

Laboratorio de datos, clase 3

Preparación, estandarización y normalización de datos

Prof. Enzo Tagliazucchi

tagliazucchi.enzo@googlemail.com

www.cocuco.org

**I HAVE NO IDEA
WHAT I'M DOING**

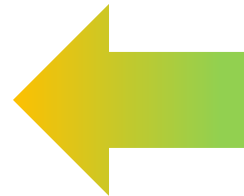


conseguir los datos
(si existen) o planear
el experimento (si no)



Visualizar y comunicar
los resultados

Explorar los datos,
buscar evidencia,
desarrollar modelos



¿Qué vamos a ver hoy?

- Por qué los datos no son lo que pensamos que son
- Cómo podemos mejorar la calidad de los datos
- Cómo podemos hacer que los datos numéricos de distintas fuentes sean más comparables entre sí

An illustration of an iceberg floating in the ocean. The tip of the iceberg is above the water line, and the much larger base is submerged. A small penguin is standing on the tip. The text is placed on the iceberg to represent different data quality issues.

Datos faltantes

100 cm vs 1 m

Datos inconsistentes

Errores de tipeo

¿3.4 o 3,4?

La persona que ingresa los datos está presa de una creencia errónea sobre el dato y va a actuar en consecuencia como si su vida dependiese de ello

La mejor manera de evitar estos problemas es reducir la libertad de quien ingresa los datos

Label

|Input text

Labo de datos 2021



¿Qué hacer con los datos faltantes?

Descartar la observación

Reemplazar por la media o la moda

Reemplazar por la media o la moda computada sobre las observaciones más similares

Usar machine learning para estimar la probabilidad de que el dato faltante tome distintos valores



A veces, los datos faltantes nos dan información en sí mismos



¿por qué nadie en el vuelo RL 239 completó la encuesta de satisfacción al cliente?

Datos faltantes \neq np.NaN

... , 'nan' , ...

.. , , ...

... , 'falta' , ...

... , -999 , ...

... , 'cosme fulanito' , ...

Rescaleo con a y b constantes

$$\tilde{x}_i = a(x_i + b)$$

Por ejemplo, pasar de un sistema de unidades a otro, o comparar los datos contra un valor de referencia

(e.g. $b = -\text{promedio hist\u00f3rico}$)

Normalización

$$\tilde{x}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Normalización min-max (entre 0 y 1)

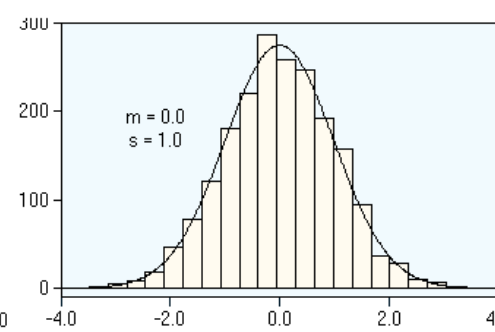
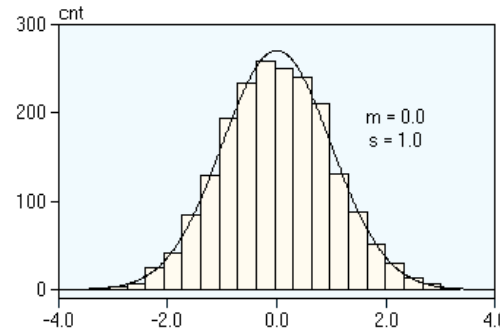
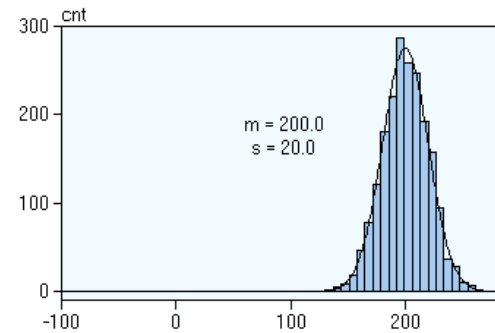
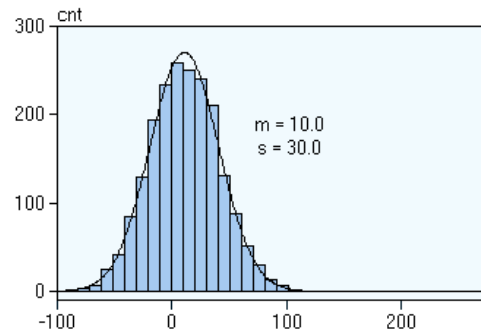
Es útil aunque los datos no sigan una distribución normal, pero es muy sensible a outliers (porque aparecen max y min)

2, 1, 3, 5, 3, 6, **10932**, 10 →

0.0001, 0, 0.0002, 0.0004, 0.0002, 0.0005, **1**, 0.0008

Estandarización

$$\tilde{X}_i = \frac{X_i - \langle x_i \rangle}{\sigma}$$

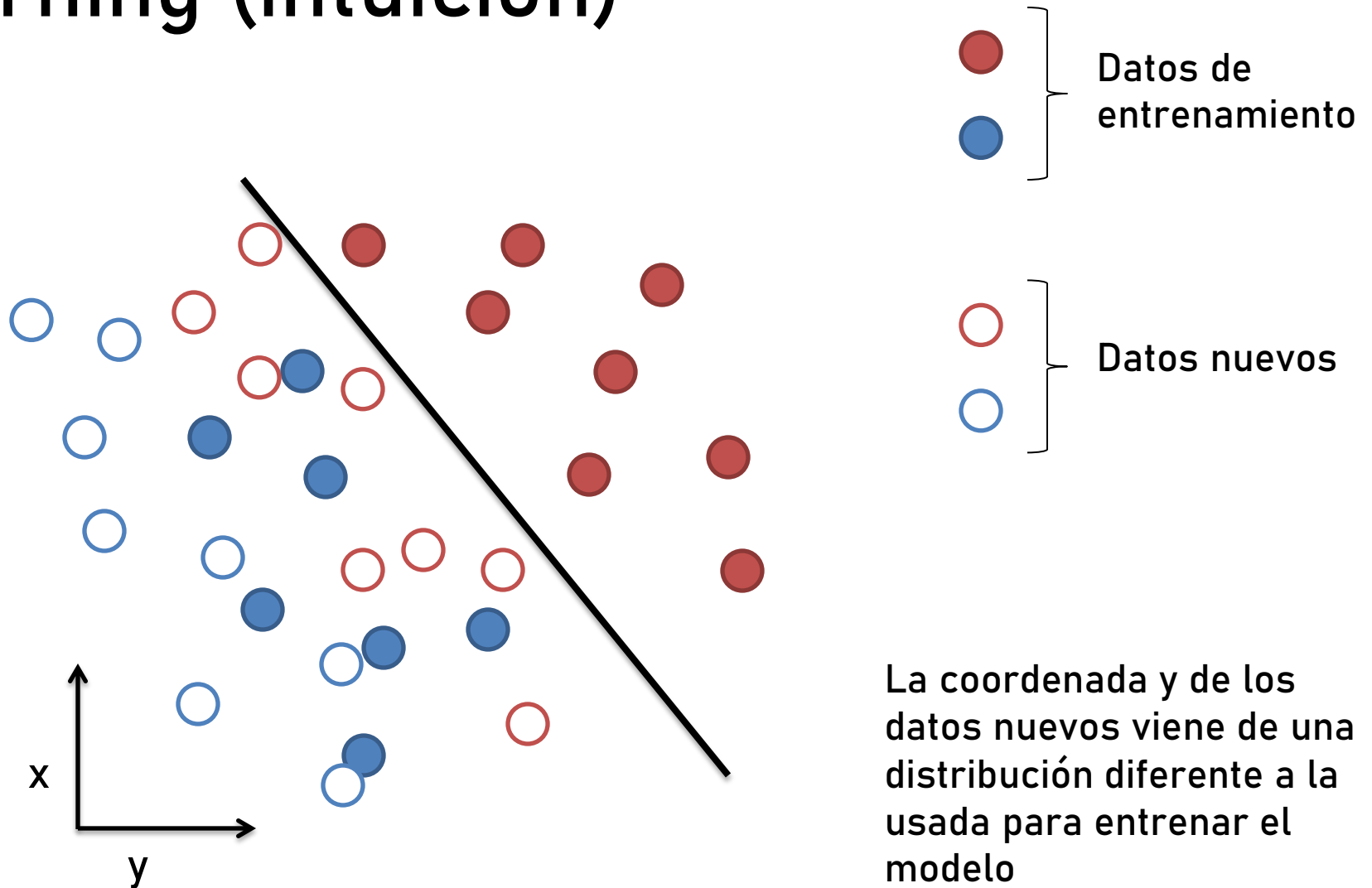


Es resistente ante la presencia de outliers, pero no es demasiado interpretable para datos con distribuciones que no sean gaussianas.



2, 1, 3, 5, 3, 6, **10932**, 10 →

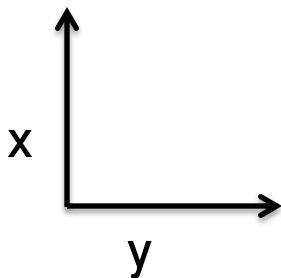
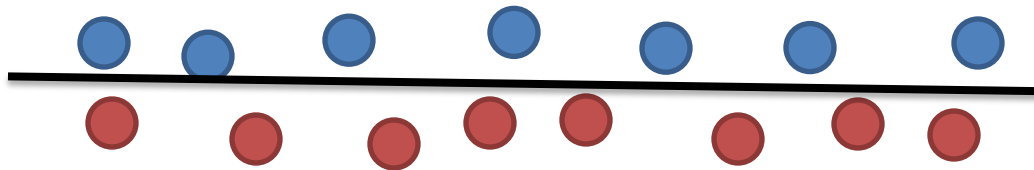
-0.3541, -0.3544, -0.3539, -0.3534, -0.3539, -0.3531, **2.4749**, -0.3521

Por qué puede ser vital para machine learning (intuición)



Por qué puede ser vital para machine learning (intuición)

  } Datos de entrenamiento



Si la coordenada y tiene un rango de valores muchísimo más amplio que x , domina el entrenamiento y siempre tengo el mismo modelo

Próxima clase:

**¿Cómo podemos caracterizar numéricamente
nuestros datos?**

Estadística descriptiva