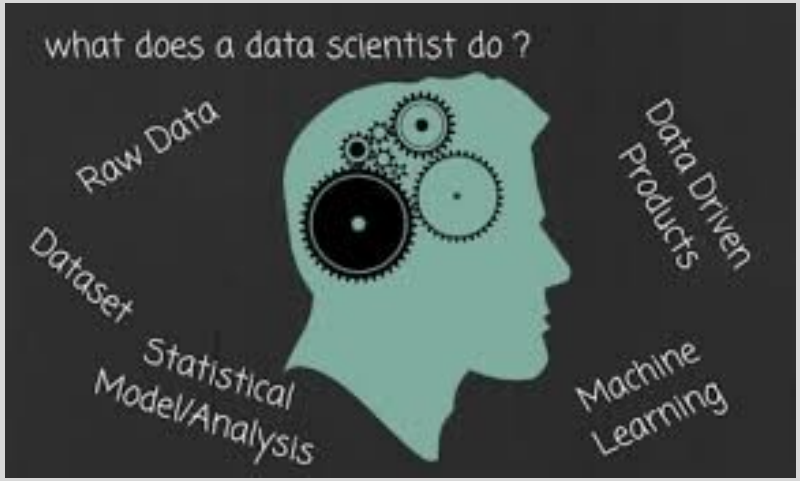


DigitalHouse > Coding School

DATA SCIENCE

UNIDAD 2
MÓDULO 4

Evaluación del ajuste de un modelo

Octubre 2017

Evaluación del ajuste de un modelo

OBJETIVOS

1. Comprender los fundamentos de los métodos de evaluación de modelos de clasificación

2. Entender las métricas de: Precision, Recall, Accuracy, F1 y curvas ROC

3. Implementar esas métricas de evaluación con SciKit-learn

Introducción: Conceptos Clave

- Los modelos de machine learning usados para clasificación se evalúan de manera diferente a las regresiones.
- En una regresión buscamos predecir una variable continua, en un clasificador en cambio el objetivo es predecir la pertenencia o la probabilidad de pertenencia a una clase.
- Existen varias maneras de evaluar la performance de un clasificador. Es importante elegir la adecuada para el problema en mano.

- Los outcomes en una clasificación en función de la tasa de acierto se pueden dividir en en cuatro clases. Pongamos como ejemplo un clasificador que determina si un individuo pertenece o no a la clase “enfermo”.
- Definiciones:
  - **Falsos Positivos (FP):** es una clase negativa que fue clasificada como positivo. Ejemplo: al individuo se lo clasificó como enfermo (y estaba sano)
  - **Falsos Negativos (FN):** el individuo fue clasificado como sano, y estaba enfermo
  - **Verdaderos Positivos (TP):** el clasificador no tuvo errores
  - **Verdaderos Negativos (TN)**: sin errores al detectar un individuo sano
- Aclaración: la noción de “positivos” o “negativos” es arbitraria y podría ser reemplazada por las de “presencia”-”ausencia”.

Explorando métricas de evaluación

Matriz de Confusión:

- Es una tabla de doble entrada donde se describen los resultados observados vs resultados esperados luego de haber aplicado del modelo

	Predice que saludable (y=0)	Predice que tiene cáncer (y=1)
Está saludable (y=0)	46 **(TN)**	85 **(FP)**
Tiene cáncer (y=1)	168 **(FN)**	31 **(TP)**

- Nos permite discernir entre los casos bien clasificados y los que fueron erróneamente clasificados por el modelo.
- Es importante porque desde acá parten las categorías de TP, TN, FP y FN.

Accuracy

- Cantidad de clases correctamente predichas por el modelo
- Notar que esta métrica puede dar positivo aun si no tengo positivos tanto verdaderos como falsos.

```
Accuracy = (True Positives + True Negatives) / Total
```

```
from sklearn.metrics import accuracy_score

acc = accuracy_score(Y_test, Y_pred)

# This is equivalent to:
acc = np.sum(Y_test == Y_pred)/len(Y_test)
```

Recall (Sensitividad o True Positive Rate):

- Nos da visibilidad de la perfomance del clasificador en cuanto aquellos casos que son correctamente clasificados. Es decir, del total de positivos en el dataset, en cuántas muestras el modelo detectó bien la clase.
- Si su valor es bajo, es porque hay *presencia de falsos negativos.* Por eso ésta medida es sensible a los FN.
- Penaliza el no detectar los casos positivos.
    - Una buena aplicación sería detectar una enfermedades. En ese caso es más “costoso” errar por los casos no detectados que por los detectados erróneamente.

```
Recall = True Positives / (True Positives + False Negatives)
```

Precisión:

- Indica cuántas de las muestras clasificadas como enfermas son realmente enfermas. Contempla la probabilidad de que la clase sea realmente positiva (1=enfermo)
- Es una métrica que está sujeta a cuán frecuente es la clase positiva en nuestro dataset. En otras palabras, nos conviene usarla cuando nos interesa más encontrar una clase positiva. Esto es útil cuando se da el problema de la "aguja en el pajar" es decir cuando los casos positivos son muchos menos que los negativos.
- Penaliza el no detectar casos positivos.
    - Una buena aplicación sería decidir dónde poner una sucursal de un negocio. En este caso es muy “costoso” errar por los casos detectados erróneamente.

```
Precision = True Positives / (True Positives + False Positives)
```

F1: Es la media armónica de las anteriores métricas

```
F1-Score = 2 * (Precision * Recall) / (Precision + Recall)
```

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

$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

$$F_1 = 2 \cdot \frac{PRE \cdot REC}{PRE + REC}$$

Fβ:

- Es la media armónica de las anteriores métricas. F1 es un caso particular donde β=1.
- Mediante el parámetro β se puede regular la importancia relativa de cada término.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

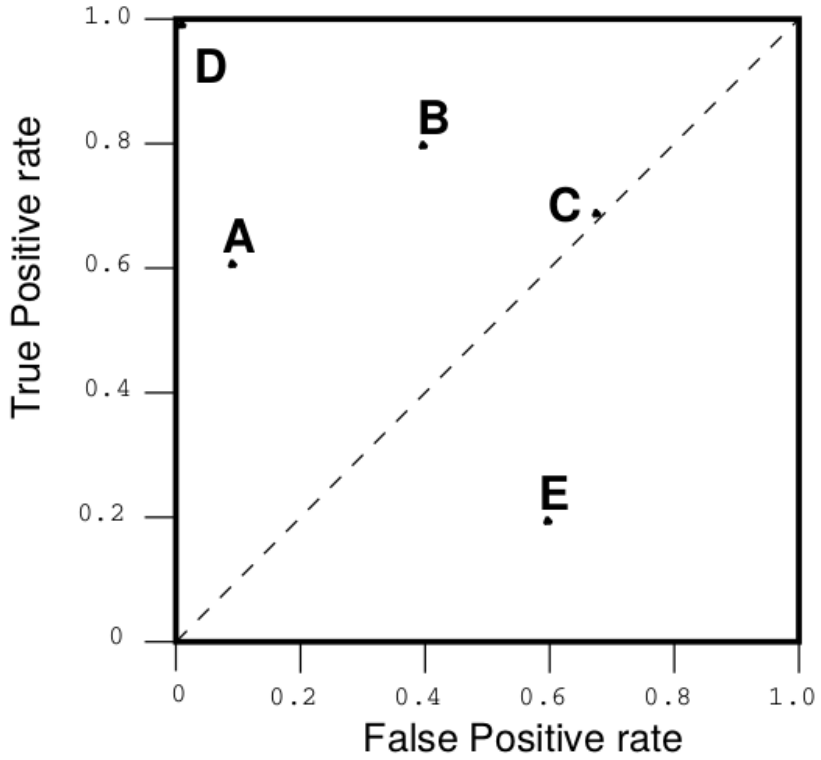
Práctica Guiada 1: explorando métricas de evaluación

Curva ROC y medida AUC...
¿de qué se tratan?

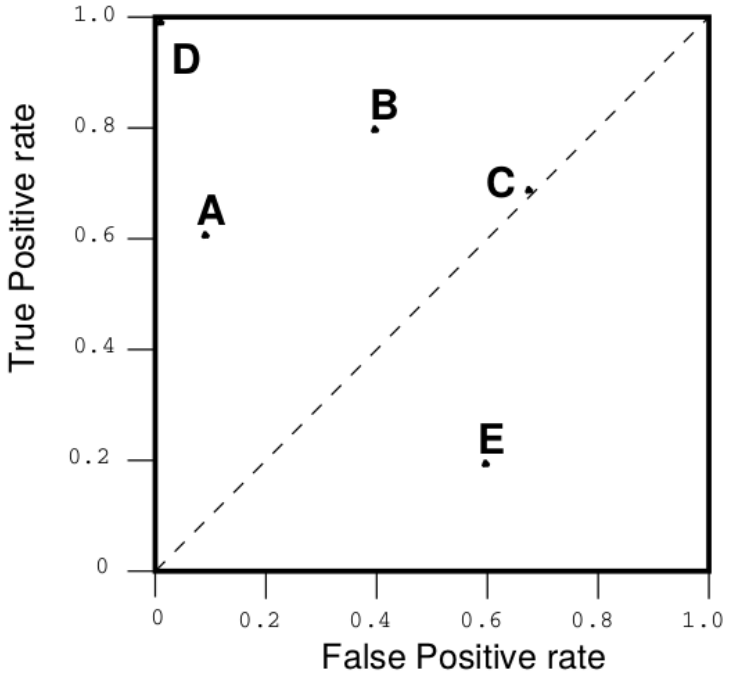
- Es una forma muy popular de representar los dos tipos de errores en un modelo de clasificación (binaria).
- Para lograr un buen entendimiento de curva ROC , es necesario tener en claro dos conceptos:
  - Sensibilidad o True Positive Ratio = **TP / (TP + FN)**
  - False Positive Ratio = **FP / (FP + TN)**

- Mundo ideal: mi modelo debería tener una Sensibilidad (TRP) de 100% y una FPR de 0%.
- El True Positive Ratio se ubica en el eje Y y el False Positive Ratio, en el eje X.
- Cada modelo representa un punto en el espacio ROC.

- Cada punto es un modelo (básicamente, la TPR y la FPR)
- ¿Cuál es el mejor modelo? ¿Por qué?
- ¿Qué puede decirse del punto **D**?
- ¿Y Del punto **E**?



- Un punto en ROC es mejor que otro si está en el “noroeste del gráfico” (cerca de D, TPR es alta y FPR es baja)
- Clasificadores cerca del eje X => **“conservadores”**: clasificaciones positivas solamente con fuerte evidencia => pocos FP pero también pocos TP
- Clasificadores derecha y arriba **“liberales”**: clasificaciones positivas con poca evidencia => muchos TP pero también muchos FP
- La diagonal => “random guess”



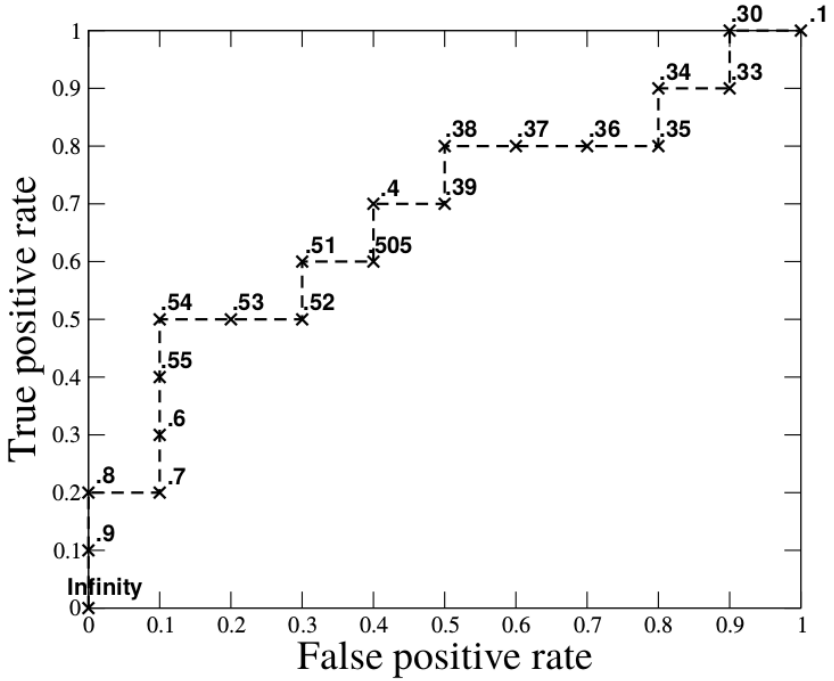
- Existen modelos (-NB, regresión logística, etc.) que predicen de forma natural scores o probabilidades de pertenencia, en lugar de etiquetas de clase.

- Tales probabilidades pueden ser utilizados para establecer un umbral (**T**) y generar las predicciones de clases (recordar regresión logística)
  - Si el caso está por encima del umbral **T**, el caso es clasificado como positivo
  - Si está por debajo, es clasificado como negativo.

- A su vez, se puede ir variando el **T** y evaluando cómo performa el modelo.

- En la slide siguiente vemos un ejemplo.

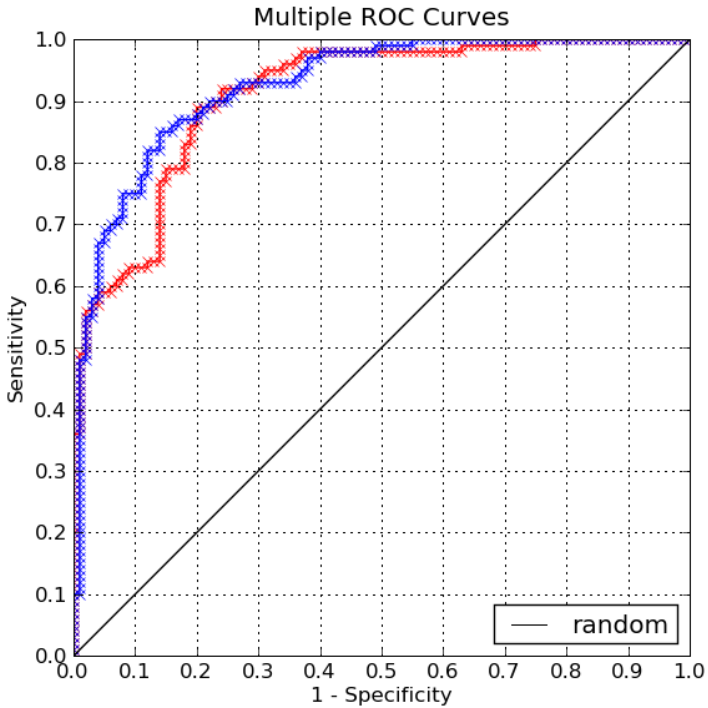
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



- A medida que modificamos **T**, el modelo funciona diferente:
  - **T** = 0.90 => FPR: 0.00 y TPR: 0.10
  - **T** = 0.80 => FPR: 0.00 y TPR: 0.20
  - **T** = 0.70 => FPR: 0.10 y TPR: 0.20
  - **T** = 0.60 => FPR: 0.10 y TPR: 0.30
  - **T** = 0.50 => FPR: 0.40 y TPR: 0.60
  - **T** = 0.40 => FPR: 0.40 y TPR: 0.70
  - **T** = 0.30=> FPR: 0.90 y TPR: 1.00

- Teniendo en cuenta esto…
  - ¿Cuál de los siguientes modelos es mejor? (Cada curva representa un modelo)

- Una buena medida es el área debajo de la curva ROC
  - Cuanto mayor sea el área… mejor será el modelo. ¿Por qué?



Práctica Guiada 2: curvas ROC

- Es posible generar métricas para evaluar modelos de clasificación en función de la matriz de confusión
- Esas métricas permiten discernir entre casos bien y mal clasificados por el modelo
- Es posible tener modelos que performen bien en el agregado (Accuracy) pero no tan bien en algunas de las clases (Recall, Precision)
- Las curvas ROC son una buena herramienta para “visualizar” la performance general del modelo