ACÁMICA

TEMA DEL DÍA

Preprocesamiento de Datos

Hoy:

- Escalado de datos
- Transformación de datos

Agenda

Daily

Explicación: Transformación de Datos

Break

Hands-on training

Cierre

Daily

Daily

Sincronizando...

Bitácora

¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

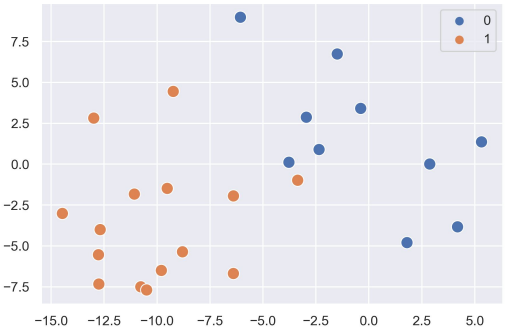
Challenge

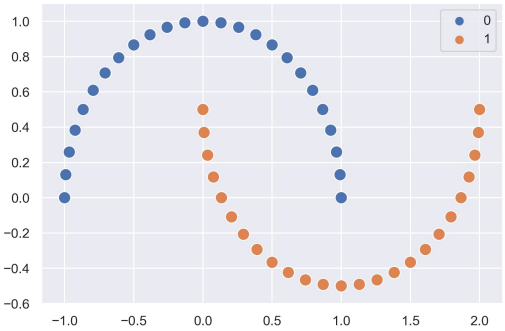
¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

Repaso general

Actividad en equipos (de 4 o 5 personas cada uno)

1. ¿Cuál es la diferencia entre aprendizaje supervisado y aprendizaje no-supervisado?
2. ¿Cuál es la diferencia entre un problema de clasificación y uno de regresión?
3. ¿Qué es el sobreajuste?
4. ¿Por qué es importante separar una porción del dataset antes de entrenar un modelo?
5. Describir el proceso por el cual un árbol de decisión *aprende* de los datos usando impureza Gini.
6. Definir Falsos Positivos, Falsos Negativos, Precisión y Exhaustividad.
7. Dibujar, aproximadamente, las fronteras de decisión que obtendrían con un árbol de decisión de profundidad uno, un árbol de decisión de profundidad dos, KNN con *k=1* y KNN con *k=número de muestras* en los siguientes casos:





*Además, elegir un caso y uno de los modelos y calcular cómo quedaría la matriz de confusión.

Repaso de la bitácora

Z-Score

Tenemos un conjunto de números x_1, x_2, x_3,... x_n. Su media es **μ**, y su desviación estándar **σ**.

$$Z = (x_i - \mu)/\sigma$$

Es una medida de cuánto se desvía un valor del promedio, medido en desviaciones estándar.

Ejemplo: $x_1 = 1$, $x_2 = 2$, $x_3 = 1.5$

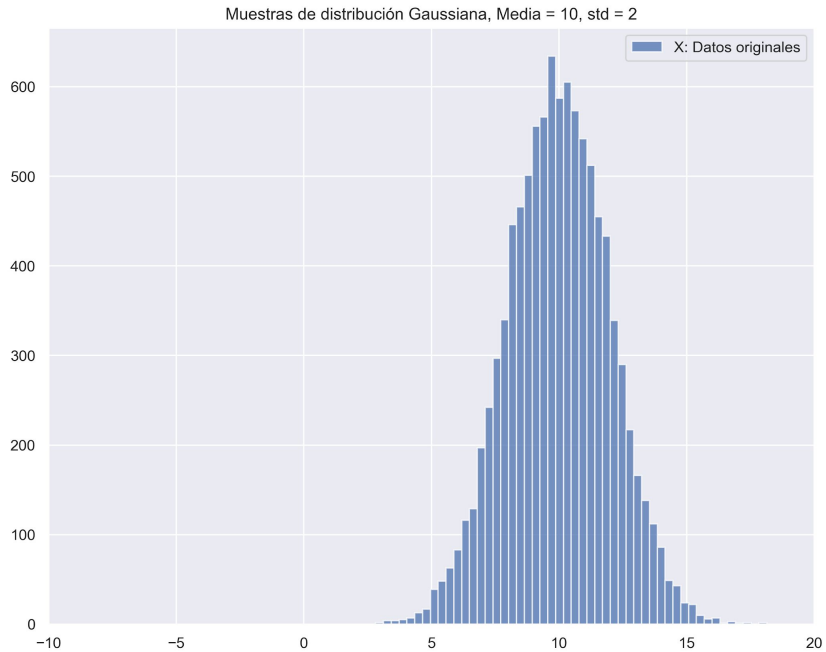
- Media, $\mu = 1.5$
- Desviación estándar, $\sigma = 0.5$

$x_1 = 1 \longrightarrow z_1 = (1 - 1.5)/0.5 = -1$

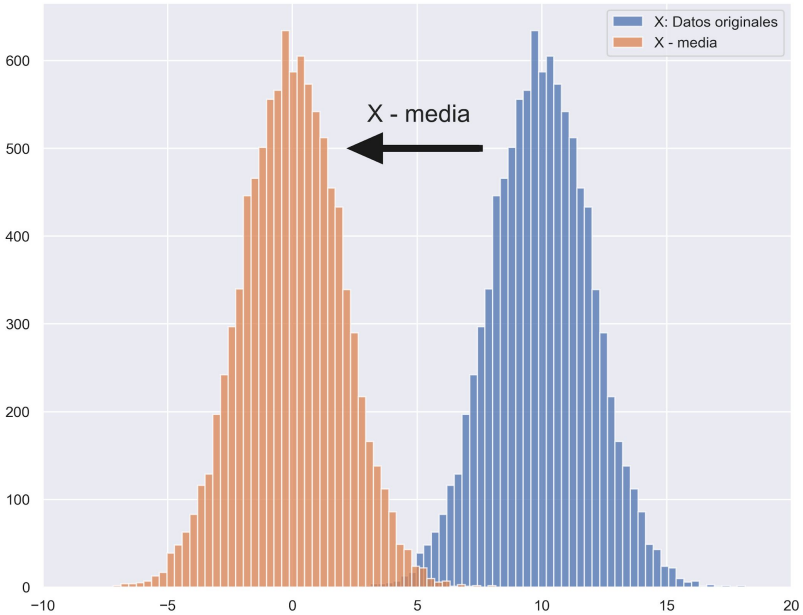
$x_2 = 2 \longrightarrow z_2 = (2 - 1.5)/0.5 = 1$

$x_3 = 1.5 \longrightarrow z_3 = (1.5 - 1.5)/0.5 = 0$

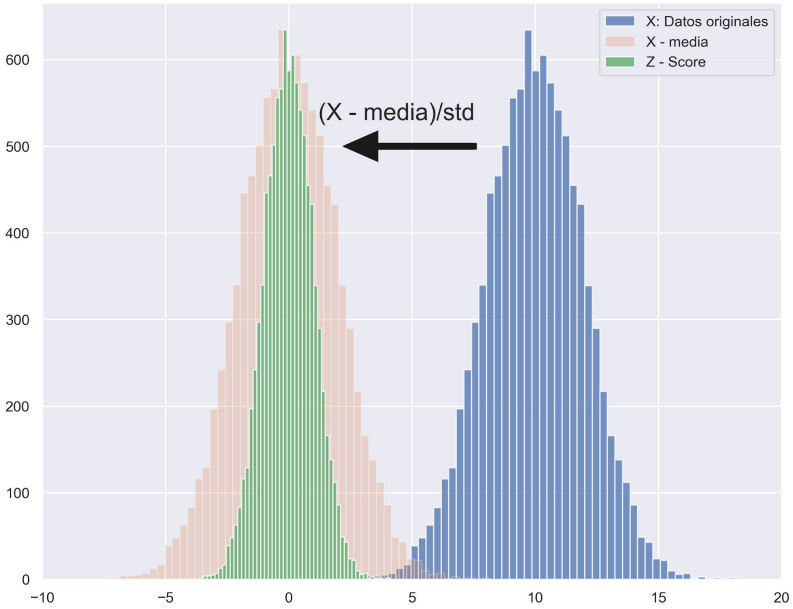
Z-Score


Muestras de distribución Gaussiana, Media = 10, std = 2
X: Datos originales
600
500
400
300
200
100
0
−10
−5
0
5
10
15
20

Z-Score


X: Datos originales
X - media
X - media
600
500
400
300
200
100
0
−10
−5
0
5
10
15
20

Z-Score


X: Datos originales
X - media
Z - Score
(X - media)/std
600
500
400
300
200
100
0
−10
−5
0
5
10
15
20

¿Recuerdas la Regla de las tres sigmas?

A veces, este método se aplica a través del Z-Score

Z-Score

- El Z-Score es una medida de cuán lejos está un dato del promedio de la muestra a cual pertenece, medido en desviaciones estándar.

- Es una medida muy usada para hablar de variables y distribuciones, en particular en ámbitos como Medicina.

- Nos sirve para **Escalado de Datos**. A veces lo podrán encontrar por el nombre de Estandarización o Normalización.

- También sirve para aplicar la regla de las tres sigmas, ya que luego de normalizar los datos, es simplemente usar una máscara $|Z| > 3$ para seleccionar valores atípicos.

- En Scikit-Learn, existe una clase *StandardScaler* del módulo *preprocessing* que lo implementa.

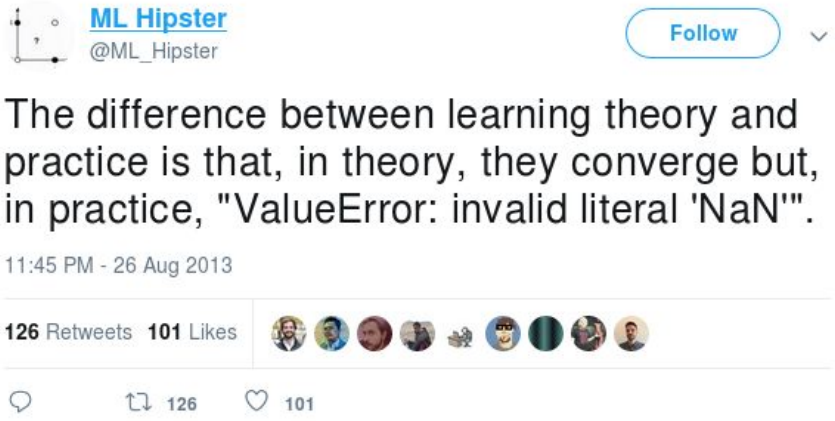
Escalado de Datos

- Es una buena práctica, antes de entrenar cualquier modelo, escalar los datos. En general, se utiliza el Z-Score para la mayoría de los atributos.

- El escalado es particularmente importante cuando se trate de modelos que involucren distancias u operaciones matemáticas donde los ordenes de magnitud pueden confundir al modelo (kNN, regresión lineal y logística). No lo es tanto en modelos como árboles de decisión.

- Si hay un atributo binario numérico (0/1), no hace falta escalarlo.

Transformación de Datos



https://twitter.com/ML_Hipster/status/372248401951215616

¡La computadora, en el fondo, sólo entiende de números!

Pero no todos los features vienen en “números”.

Por ejemplo, en el dataset de críticas de vinos tenemos:

- País, Provincia y Región de Origen
- Variedad
- Bodega
- Etc.

Supongamos que queremos clasificar los vinos según algunos atributos.

¿Cuáles creen que serían útiles?

¿Cuáles creen que faltan?

Supongamos que queremos clasificar los vinos según algunos atributos.

¿Cuáles creen que serían útiles?

¿Cuáles creen que faltan?

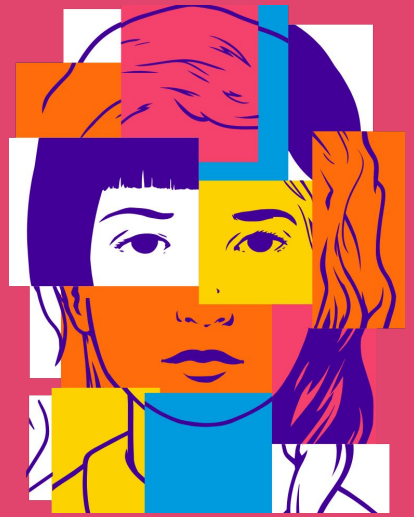
¿Y si queremos predecir el precio?¿O si es bueno (puntaje alto) o malo (puntaje bajo)?

La pregunta que querramos responder nos va a indicar cómo tenemos que trabajar con nuestro dataset.

Pero, en general, hay algunas cosas que **no pueden faltar:**

- Conversión de features.
- Tratamiento de datos faltantes.
- Selección y combinación de variables.
- Y más.

Tipos de datos



Variables numéricas

Son aquellas variables que se miden o se cuentan (discretas o continuas).

- Hay una relación de orden entre ellas
- Se pueden sumar (en algunas circunstancias)

Ejemplos:

- Edad, Altura y Peso de una persona
- Puntaje, precio de un vino
- Valor de un pasaje
- Etc.

Tratamiento

En general, ya vienen en un formato “cómodo” para trabajar, pero a veces queremos agruparlas según grupos o rangos.

Ejemplo: agrupar edades en rangos (bebés, niños, adolescentes, adultos, ancianos)

→ **Discretización y Binning**

Variables ordinales

Sus posibles valores son categorías, pero hay una relación de orden.

¡Notar que no se pueden sumar!

Ejemplos:

- Tamaño de una prenda de ropa: XS, S, M, L, XL
- Tipo de Nafta por octanaje: 95, 98, más de 98.
- Rangos etarios: bebé, niño/a, adolescente, adulto/a, anciano/a

Tratamiento

Podemos hacer una asignación a número enteros manteniendo el orden:

$$S \rightarrow 0$$
$$M \rightarrow 1$$
$$L \rightarrow 2$$

Pero, ¡cuidado!, recordar que no se pueden sumar.

Variables nominales

Sus posibles valores pertenecen a una de varias categorías.

- Las categorías no siguen una relación de orden
- Ninguna es mayor que otra

Ejemplos:

- Nacionalidad
- Tipo de vino
- Color de una prenda de ropa
- Género: femenino, masculino, no binario, etc.

Tratamiento

- Llevar a variables dummies/One-Hot Encoding.
- Hay que tener cuidado porque puede hacer que nuestro dataset crezca mucho.

Resumiendo

Conversión de variables: Los modelos sólo entienden de números. ¿Si los atributos no son números?

Tipos de variables a tratar:

- Numéricas: edad, altura, puntaje.
- Ordinales: tamaño de una prenda de ropa.
- Categóricas/nominales: nacionalidad, color de una prenda de ropa.
- Y más...

Tratamiento

- Discretización/binning
- “Labelización”[1]
- Variables dummies/One-Hot encoding
- Y más...

[1] Tal vez inventamos una palabra

Scikit-Learn y Pandas

Transformación de datos con Pandas

Googlear:

- **Discretización y binning:**
  - ¿Qué funciones podemos usar?
  - ¿Cuáles son sus argumentos?
  - ¿En qué situaciones será conveniente?

- **¿Qué hace la función map()?**
  - ¿Qué toma como argumento?
  - ¿Sobre qué objeto opera?

- **¿Qué es una *variable dummy*?**
  - ¿Cómo funciona *get_dummies()* de Pandas?

¡BREAK!

Ph. Credit: Drew Coffmann

Hands-on training

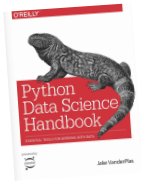
Hands-on training

DS_Bitácora_19_20_Preprocesamiento.ipynb

Descarga aquí los datos sin valores faltantes.

Recursos

Recursos



- Valores faltantes: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
- Capítulo 3, “Data Manipulation With Pandas”, de Python Data Science Handbook
- Estadística en general: Serie de cinco artículos sobre Estadística en Data Science: https://towardsdatascience.com/statistics-is-the-grammar-of-data-science-part-2-8be5685065b5

Para la próxima

- Termina el notebook de hoy.
- Lee la bitácora 21 y carga las dudas que tengas al Trello.

En el encuentro que viene uno/a de ustedes será seleccionado/a para mostrar cómo resolvió el challenge de la bitácora. De esta manera, ¡aprendemos todos/as de (y con) todas/as, así que vengan preparados/as.

ACÁMICA