

(Hitz Research Center), Departamento de LSI, EHU

Sist. de Ayuda a la Decisión: Preprocesado de los datos

AitZiber AtutXa

aitziber.atucha@ehu.eus

Índice

- 1 Objetivos de Aprendizaje
- 2 Exploración de los datos
- 3 Preprocesado de los Datos

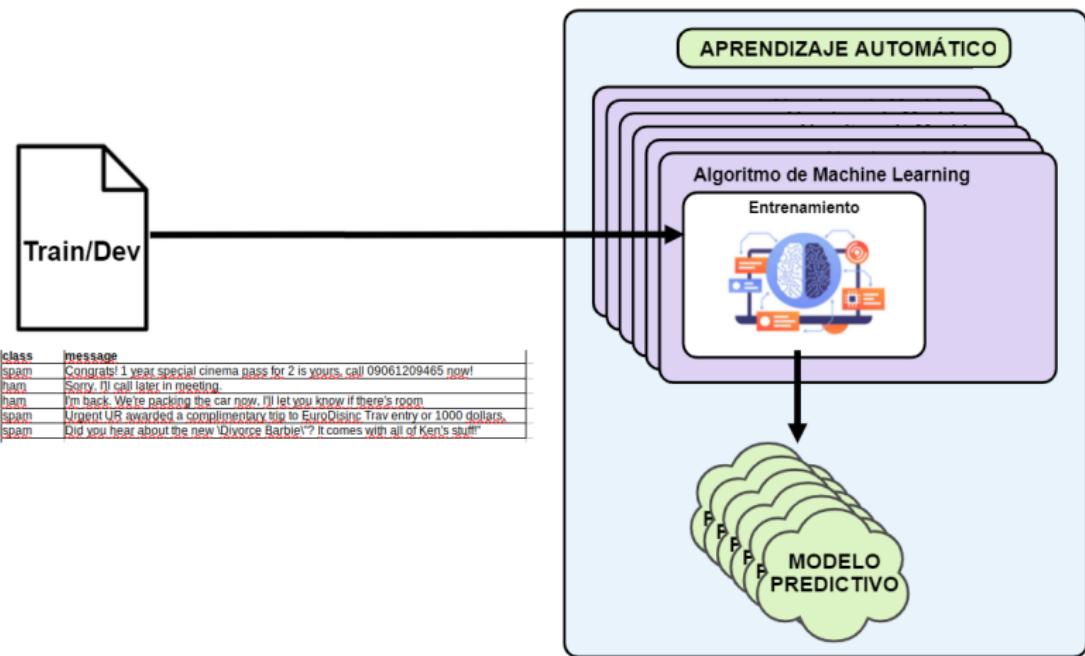
Objetivos de Aprendizaje

- Entender la necesidad del preprocesado de los datos.
- Entender la necesidad de la exploración previa
 - Saber caracterizar los diferentes tipos de datos
 - Saber caracterizar una muestra a través de medidas estadísticas (media, mediana, desviación estandar,...)
 - Saber emplear el gráfico apropiado para representar cada tipo de dato

Objetivos de Aprendizaje

- Conocer técnicas de preprocesado: su entrada y salida
 - Discretización
 - Normalización
 - Tratamiento de valores faltantes
- Comprender la necesidad de transformar los datos en valores numéricos para realizar predicciones
- Conocer distintos métodos de muestreo: implicaciones

Terminología



Terminología

Modelo Predictivo (parametrizable¹)

Algoritmo matemático: aprende comportamiento de ejemplos (muestra entrenamiento) y predice el comportamiento de ejemplos no vistos anteriormente (test).

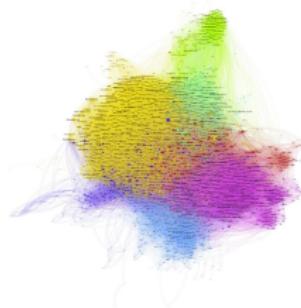
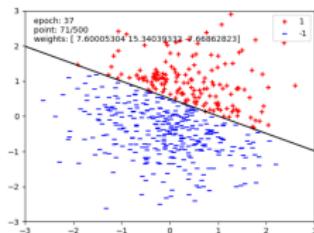
Ejemplos

- Modelo de predicción:
 - Editorial: Saber cuál será el número bestseller
 - KONNE: Saber cuando sucederá una avería
 - Sistemas Dialogo: Saber cuál es el sujeto y cuál el objeto
 - Predicción de FA: Saber si recurrencia de FA

¹Aprende parámetros que suelen ser pesos "cantidad de importancia" de 6/37 cada atributo

Exploración de los datos

- Aprendizaje Supervisado:
 - Clasificación
 - Regresión
- Aprendizaje No-Supervisado:
 - Clustering
 - Asociación

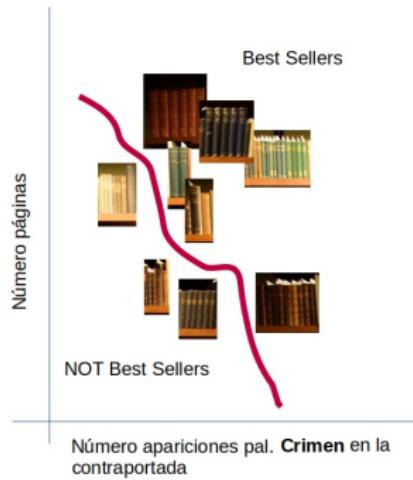


Clasificación



Origen:

Objetivo:



Clustering

Origen:

							
Edad: 42 Suscripción: 7	Edad: 18 Suscripción: 3	Edad: 23 Suscripción: 3	Edad: 49 Suscripción: 1	Edad: 37 Suscripción: 7	Edad: 51 Suscripción: 1	Edad: 40 Suscripción: 6	Edad: 20 Suscripción: 4

Objetivo:



Terminología

- Identificar los atributos(rasgos,features) para caracterizar (representar) cada ejemplo (instancia) de aprendizaje.
 - Editorial: Saber cuál será el número bestseller
 - KONNE: Saber cuando sucederá una avería
 - Sistemas Dialogo: Saber cuál es el sujeto y cuál el objeto
 - Predicción de FA: Saber si recurrencia de FA
- Si es posible establecer (cuantificar) cuantas instancias vamos a necesitar

Terminología

Ejemplo: BestSeller

Muestra: Un conjunto de 10000 libros

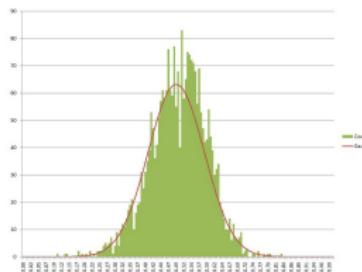


- Cada libro es una instancia de entrenamiento
- Por cada instancia (libro): Varios atributos ¿Qué información es relevante? (*feature engineering*)
 - pag
 - freq. palabra *Crimen*
 - ...

BestSeller	#pag.	FreqCrimen
0	1000	0
0	700	1
0	750	0
1	300	10
0	800	2

Exploración de los datos

- Tamaño de la muestra (num. instancias)
- Número atributos
- Tipos de atributos:
 - Númericos: Continuos y Discretos
 - Categóricas: Ordinales y Nominales (del latín *nomine*)
- Medidas estadísticas (atributos numéricos):
 - Centrales: Media, Mediana y Moda
 - De Dispersion: Varianza, desviación std
- Otras caraterizaciones:
 - Número de valores únicos
 - Outliers (valores extremos) Missing o valores faltantes



Tipos de Atributos

- Atributos Numéricos (Cuantitativos):
 - Discretos: Muestran un número finito de valores entre dos valores. Ejemplos: [1,2,3..10] o [1.1,1.2,1.3,1.4,1.5,1.6]
 - Continuos: Muestran un número infinito de valores entre dos valores. 1.2, 1.25, 1.3, 1.5, etc. Por ejemplo, el redondeo es una forma de convertir una variable continua en discreta.
- Datos Categoriales (Cualitativos):
 - Nominales: No muestran ningún tipo de orden importante en la tarea. [Americano, Africano, Asiatico]
 - Ordinales: Muestran algún tipo de orden importante en la tarea, se pueden ordenar. Por ejemplo, rangos numéricos [1-10,10-20,20-30], [fácil,difícil]

Tipos de Atributos

Ejercicio: ¿Qué tipos de Datos?

Identifica cada tipo de dato que aparece en esta tabla:

Altura	Edad	Color Cabello	Dominio del Inglés
1.785	18	Rubio	Alto
1.80	20	Moreno	Medio
1.90	22	Pelirojo	Bajo
1.9267	22	Rubio	Bajo
1.60	42	Moreno	Alto

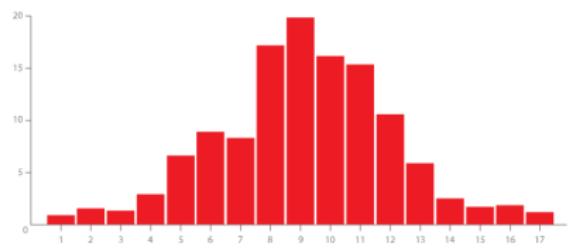
Cuadro: Ejemplos de tipos de datos

Continuo, Discreto, Nominal, Ordinal

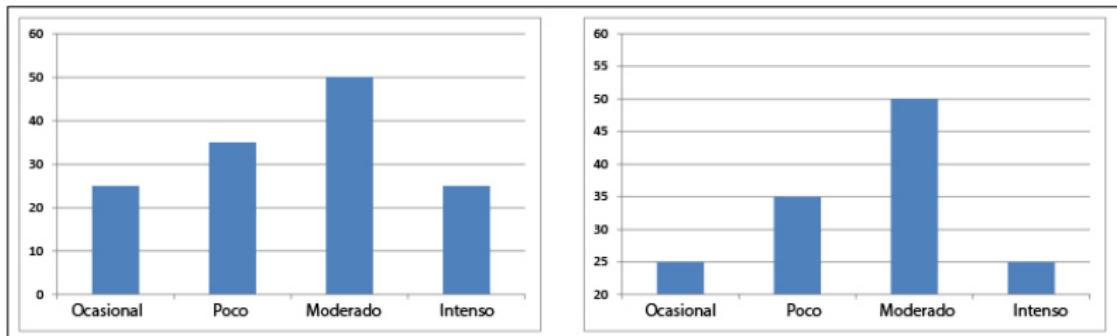


Representación de una variable X

- Histogramas: Variable es cuantitativa

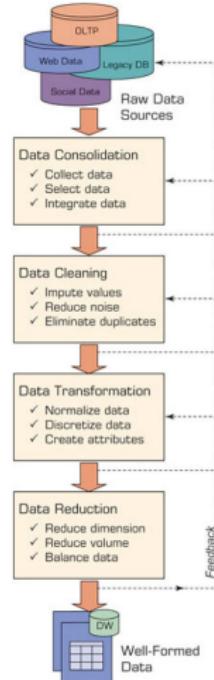


- Gráfico de Barras: Variable es cualitativa



Preprocesado de los Datos

- Obtención de los Datos
- Limpieza de los Datos
- Transformación de los Datos
- Reducción de los Datos



El preprocessado de los Datos

Tarea	Subtarea	Métodos
Obtención Datos	Acceder y recopilar los datos Seleccionar y filtrar los datos Integrar y unificar los datos	Consultas SQL, agentes de software, servicios web. Experiencia en el dominio, consultas SQL, pruebas estadísticas. Consultas SQL, experiencia en el dominio, mapeo de datos basado en ontologías.
Limpieza de Datos	Tratar los valores faltantes en los datos (missing values)	Completar los valores faltantes imputándoles los valores más apropiados (media, mediana, mín./máx., moda, etc.); sustituir los valores que faltan con una constante como NA, o un valor no posible (-1); eliminar la instancia que contiene el valor faltante; No hacer nada.
	Identificar y reducir el ruido en los datos.	Identificar los valores atípicos (outliers) en los datos con técnicas estadísticas simples (como valores centrales y desviaciones estándar) o con análisis de conglomerados; una vez identificados, eliminar los valores atípicos o suavizarlos mediante el uso de intervalos, regresión o promedios simples.
	Encontrar y eliminar datos erróneos	Identificar los valores erróneos en los datos (que no sean valores atípicos), como valores impares, etiquetas de clase inconsistentes, distribuciones impares; una vez identificados, emplear la experiencia en el dominio para corregir los valores o eliminar las instancias que contienen los valores erróneos.

El preprocessado de los Datos

Tarea	Subtarea	Métodos
Transformación de los Datos	Normalizar los datos (reducir la dispersión)	Reducir el rango de valores en cada variable valorada numéricamente a un rango estándar (por ejemplo, entre 0 y 1 o -1 y +1) usando una variedad de técnicas de normalización o escala.
	Discretizar o agregar los datos	Si es necesario, convierta las variables numéricas en representaciones discretas utilizando técnicas de binning basadas en rango o frecuencia; para las variables categóricas, reduzca el número de valores aplicando jerarquías de conceptos adecuadas.
	Generar nuevos atributos	Generar nuevos atributos más informativos a partir de los existentes utilizando una amplia gama de funciones matemáticas (tan simples como la suma y la multiplicación o tan complejas como una combinación híbrida de transformaciones logarítmicas).
Reducción Datos	Reducir el número de Atributos	Emplear PCA (análisis de componentes principales), el análisis de componentes independientes, las pruebas de chi-cuadrado, el análisis de correlación y la inducción del árbol de decisión.
	Reducir el número de registros	Realizar un muestreo aleatorio, muestreo estratificado, muestreo intencionado basado en el conocimiento de expertos.
	Equilibrar datos desbalanceados (sesgados)	Sobremuestree (Oversampling) las clases menos representadas o submuestree las (undersample) clases más representadas.

Limpieza de los datos

- Reducir el ruido:
 - Tratamiento de valores atípicos
 - Suavizado
 - Eliminación
 - Eliminación de valores erroneos
- Tratar Valores Faltantes:
 - Eliminar las instancias que los contienen
 - Imputar un valor (media, moda,predicción)



Limpieza de los datos: Discretizar Valores

Cuando por ejemplo trabajamos con valores continuos, o nuestra muestra es muy pequeña.

- Los valores continuos crean dificultades para encontrar patrones porque son muy dispersos.
- cuando la muestra es muy pequeña y disponemos de pocos ejemplos representativos de un valor

Solución: Discretizar (Binning)

- Por frecuencia: Agrupamos elementos que tienen la misma frecuencia
- Por amplitud: $[min + w], [min + 2w] \dots [min + nw]$ where $w = (max - min) / (\text{núm. of bins})$

Limpieza de los datos

- Transformación de los Datos:
 - Escalar valores
 - Discretizar valores
 - Generar nuevos atributos



Limpieza de los datos: Escalar Valores

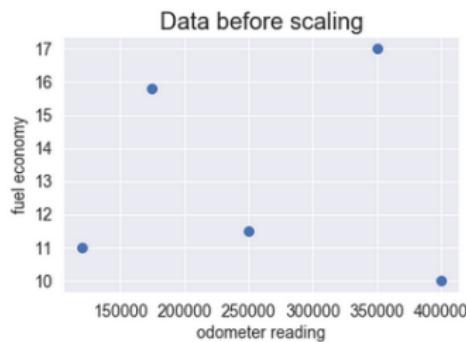
- Escalar los valores
 - Relativizar con respecto al max
 - Relativizar con respecto al max y min
 - z-scale



Limpieza de los datos: Escalar Valores

Datos antes de escalar

	odometer_reading	fuel_economy
0	120000	11.0
1	250000	11.5
2	175000	15.8
3	350000	17.0
4	400000	10.0



Limpieza de los datos: Escalar Valores

- Relativizar con respecto al max

$$maxScale(x) = \frac{x}{max}$$

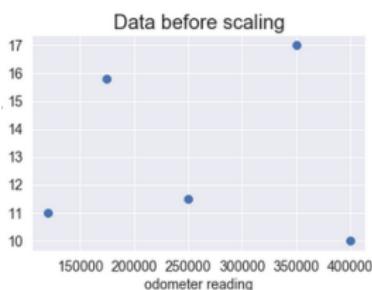
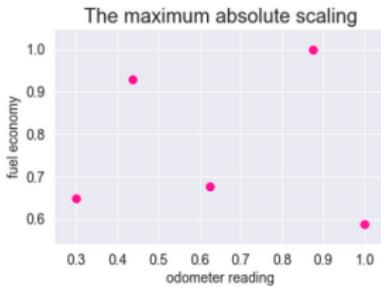


Limpieza de los datos: Escalar Valores

Datos después y antes de escalar relativizando con respecto al max

	odometer_reading	fuel_economy
0	0.3000	0.647059
1	0.6250	0.678471
2	0.4375	0.929412
3	0.8750	1.000000
4	1.0000	0.588235

	odometer_reading	fuel_economy
0	120000	11.0
1	250000	11.5
2	175000	15.8
3	350000	17.0
4	400000	10.0



Limpieza de los datos: Escalar Valores

- Relativizar con respecto al min y max

$$\text{minMaxScale}(x) = \frac{x - \text{min}}{\text{max} - \text{min}}$$

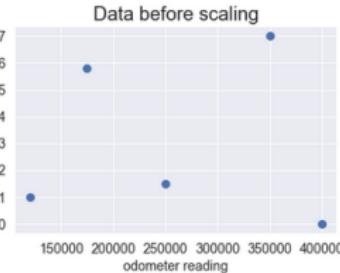
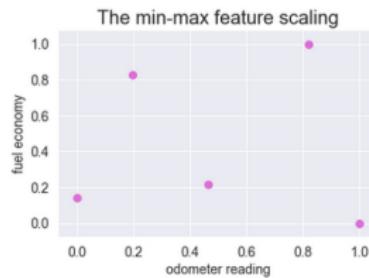


Limpieza de los datos: Escalar Valores

Datos después y antes de escalar relativizando con respecto al max y min

	odometer_reading	fuel_economy
0	0.000000	0.142857
1	0.464286	0.214286
2	0.196429	0.828571
3	0.821429	1.000000
4	1.000000	0.000000

	odometer_reading	fuel_economy
0	120000	11.0
1	250000	11.5
2	175000	15.8
3	350000	17.0
4	400000	10.0



Limpieza de los datos: Escalar Valores

- z-scale: Relativizar con respecto a la media μ y desviación std σ

$$z-score(x) = \frac{x-\mu}{\sigma}$$

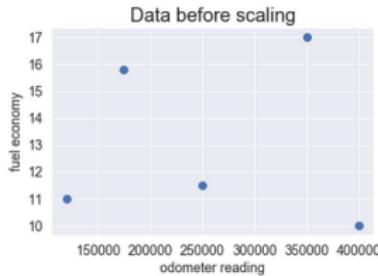
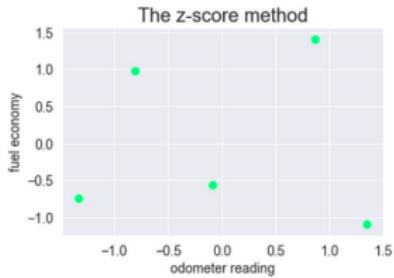


Limpieza de los datos: Escalar Valores

Datos después y antes de escalar relativizando con respecto a la media y la desviación estandar

	odometer_reading	fuel_economy
0	-1.189512	-0.659120
1	-0.077019	-0.499139
2	-0.718842	0.876693
3	0.778745	1.260647
4	1.206628	-0.979081

	odometer_reading	fuel_economy
0	120000	11.0
1	250000	11.5
2	175000	15.8
3	350000	17.0
4	400000	10.0



Limpieza de los datos: Escalar Valores

Ejercicio: ¿Qué valores son faltantes, erroneos, necesitan ser escalados?

Nombre	Edad	Sexo (0/1)	Núm. alquileres-anual
Asier Arrutia	35	Varón (0)	350
Xabier Axular	28	Varón (0)	240
Miren Lopetegi		Mujer (1)	300
Unai Lopetegi	22	Varón (0)	350
Jorge Zelaia		Varón (0)	240
Joseba Lakarra	460	Varón (0)	290
Jorge Goenaga	45	Varón (0)	280
Mireia Baztan	19	Mujer (1)	
Carlos Fdz	26	Varón (0)	280
Luis Garcia	45	Varón (0)	290
Itziar Ganbara	-20	Mujer (1)	270
Idoia Ara	20	Mujer (1)	270



Transformación de datos: Balanceo

En tareas de clasificación cuando las clases están desbalanceadas:

Ejemplos

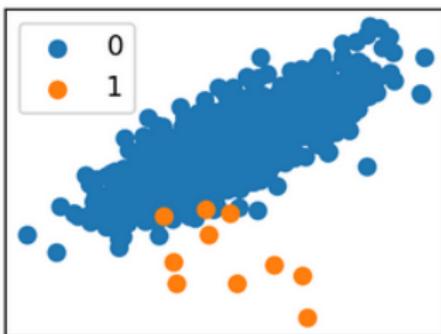
- Predictor de Enfermedades Raras (Enfermedad rara es aquella cuya prevalencia es inferior a 5 casos por cada 10.000 personas en la Comunidad Europea.)
- Predictor de bestsellers
- Predictor de riesgo de impago de hipoteca en una sucursal de una zona donde la renta per capita es superior a los 200000 euros anuales

Transformación de datos: Balanceo

Preguntas.

¿Que se os ocurre intuitivamente que se puede hacer?

¿Cómo puedo balancear los ejemplos positivos y negativos?



Transformación de datos: Balanceo

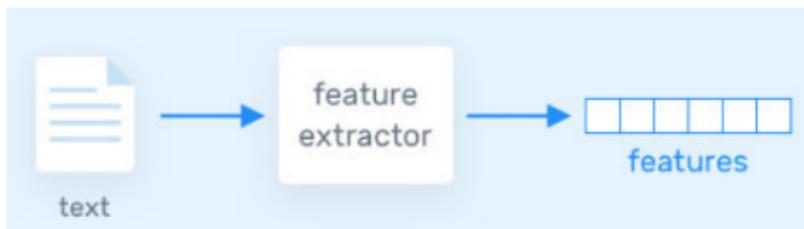
Aplicación de técnicas de over y undersampling:

- Undersampling
 - Eliminar instancias de la clase mayoritaria al azar.
Problemas: Podemos estar eliminando ejemplos relevantes
 - Eliminar instancias que estén cercanas a instancias de la clase minoritaria (Tomek links). *Border examples.*
- Oversampling
 - Copiar instancias de la clase mayoritaria
 - Generar instancias de forma sintética. Por ejemplo medias entre dos puntos de la clase mayoritaria.

Limpieza de los datos: Tratar texto

¿Cómo tratamos el texto?

Técnicas para convertir texto en vectores numéricos



Técnicas de conversión txt2vec

- BOW (Bag of Words)
- tf-idf
- hashing

Limpieza de los datos: Tratar texto

BOW (Bag of Words)

- Se toma el diccionario de todas las palabras aparecidas en los documentos y se genera un vector donde cada dimensión es una palabra cuyo contador se inicializa a 0
- Por cada documento (instancia de entrenamiento) se rellena ese vector actualizandolo, y sustituyendo los 0s por la frecuencia de aparición

Limpieza de los datos: Tratar texto



Se suele emplear el ***tf-idf***.

$$tfidf(t_x, d, D) = tf(t, d) \times idf(t_x, D)$$

$$tf(t_x, d) = \frac{\#t_x}{\sum_{i=1}^n t_i}$$

$$idf(t_x, D) = \log(D / (1 + \#t_x))$$

Limpieza de los datos: Tratar texto

tf-idf: $tf \times idf$

- El término frecuencia (tf) de una palabra en un documento: La versión más simple es un recuento de las apariciones de la palabra en el documento. Luego, se normaliza, dividiendo por la longitud de un documento.
- idf (inverse document frequency): Forma de ponderar lo común/frecuente o rara es una palabra en los documentos. Cuanto más cerca está de 0, más común. Se puede calcular tomando el número total de documentos, dividiéndolo por el número de documentos que contienen una palabra. Se aplica el logaritmo para escalar el valor. $\frac{\text{NumDocs}}{\text{FrecAparicionPalEnLosDocs}}$. Si el número de documentos es muy grande el idf explotaría. Se emplea el logaritmo para coseguir escalar el valor, reduciendo su dimensionalidad. Por ejemplo $\log_{10} 100 = 2$