problema **multiclase**, la precisión global se calcula como la suma de los **elementos de la diagonal** de la matriz de confusión, dividido por la suma de todos los elementos de la matriz.

$$CCR = \frac{1}{N} \sum_{j=1}^{J} n_{jj},$$

donde N es el número de patrones en generalización, J es el número de clases, y n_{jj} (elemento de la diagonal) es el número de patrones de la clase j-th que están correctamente clasificados.

Métricas para Clasificación Bi-clase

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

TP Rate, TN Rate, Precision, FP Rate y *F-Measure*. $[0, 1]$

- ***TP Rate*, *Recall*, *Sensitivity*, *Precisión positiva* (A Maximizar)**:
  Porcentaje de patrones positivos predichos como positivos.

$$TPRate = Recall = \frac{TP}{TP+FN}$$

- ***TN Rate*, *Specificity*, *Precisión negativa* (A Maximizar)**:
  Porcentaje de patrones negativos predichos como negativos.

$$Specificity = \frac{TN}{FP+TN}$$

- ***Precision* (A Maximizar)**:
  Porcentaje de patrones positivos predichos como positivos, frente al total de patrones predichos como positivos.

$$Precision = \frac{TP}{TP+FP}$$

Métricas para Clasificación Bi-clase

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

TP Rate, TN Rate, Precision, FP Rate, F-Measure. $[0, 1]$

- ***FP Rate* (A Minimizar):**
  Porcentaje de patrones negativos predichos como positivos.
  Equivale a (1-*Specificity*).

$$FPRate = \frac{FP}{FP+TN}$$

- ***FN Rate* (A Minimizar):**
  Porcentaje de patrones positivos predichos como negativos.

$$FNRate = \frac{FN}{FN+TP}$$

- ***F-Measure* o *F-Score* (A Maximizar)**:
  Combina las métricas *Recall* y *Precision*.

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$F - Measure = \frac{2TP}{2TP + FP + FN}$$

Métricas para Clasificación Bi-clase

Resumen gráfico en la matriz de confusión Bi-clase:

		Predicción C_P	Predicción C_N
Real	C_P	TP	FN
Real	C_N	FP	TN

Accuracy

		Predicción C_P	Predicción C_N
Real	C_P	TP	FN
Real	C_N	FP	TN

TP Rate (Recall)

		Predicción C_P	Predicción C_N
Real	C_P	TP	FN
Real	C_N	FP	TN

FP Rate y Specificity

		Predicción C_P	Predicción C_N
Real	C_P	TP	FN
Real	C_N	FP	TN

Precision

		Predicción C_P	Predicción C_N
Real	C_P	TP	FN
Real	C_N	FP	TN

F-measure

Métricas para Clasificación Multiclase

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

¿Cómo obtener los valores de las métricas anteriores para cada clase en clasificación Multiclase?

- Se pueden obtener todas las métricas anteriores pero siempre en función de una de las clases, que es la que se considera la **positiva, contra el resto.**

Una clase contra el resto en la matriz de confusión

	Clase predicha		
	a	b	c
Clase real	49	1	0
	0	47	3
	0	2	48

	Clase predicha		
	a	b	c
Clase real	49	1	0
	0	47	3
	0	2	48

clase a

	Clase predicha		
	a	b	c
Clase real	49	1	0
	0	47	3
	0	2	48

clase b

	Clase predicha		
	a	b	c
Clase real	49	1	0
	0	47	3
	0	2	48

clase c

TP FN FP TN

Estadístico KAPPA (binario y multiclase)

Kappa. Valor entre $[-1, 1]$ a maximizar

- Compara la **concordancia** observada en un conjunto de datos por un modelo, respecto a la que **podría ocurrir por mero azar**.
- Se calcula de igual manera para problemas binarios y multiclase (ver siguiente diapositiva).
- Puede tomar valores en el rango $[-1, 1]$. A maximizar.
  - -1 = Discordancia total, peor que una clasificación al azar.
  - 1 = Concordancia perfecta, sin azar.
  - > 0 = Mayor concordancia que la que se esperaría por el puro azar.
  - 0 = No existe relación, la concordancia observada coincide con la que ocurriría por puro azar.

Estadístico KAPPA (binario y multiclase)

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

$$Kappa = \frac{p_o - p_e}{1 - p_e}$$

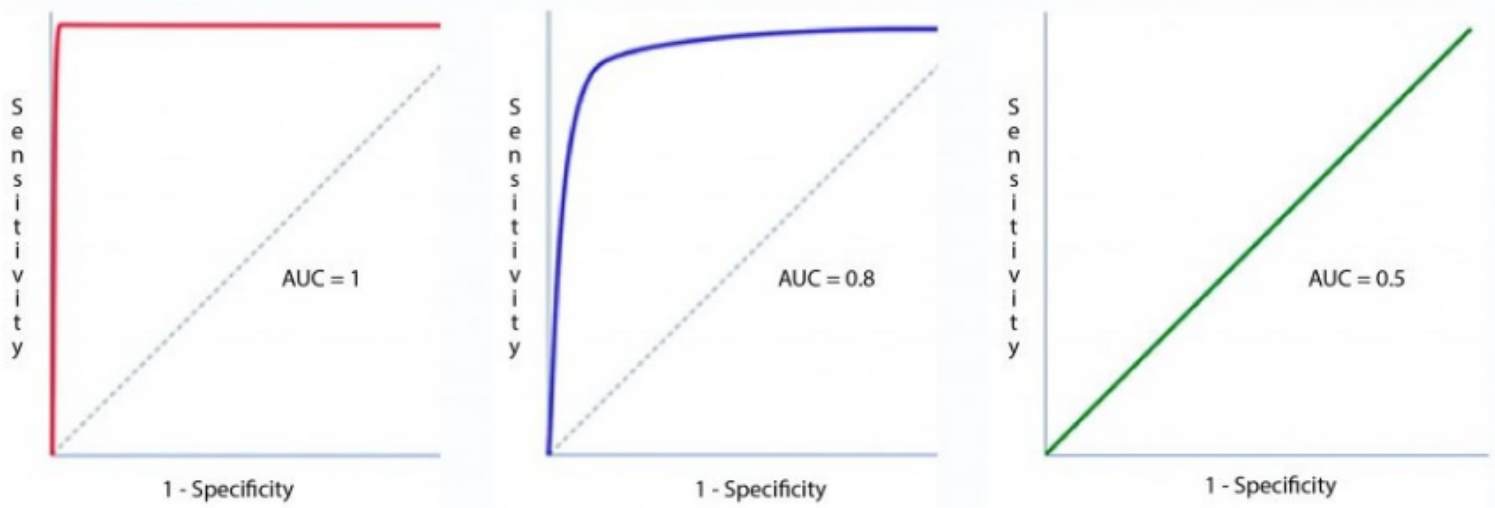
$$p_o = CCR = \frac{1}{n}\sum_{j=1}^{J} n_{jj} \qquad p_e = \frac{1}{n^2}\sum_{j=1}^{J} n_{j\bullet} n_{\bullet j}$$

- n_{jj} es un elemento de la matriz de confusión.
- J es el número de clases.
- n es el número de patrones en *testing*.
- $n_{j\bullet}$ es la suma de todos los elementos de la fila j
- $n_{\bullet j}$ es la suma de todos los elementos de la columna j

Area Under the ROC Curve (*AUC*)

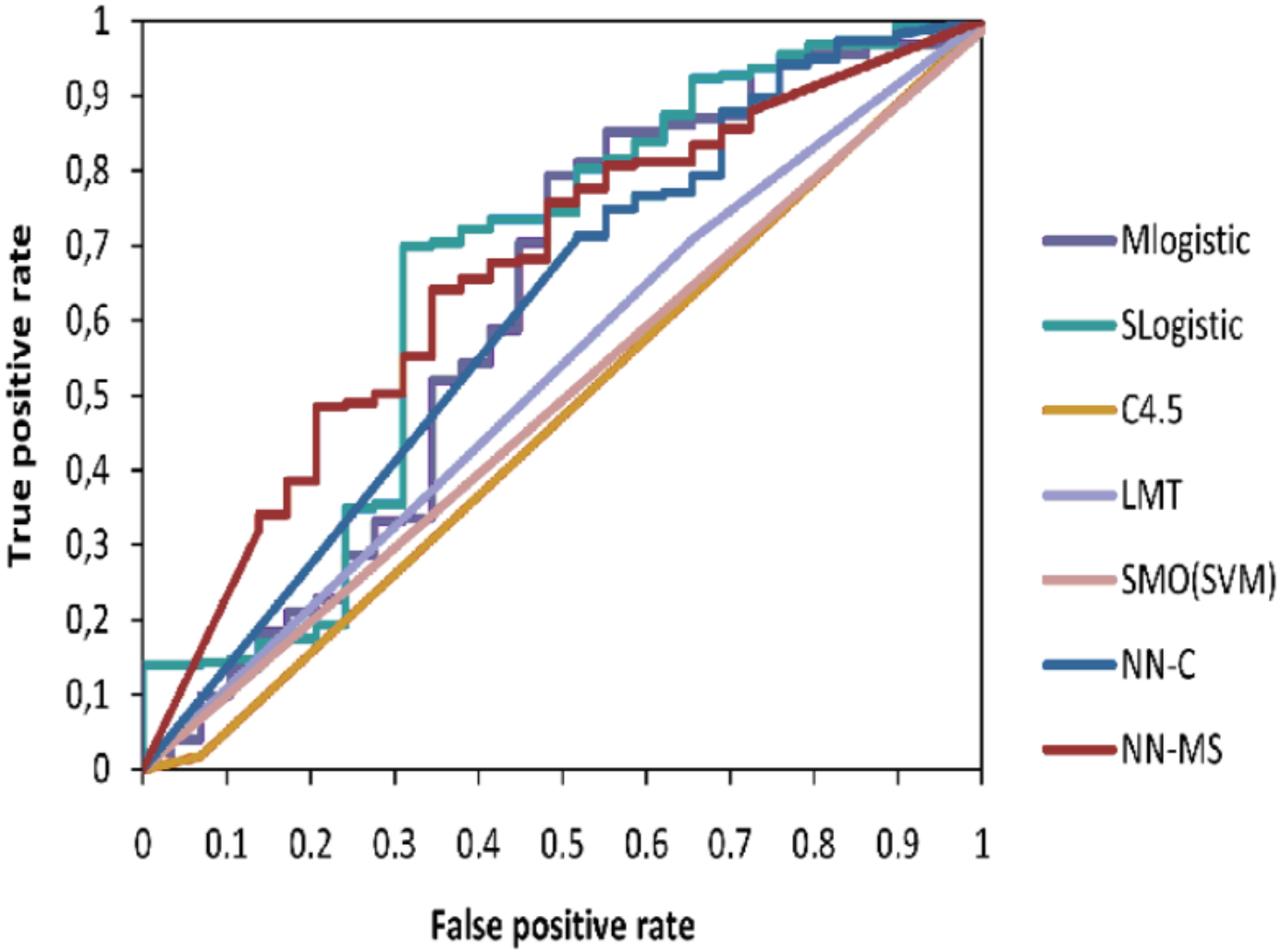
AUC o Curva ROC. Valor entre $[0, 1]$ a maximizar

- Más usada en **clasificación binaria**.
- Espacio bidimiensional ROC: *FP Rate* eje X ; *TP Rate* eje Y .
- Se calcula obteniendo el **área que queda por debajo de una curva que se calcula a partir de los patrones** y su probabilidad de pertenencia a la clase positiva (consultar la web si está interesado en su cálculo matemático).
- La **linea representa un clasificador binario trivial** (como lanzar una moneda con cara y cruz al aire). Deberíamos **conseguir valores mayores a 0.5**.



Area Under the ROC Curve (*AUC*)

Diferentes clasificadores y su curva ROC: Seleccionar el clasificador con mayor área bajo la curva (AUC).



Recordatorio: Problema del Accuracy o Correct Classification Rate (CCR)

Problema de predicción con **2 clases** (tumor, no tumor) y **1000 patrones**.

$$\text{Clase real} \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix} = \begin{pmatrix} 0 & 10 \\ 0 & 990 \end{pmatrix} \quad \text{(Clase predicha)}$$

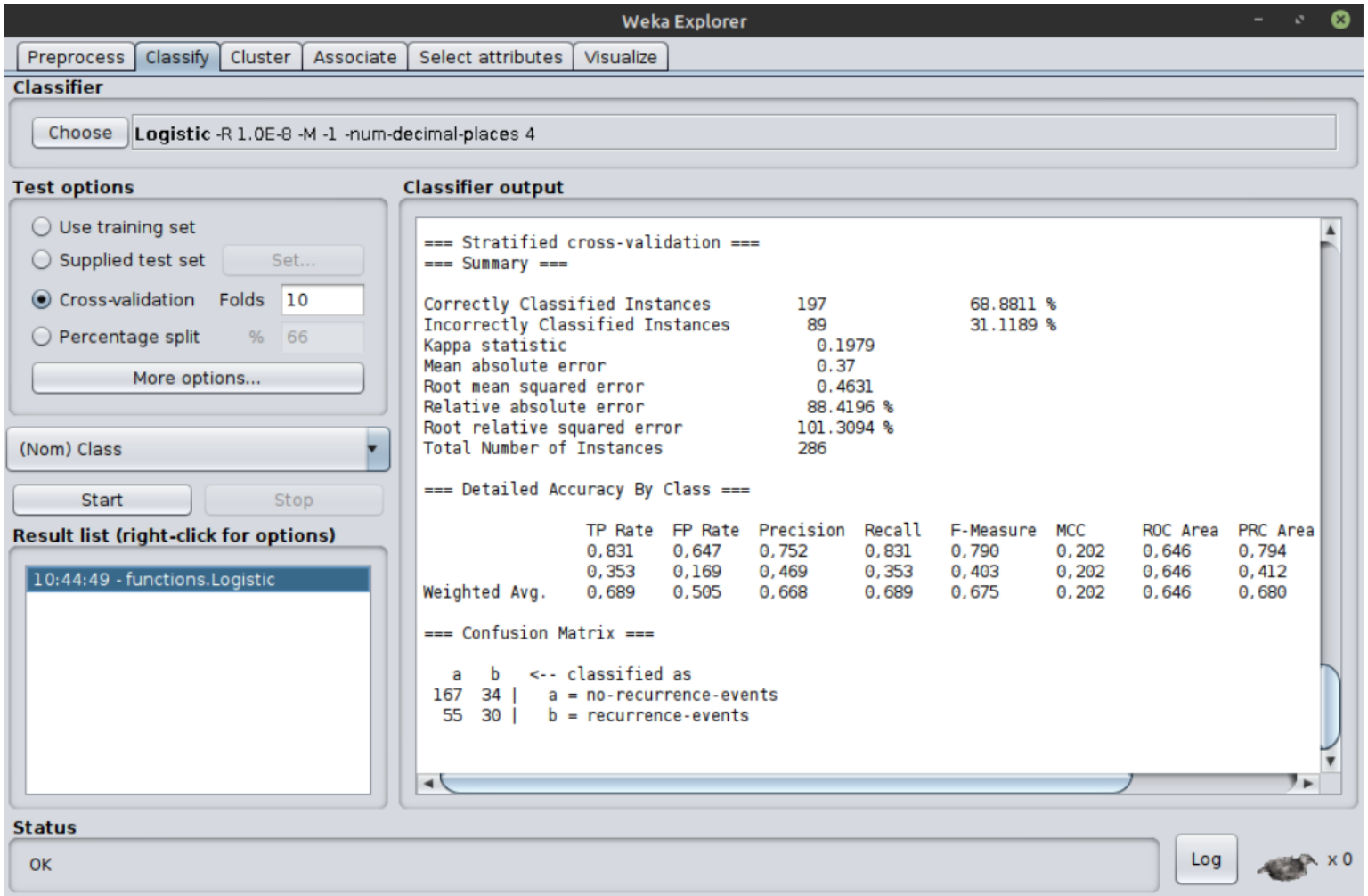
- **10 ejemplos** de la **clase 1 o positiva** (pacientes con tumor).
- **990 ejemplos** de la **clase 2 o negativa** (pacientes sin tumor).
- Si el modelo siempre dice que los ejemplos son de la clase 2, su precisión global es:

$$CCR = \frac{990}{1000} = 99{,}9\,\%$$

- Sensibilidad $= 0$; Especificidad $= \frac{990}{990} = 1$;
- Valor engañoso, ya que **nunca detecta patrones de la clase 1**.
- Mirar el CCR de cada una de las clases. En Weka es la columna *TP Rate*.

Recordatorio: Problema del Accuracy o Correct Classification Rate (CCR)

Problema en Weka con un **83,1 % de aciertos para la clase positiva** y un **35,3 % para la clase negativa**. La clase negativa no la predice tan bien como la positiva.



Índice

Métricas a usar en Weka

- Observará que en Weka, **a pesar de que un problema no sea binario**, es decir, sea de tres o más clases, **aparecen métricas** que solo recomiendan utilizar en clasificadores binarios, como por ejemplo el FP Rate, Precision, AUC, ...
  - Esto es porque para calcularla se enfrenta **una clase frente al resto**, de forma que el resto de clases se toma como la clase negativa.
- **Métricas** que usaremos en **Weka** para hacer una análisis de resultados básico en **Regresión**:
  - Correlation coefficient (Correlation coefficient, R^2).
  - Mean absolute error (MAE).
  - Root mean squared error (RMSE).
- **Métricas** que usaremos en **Weka** para hacer una análisis de resultados básico en **Clasificación**:
  - Correctly Classified Instances (CCR).
  - Kappa Statistic (Kappa).
  - TP Rate (El CCR por clase).

¿Preguntas?

¡Gracias!