

ACÀMICA

¡Bienvenidos/as a Data Science!



Agenda

¿Cómo anduvieron?

Repaso

Hands-On

Explicación: Datasets Desbalanceados

Break

Hands-On

Explicación: Teorema de Bayes

Cierre



¿Dónde estamos?



¿Cómo anduvieron?



Proyecto 2: Modelado



Repaso



Aprendizaje supervisado: **Clasificación**

Modelos

- **Árboles de decisión** (Hiperparámetros: profundidad, criterio de entrenamiento, etc.)
- **KNN** (Hiperparámetros: cantidad de vecinos, distancia, etc.)

Métricas de evaluación

- Exactitud
- Precisión/Exhaustividad
- F-Score
- Matriz de Confusión¹

¹Bueno, técnicamente no es una métrica

Hands-on training



DS_Encuentro_21_DDDesb.ipynb

Parte 1 y 2



Datasets Desbalanceados



Un **dataset balanceado** es aquel que tiene - aproximadamente - la misma proporción de instancias de cada clase.
Por ejemplo, en el caso, binario, alrededor de 50:50 (1:1) de cada clase.

Un **dataset desbalanceado** - en el caso binario - es aquel que tiene muchas instancias de una clase y muy pocas de la otra, dificultando el entrenamiento.
Por ejemplo, 80:20, 90:10, 99:1, y peor.

Un poco de desbalance de clases es esperable, y no afecta a nuestro análisis.

Pero en algunas áreas suelen haber datasets muy desbalanceados:

- Detección de fraudes
- Diagnóstico médico
- Deforestación

Cuando trabajemos con estos datasets, tenemos que tener cuidado con:

- Cómo entrenamos nuestros modelos.
- Qué métricas usamos para evaluarlo.

Cuando trabajemos con estos datasets, tenemos que tener cuidado con:

- Cómo entrenamos nuestros modelos.
- Qué métricas usamos para evaluarlo.

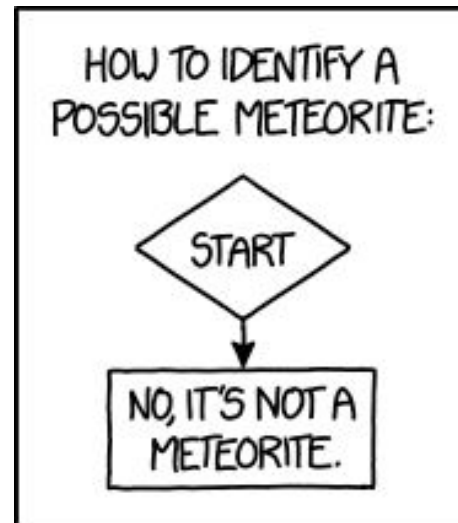
WARNING

Uno de los malos de la película:

La paradoja de la exactitud (suena mejor en inglés, Accuracy Paradox)

A medida que el desbalance de clases es mayor, la exactitud aumenta, por más que nuestro modelo sea muy malo.

¿Por qué? <https://xkcd.com/1723/> *



Algunas técnicas para trabajar correctamente con estos datasets:

1

¿Podemos **recolectar nuevos datos**?

2

Elegir la **métrica de performance** apropiada para nuestro problema (¡Olvidarse de Exactitud!). Matriz de Confusión, Precisión y Exhaustividad (recall) suelen ser las primeras opciones, pero hay más. ¿Un Falso Positivo tiene el mismo costo que un Falso Negativo?

3

Resamplear el dataset.

- a. Oversampling: generar nuevas instancias de la clase minoritaria, ya sea copiando instancias preexistentes, o generando instancias sintéticas (ver SMOTE).
 - b. Undersampling: eliminar instancias de la clase sobrerrepresentada.
-

4

Probar diferentes modelos (modelos de ensamble suelen ser buenos) y/o agregarle peso a la clase subrepresentada (fácil desde Scikit-Learn).

5

Las **opciones no se terminan acá**. Es un área de continuo desarrollo. Para tener en cuenta: One-Class classification.

Recursos

[8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset](#): El artículo en el que nos basamos. Claro y sencillo.

[Credit Fraud Detector](#): Un muy lindo kernel de Kaggle aplicado a un dataset MUY desbalanceado.

[Handling imbalanced datasets in machine learning](#): Explicación mucho más técnica, pero exhaustiva.



A close-up photograph of a white ceramic cup filled with a latte. The surface of the milk is decorated with intricate latte art, featuring a central heart shape surrounded by concentric, wavy lines. The cup is placed on a matching white saucer. In the background, a white napkin and a silver fork are visible, though they are out of focus. The overall lighting is soft and even, highlighting the textures of the coffee and the smooth surface of the cup.

¡BREAK!



Hands-on training



DS_Encuentro_21_DDesb.ipynb

Parte 3 y 4



Teorema de Bayes



Teorema de Bayes • Ejemplo típico

Un examen médico tiene una probabilidad de detección de 0.99 y una probabilidad de Falso Positivo de 0.01. El objetivo del Test es detectar una enfermedad de relativa baja prevalencia, que solo la tiene una persona en mil. Hacer el examen a 100000 personas y completar la matriz de confusión (es decir, calcular, en promedio, cuántos aciertos esperan obtener, cuántos Falsos Negativos, FP y Verdaderos Negativos).

Teorema de Bayes • Ejemplo típico

Un examen médico tiene una probabilidad de detección de 0.99 y una probabilidad de Falso Positivo de 0.01. El objetivo del Test es detectar una enfermedad de relativa baja prevalencia, que solo la tiene una persona en mil. Hacer el examen a 100000 personas y completar la matriz de confusión (es decir, calcular, en promedio, cuántos aciertos esperan obtener, cuántos Falsos Negativos, FP y Verdaderos Negativos).

Pistas:

En 100.000 mil personas, hay ____ casos positivos. De esos ____ casos positivos, si hago el test, espero que dé positivo en ____.

De los casos negativos, espero que ____ sean identificados correctamente como negativos y ____ como falsos positivos.

Teorema de Bayes • Ejemplo típico

Un examen médico tiene una probabilidad de detección de 0.99 y una probabilidad de Falso Positivo de 0.01. El objetivo del Test es detectar una enfermedad de relativa baja prevalencia, que solo la tiene una persona en mil. Hacer el examen a 100000 personas y completar la matriz de confusión (es decir, calcular, en promedio, cuántos aciertos esperan obtener, cuántos Falsos Negativos, FP y Verdaderos Negativos).

Pistas:

En 100.000 mil personas, hay ____ casos positivos. De esos ____ casos positivos, si hago el test, espero que dé positivo en ____ .
De los casos negativos, espero que ____ sean identificados correctamente como negativos y ____ como falsos positivos.

¿Cuál es la probabilidad de que **una persona tenga la enfermedad** si el examen dio positivo?



Predicha / Verdadera	Positivos	Negativos
Positivos	99	999
Negativos	1	98901

Teorema de Bayes • Ejemplo típico

Un examen médico tiene una probabilidad de detección de 0.99 y una probabilidad de Falso Positivo de 0.01. El objetivo del Test es detectar una enfermedad de relativa baja prevalencia, que solo la tiene una persona en mil. Hacer el examen a 100000 personas y completar la matriz de confusión (es decir, calcular, en promedio, cuántos aciertos esperan obtener, cuántos Falsos Negativos, FP y Verdaderos Negativos).

Pistas:

En 100.000 mil personas, hay ____ casos positivos. De esos ____ casos positivos, si hago el test, espero que dé positivo en ____ .
De los casos negativos, espero que ____ sean identificados correctamente como negativos y ____ como falsos positivos.

¿Cuál es la probabilidad de que **una persona tenga la enfermedad** si el examen dio positivo?

Predicha / Verdadera	Positivos	Negativos
Positivos	99	999
Negativos	1	98901

Teorema de Bayes • Ejemplo típico

¿Cuál es la probabilidad de que
una persona tenga la enfermedad
si el examen dio positivo?

Predicha / Verdadera	Positivos	Negativos
Positivos	99	999
Negativos	1	98901

¡De 1098 predichos, 99 eran verdaderos positivos!
Es decir, **~9% (o probabilidad = 0.0902)**.

Teorema de Bayes • Ejemplo típico

¿Cuál es la probabilidad de que
una persona tenga la enfermedad
si el examen dio positivo?

Predicha / Verdadera	Positivos	Negativos
Positivos	99	999
Negativos	1	98901

¡De 1098 predichos, 99 eran verdaderos positivos!
Es decir, **~9% (o probabilidad = 0.0902)**.

¿A qué métrica vista corresponde este resultado?



que afecta al 0.1% de la población y es una desagradable enfermedad de horribles consecuencias,

The Bayesian Trap

Teorema de Bayes

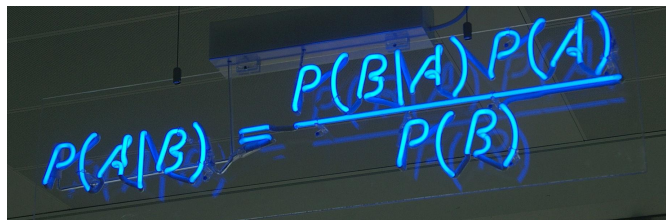
Dados dos eventos A y B:

- $P(A)$ es la probabilidad del evento A
- $P(B)$ es la probabilidad del evento B
- $P(A|B)$ es la probabilidad condicional del evento A dado que *ocurrió* B
- $P(B|A)$ es la probabilidad condicional del evento B dado que *ocurrió* A
- Si $P(A|B) = P(A)$ y $P(B|A) = P(B)$, los eventos son independientes.

¡No implica
causalidad!

En general, $P(A|B) \neq P(B|A)$.

Para obtener uno dado el otro,
necesitamos el Teorema de Bayes:


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Teorema de Bayes

Dados dos eventos A y B:

- $P(A)$ es la probabilidad del evento A
- $P(B)$ es la probabilidad del evento B

Teorema de Bayes

Dados dos eventos A y B:

- $P(A)$ es la probabilidad del evento A
- $P(B)$ es la probabilidad del evento B
- $P(A|B)$ es la probabilidad condicional del evento A dado que *ocurrió* B
- $P(B|A)$ es la probabilidad condicional del evento B dado que *ocurrió* A

Teorema de Bayes

Dados dos eventos A y B:

- $P(A)$ es la probabilidad del evento A
- $P(B)$ es la probabilidad del evento B
- $P(A|B)$ es la probabilidad condicional del evento A dado que *ocurrió* B
- $P(B|A)$ es la probabilidad condicional del evento B dado que *ocurrió* A
- Si $P(A|B) = P(A)$ y $P(B|A) = P(B)$, los eventos son independientes.

¡No implica
causalidad!

Teorema de Bayes

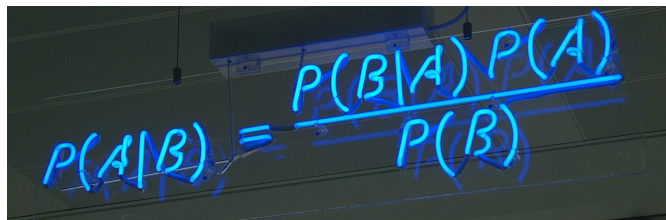
Dados dos eventos A y B:

- $P(A)$ es la probabilidad del evento A
- $P(B)$ es la probabilidad del evento B
- $P(A|B)$ es la probabilidad condicional del evento A dado que *ocurrió* B
- $P(B|A)$ es la probabilidad condicional del evento B dado que *ocurrió* A
- Si $P(A|B) = P(A)$ y $P(B|A) = P(B)$, los eventos son independientes.

¡No implica
causalidad!

En general, $P(A|B) \neq P(B|A)$.

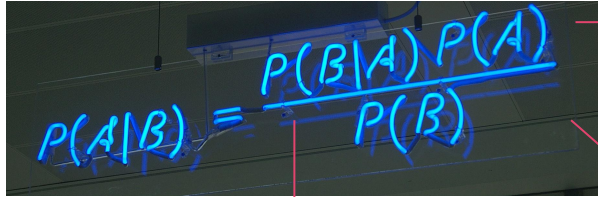
Para obtener uno dado el otro,
necesitamos el Teorema de Bayes:


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Teorema de Bayes

A: estar enfermo **E+**

B: dio positivo **T+**


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

P(A): prior o probabilidad a priori de A

P(B): probabilidad marginal.

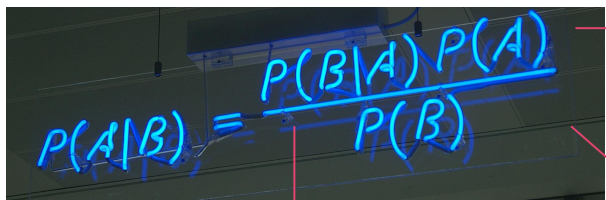
P(B|A): verosimilitud

P(A|B): posterior o probabilidad a posteriori

Volviendo al
ejemplo anterior...

Teorema de Bayes

A: estar enfermo **E+**
B: dio positivo **T+**


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

P(A): prior o probabilidad a priori de A

P(E+): El prior es la prevalencia de la enfermedad en la población = **1/100**

P(B): probabilidad marginal.

P(T+). La probabilidad de que el test dé positivo. Esto puede ocurrir si una persona está enferma pero también si no lo está. = **$P(T+|E+) * P(E+) + P(T+|E-) * P(E-) = 0.99 * 0.001 + 0.01 * 0.999 = 0.01098$**

P(B|A): verosimilitud

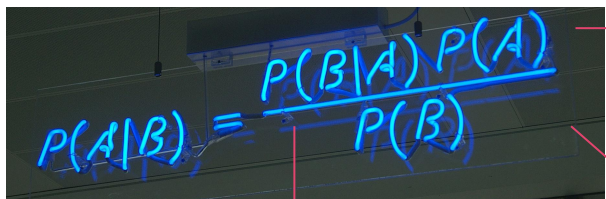
P(T+|E+): probabilidad de que el test de positivo dado que la persona está enferma. ¡Es la probabilidad de detección! = **0.99**

P(A|B): posterior o probabilidad a posteriori

P(E+|T+): probabilidad de estar enfermo dado que el test dio positivo.

Teorema de Bayes

A: estar enfermo **E+**
B: dio positivo **T+**


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A): prior o probabilidad a priori de A

P(E+): El prior es la prevalencia de la enfermedad en la población = **1/100**

P(B): probabilidad marginal.

P(T+). La probabilidad de que el test dé positivo. Esto puede ocurrir si una persona está enferma pero también si no lo está. = $P(T+|E+) \cdot P(E+) + P(T+|E-) \cdot P(E-) = 0.99 \cdot 0.001 + 0.01 \cdot 0.999 = 0.01098$

P(B|A): verosimilitud

P(T+|E+): probabilidad de que el test de positivo dado que la persona está enferma. ¡Es la probabilidad de detección! = **0.99**

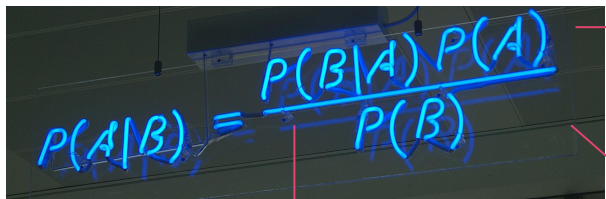
¿Quién es?

P(A|B): posterior o probabilidad a posteriori

P(E+|T+): probabilidad de estar enfermo dado que el test dio positivo.

Teorema de Bayes

A: estar enfermo **E+**
B: dio positivo **T+**


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

P(A): prior o probabilidad a priori de A

P(E+): El prior es la prevalencia de la enfermedad en la población = **1/100**

P(B): probabilidad marginal.

P(T+). La probabilidad de que el test dé positivo. Esto puede ocurrir si una persona está enferma pero también si no lo está. = $P(T+|E+) * P(E+) + P(T+|E-) * P(E-) = 0.99 * 0.001 + 0.01 * 0.999 = 0.01098$

P(B|A): verosimilitud

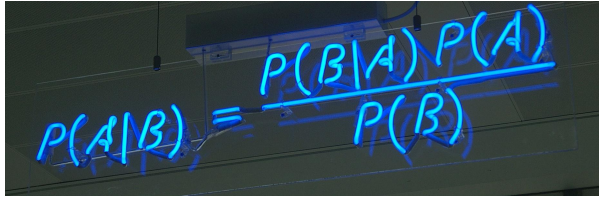
P(T+|E+): probabilidad de que el test de positivo dado que la persona está enferma. ¡Es la probabilidad de detección! = **0.99**

¡La probabilidad de Falso Positivo!

P(A|B): posterior o probabilidad a posteriori

P(E+|T+): probabilidad de estar enfermo dado que el test dio positivo.

Teorema de Bayes


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

A: estar enfermo **E+**

B: dio positivo **T+**

Si ponemos todo junto:

$$P(E+|T+) = P(T+|E+)P(E+)/[P(T+|E+)P(E+) + P(T+|E-)P(E-)] = 0.99*0.001/0.01098 = 0.0902$$



¿Para qué sirve?

1. El Teorema de Bayes tiene en cuenta automáticamente la prevalencia de las clases (en el caso visto Enfermos/No-Enfermos)
2. Dada una clasificación Binaria entre C+ y C-, llamamos X a los atributos, la formulación más general del problema de clasificación es:

$$P(C+|X) = P(X|C+)P(C+)/P(X) \text{ y } P(C-|X) = P(X|C-)P(C-)/P(X)$$



¿Para qué sirve?

1. El Teorema de Bayes tiene en cuenta automáticamente la prevalencia de las clases (en el caso visto Enfermos/No-Enfermos)
2. Dada una clasificación Binaria entre C+ y C-, llamamos X a los atributos, la formulación más general del problema de clasificación es:

$$P(C+|X) = P(X|C+)P(C+)/P(X) \text{ y } P(C-|X) = P(X|C-)P(C-)/P(X)$$

En general, es muy difícil formular de manera completa este problema. Necesitaríamos saber de qué tipo de distribución viene cada feature, cómo están correlacionados, etc.

Para la próxima

1. Ver los videos de la plataforma “Ajustes del Modelo”
2. Completar Notebook de hoy y atrasados.
3. Trabajar en la Entrega 03.

ACÀMICA