



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

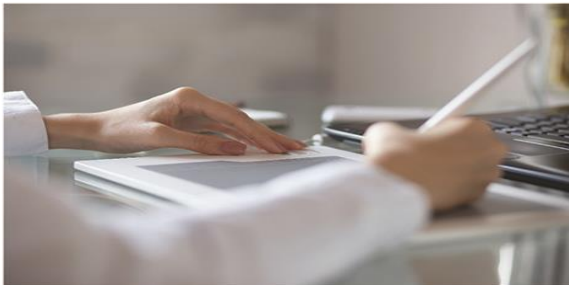
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

Otros temas de su interés

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau



Otros temas de su interés

Reducción de Dimensionalidad

Clustering Jerárquico

Series de Tiempo

Minería de Texto

Procesamiento de Imágenes

Reducción de Dimensionalidad

Reducción de dimensionalidad

Supongamos que tenemos m variables explicativas

El objetivo es construir n nuevas variables (con $n < m$) que sean capten de la mejor manera posible la información contenida en las m variables originales

Un modelo que considere las n nuevas variables será más sencillo y tendrá un rendimiento similar a un modelo que considere las m variables originales

Análisis de Componentes Principales

El objetivo es explicar tanta variabilidad de las variables originales como sea posible

Mientras más componentes tengamos, más variabilidad explicaremos

Las componentes son una combinación lineal de las variables originales

Las componentes generadas son independientes, lo que facilita su inclusión en los modelos

Ejemplo

Encuesta de satisfacción (medida en escala de 0 a 100) a viajeros de una aerolínea

10 preguntas específicas con escala de 0 a 100 respecto de:

Puntualidad

Entretenimiento a bordo

Comodidad de asientos

Programa de viajero frecuente

Comida

Frecuencia de vuelos

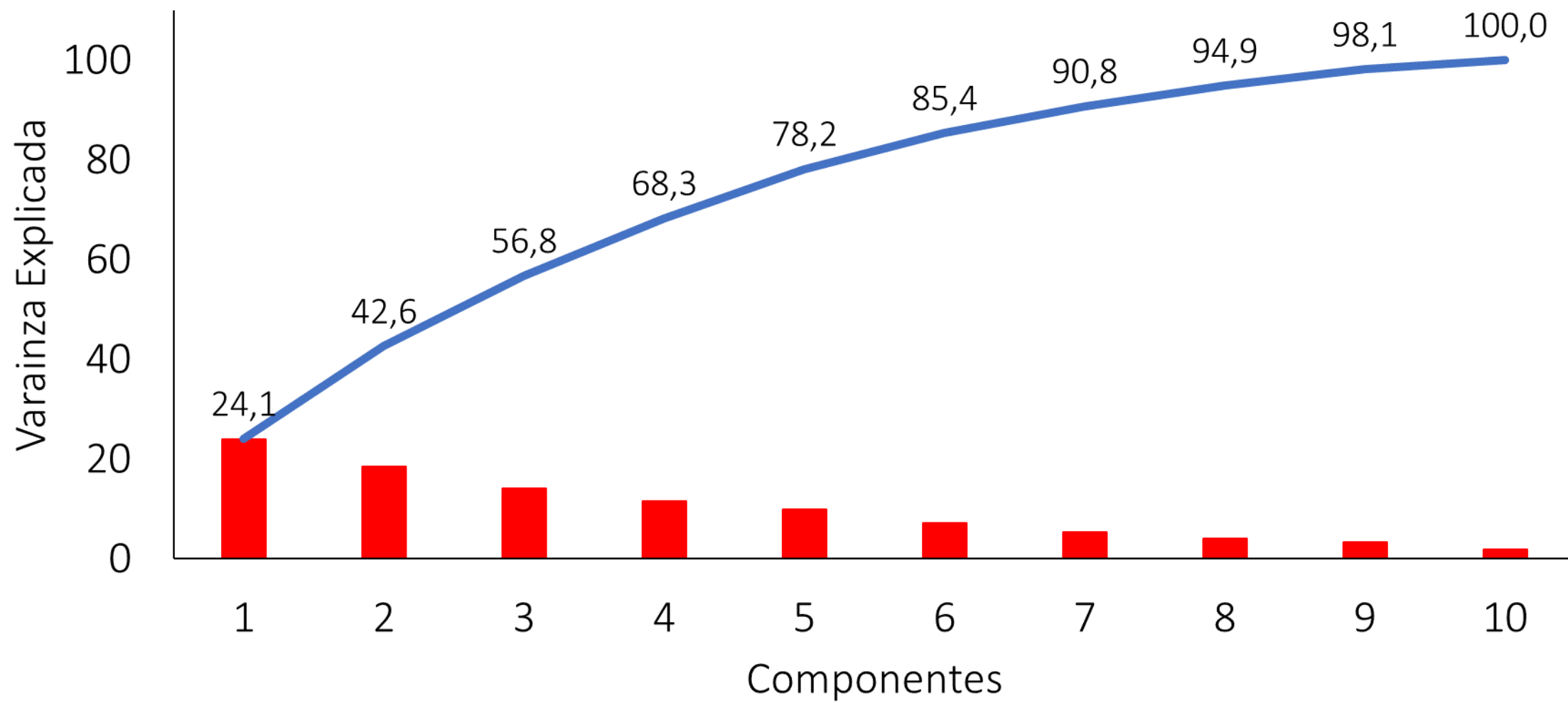
Simpatía de los tripulantes

Facilidad de maletas

Precio

Facilidad de pago

Ejemplo



Ejemplo

Variable	Componentes			
	1	2	3	4
Puntualidad	-0,054	0,453	-0,360	0,346
Comodidad Asientos	-0,548	0,027	-0,004	-0,044
Comida	-0,588	-0,024	0,165	0,092
Simpatía Tripulación	0,011	0,028	0,084	-0,244
Precio	-0,196	0,203	-0,445	-0,418
Entretenimiento a Bordo	-0,537	-0,012	0,229	0,109
Programa Viajero Frecuente	0,085	0,457	0,494	-0,239
Frecuencia de Vuelos	-0,026	0,472	-0,319	0,371
Facilidad de Maletas	0,074	0,559	0,377	-0,122
Facilidad de Pago	-0,103	0,090	-0,313	-0,643

Ejemplo

Variable	Componentes			
	1	2	3	4
Puntualidad		0,453		
Comodidad Asientos	-0,548			
Comida	-0,588			
Simpatía Tripulación				
Precio			-0,445	-0,418
Entretenimiento a Bordo	-0,537			
Programa Viajero Frecuente		0,457	0,494	
Frecuencia de Vuelos		0,472		
Facilidad de Maletas		0,559		
Facilidad de Pago				-0,643

Clustering Jerárquico

Clustering Jerárquico

Otro método de clustering, alternativo a K Means

Existen dos tipos:

Aglomerativo: las observaciones se van juntando de una a la vez

Divisivo: las observaciones se van separando de una a la vez

Clustering Jerárquico Aglomerativo

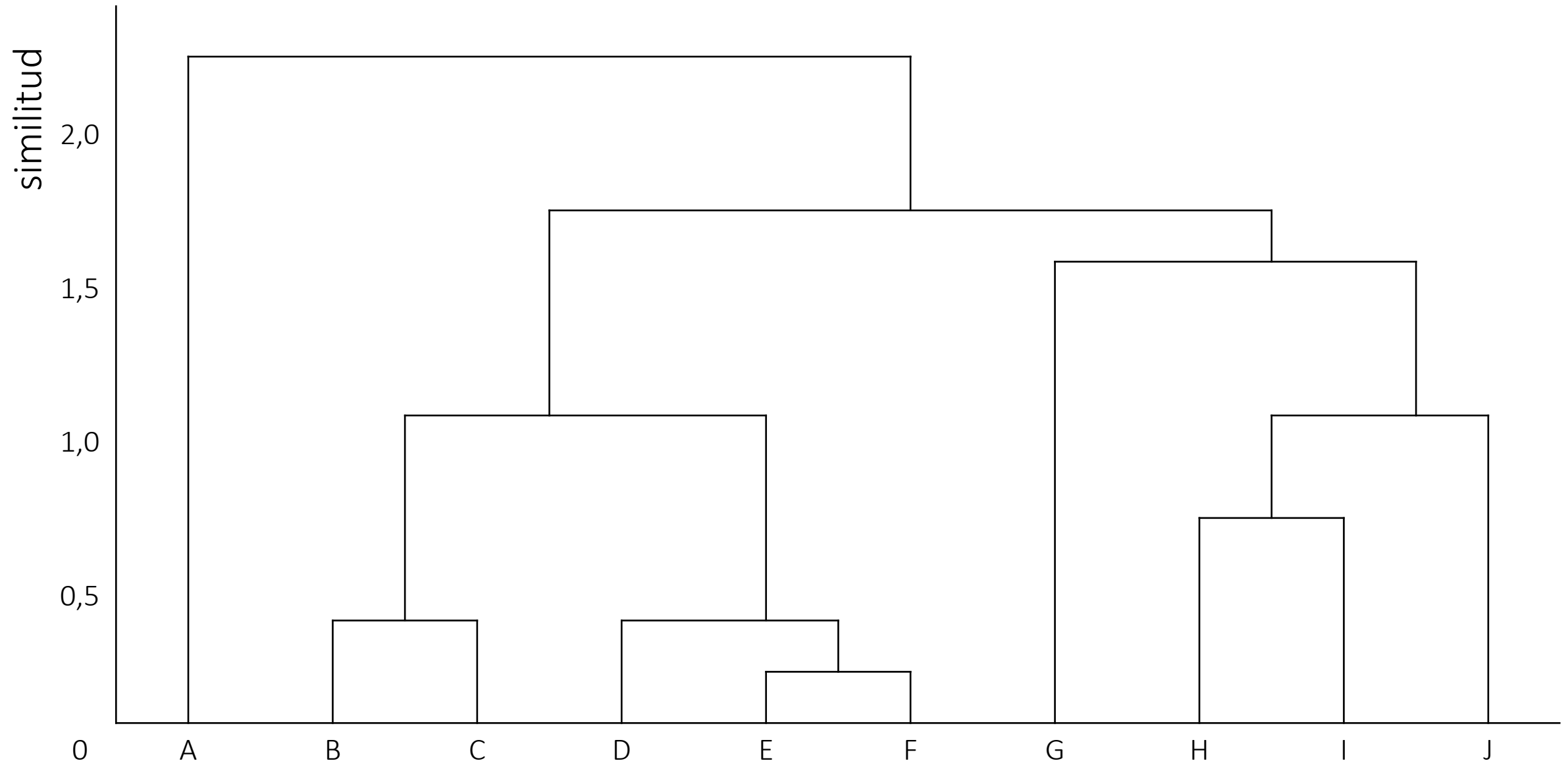
Inicialmente cada observación es su propio clúster

En cada iteración unimos dos clústeres de acuerdo a una métrica de similitud

Tras cada iteración actualizamos la métrica de similitud

El algoritmo termina cuando sólo queda un clúster con todas las observaciones

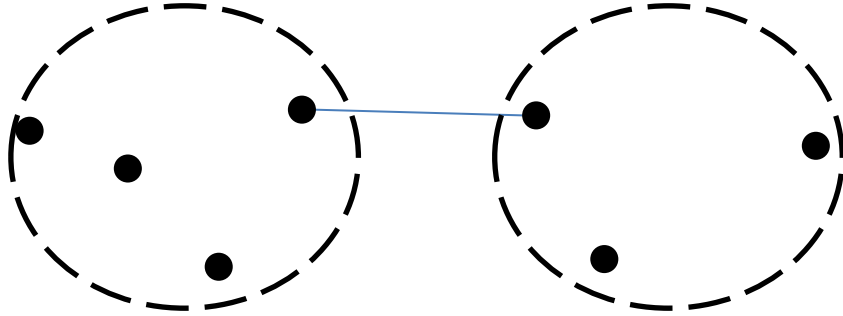
Dendrograma



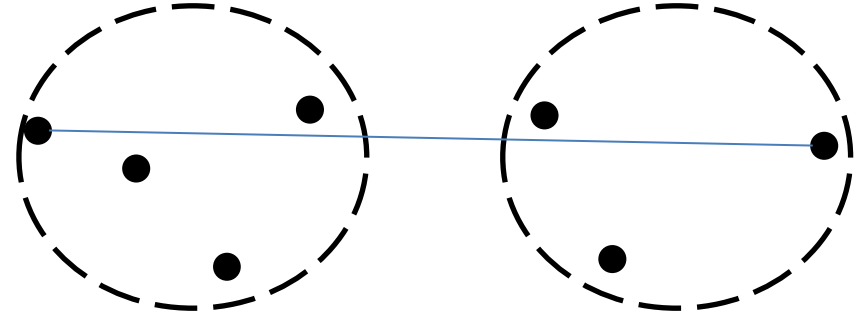
Métricas de similitud

Métricas de distancia

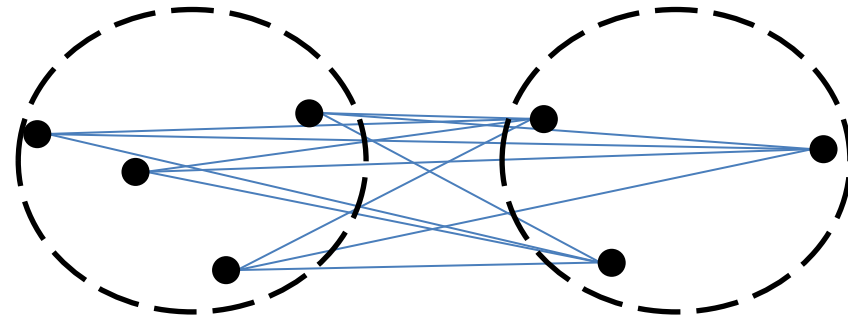
Mínimo



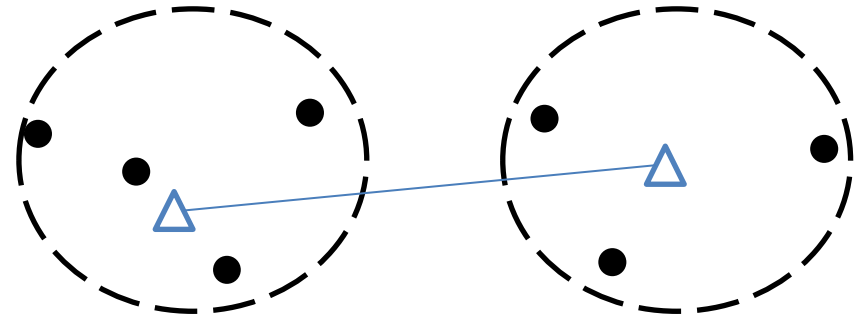
Máximo



Promedio



Centroides



Otras posibles métricas (correlación, desviación estándar, etc.)

Series de Tiempo

Series de Tiempo

Modelos autorregresivos (AR)

Son modelos lineales donde la variable de interés Y_t en cierto período t depende de los valores de sí misma en períodos anteriores

$$Y_t = \phi_1 \cdot Y_{t-1} + \phi_2 \cdot Y_{t-2} + \dots + \phi_p \cdot Y_{t-p} + \varepsilon_t$$

Series de Tiempo

Modelos de media móvil (MA)

Son modelos lineales donde la variable de interés Y_t en cierto período t depende de los errores de predicción en períodos anteriores

$$Y_t = \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \dots + \theta_q \cdot \varepsilon_{t-q} + \varepsilon_t$$

Series de Tiempo

Estacionalidad

El patrón de fluctuaciones periódicas de la serie de datos dentro de un período determinado (diariamente, semanalmente, mensualmente, anualmente, etc.)

La estacionalidad puede ser modelada y trabajada por separado, descomponiendo la serie de tiempo

Series de Tiempo

Estacionariedad

Es necesario que los datos no tengan tendencias y su media sea constante

Si los datos no son estacionarios, podría ser posible trabajar con una serie diferenciada

$$Z_t = Y_t - Y_{t-1}$$

Series de Tiempo

Modelos ARIMA(p,d,q)

ARIMA p indica el orden de la parte autorregresiva

AR**I**MA d indica la cantidad de diferenciaciones

ARIMA**A** q indica el orden de la parte de media móvil

Series de Tiempo

Ejemplo: ARIMA(3,1,2)

$$Z_t = Y_t - Y_{t-1}$$

$$Z_t = \phi_1 \cdot Z_{t-1} + \phi_2 \cdot Z_{t-2} + \phi_3 \cdot Z_{t-3} + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \varepsilon_t$$

Los parámetros del modelo se pueden estimar con algún método estadístico, como Máxima Verosimilitud

Minería de Texto

Minería de Texto

Las bases de datos con las que hemos trabajado se ven así:

Obs.	Var. 1	Var. 2	Var. 3	Var. 4	Var. 5	...
1	6	1	8	1	5	...
2	5	3	2	4	8	...
3	7	6	1	3	7	...
4	3	0	4	9	6	...
5	1	2	6	0	3	...
6	9	1	2	3	4	...
7	4	4	2	1	5	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Minería de Texto

Pero los documentos se ven así

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lantejas los viernes, algún palomino de añadidura los domingos, consumían las tres cuartas partes de su hacienda. El resto della concluían sayo de velarte, calzas de velludo para las fiestas, con sus pantuflos de lo mismo, y los días de entresemana se honraba con su vellorí de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y una sobrina que no llegaba a los veinte, y un mozo de campo y plaza, que así ensillaba el rocín como tomaba la podadera. Frisaba la edad de nuestro hidalgo con los cincuenta años; era de complexión recia, seco de carnes, enjuto de rostro, gran madrugador y amigo de la caza. Quieren decir que tenía el sobrenombre de Quijada, o Quesada, que en esto hay alguna diferencia en los autores que deste caso escriben; aunque, por conjeturas verosímiles, se deja entender que se llamaba Quejana. Pero esto importa poco a nuestro cuento; basta que en la narración dél no se salga un punto de la verdad.

¿Cómo representar documentos en forma numérica?

Corpus

Un corpus es un conjunto de documentos

Ejemplo:

Documento 1: “un auto rojo”

Documento 2: “un tomate rojo y un globo rojo”

Documento 3: “un plátano amarillo y un tomate verde”

Vocabulario

Un vocabulario es una secuencia ordenada de palabras con un identificador único

En nuestro ejemplo:

ID	Palabra
1	amarillo
2	auto
3	globo
4	plátano
5	rojo
6	tomate
7	un
8	verde
9	y

Bolsa de palabras

Podemos representar cada documento como una bolsa de palabras en forma matricial

										ID	Palabra
Documento 1:	“un auto rojo”									1	amarillo
Documento 2:	“un tomate rojo y un globo rojo”									2	auto
Documento 3:	“un plátano amarillo y un tomate verde”									3	globo
										4	plátano
										5	rojo
										6	tomate
										7	un
Doc.	1	2	3	4	5	6	7	8	9	8	verde
1	0	1	0	0	1	0	1	0	0	9	y
2	0	0	1	0	2	1	2	0	1		
3	1	0	0	1	0	1	2	1	1		

Minería de Texto

¿Qué podemos hacer con estos datos?

Aprendizaje supervisado (e.g. análisis de sentimientos)

Clasificación

Análisis no supervisado

Clustering

Minería de Texto

Limpieza y preprocesamiento de los datos

Mayúsculas y minúsculas

Pronombres, preposiciones y otras palabras vacías

Acentos y puntuación

Emojis

Stemming

Procesamiento de Imágenes

Procesamiento de Imágenes


Así como la unidad base de un texto son sus palabras, la unidad base de una imagen son sus píxeles

Cada píxel tiene básicamente dos tipos de información:


- Información espacial

- Información cromática


Color




R = 73
G = 86
B = 102



R = 124
G = 132
B = 149

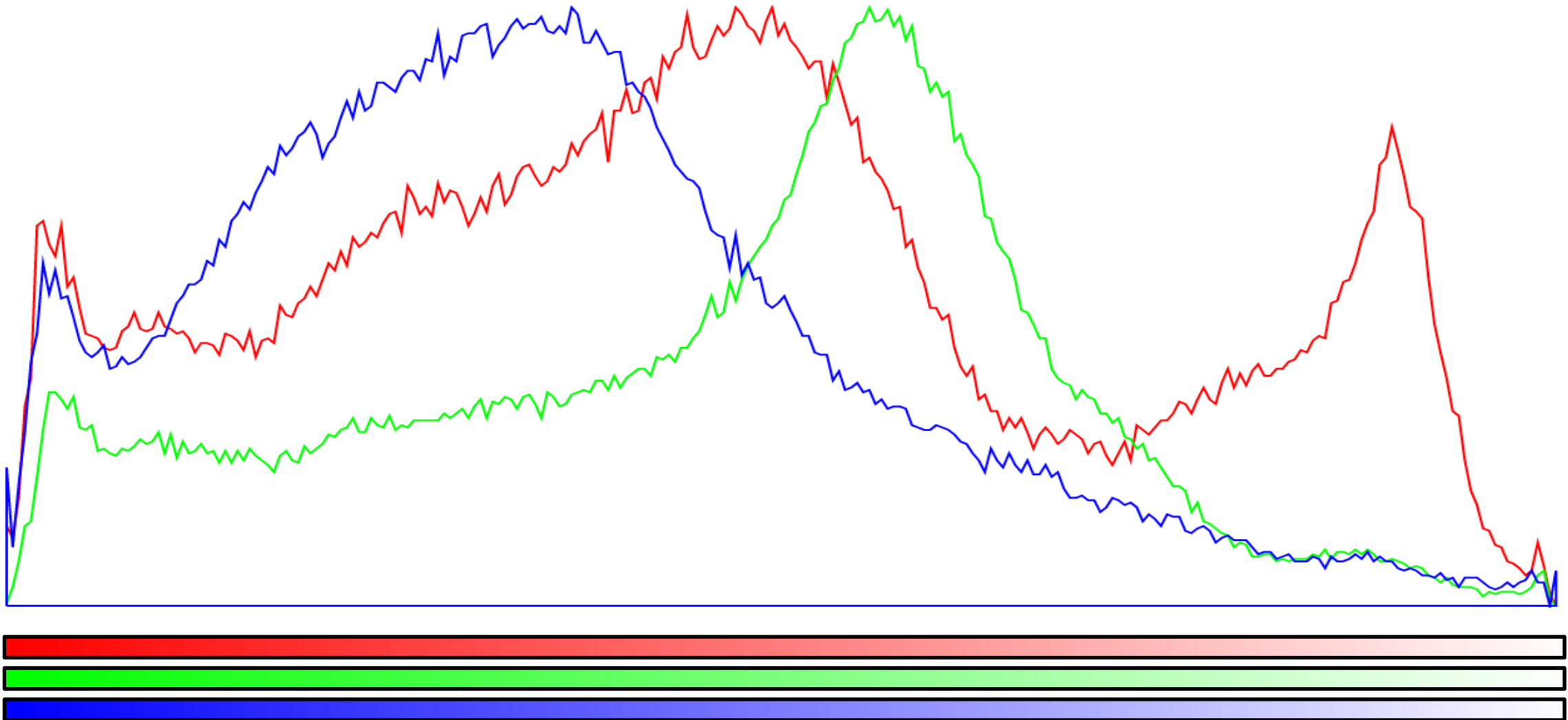


R = 223
G = 149
B = 112



R = 233
G = 238
B = 242

Histograma de colores



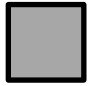
Intensidad



R = 84
G = 84
B = 84



R = 125
G = 125
B = 125

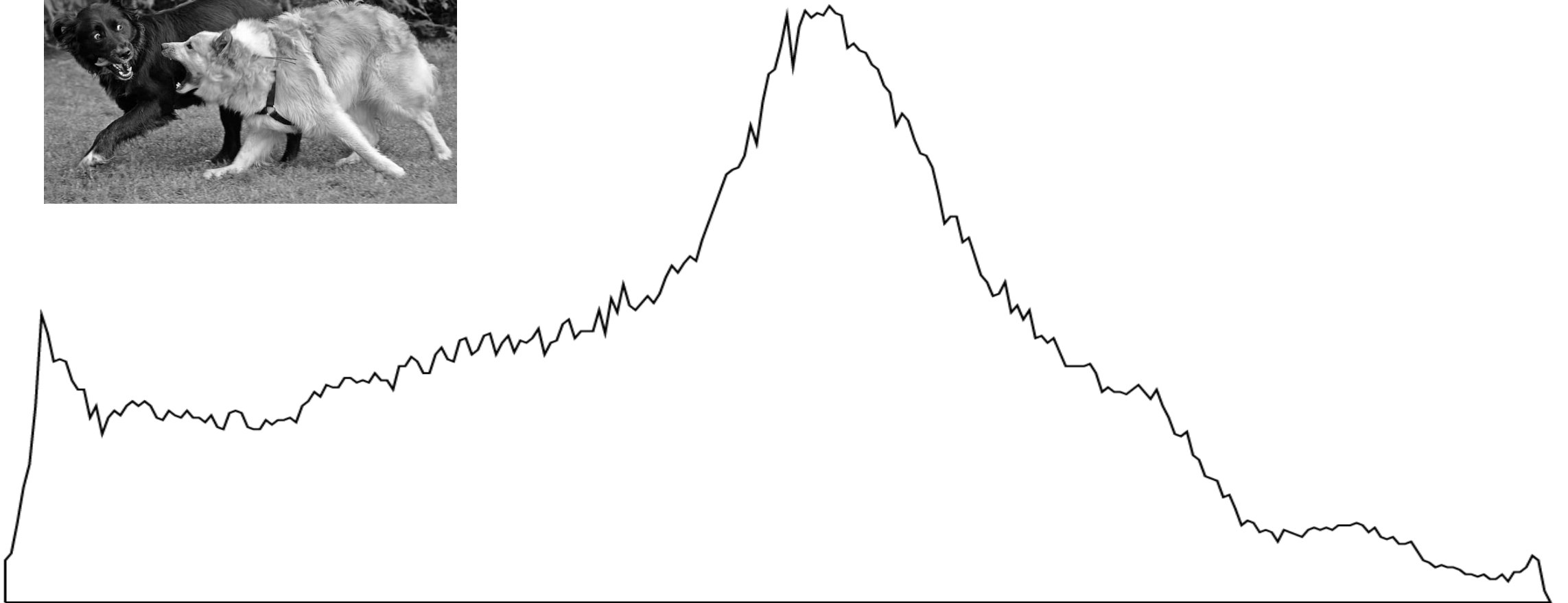


R = 167
G = 167
B = 167



R = 237
G = 237
B = 237

Histograma de intensidad



Procesamiento de Imágenes

Otros indicadores de textura basados en intensidad:

Bordes

Segundo momento angular

Contraste

Correlación

Entropía

Ejemplo: Clustering

$K = 2$



$K = 3$



$K = 4$



$K = 5$

