

Datos II

Datos, Pre-procesamiento

Bárbara Poblete

Calidad de los datos

- No poseen la calidad deseada a priori, los algoritmos de DM se enfocan en:
 1. Detección y corrección de problemas de calidad
 2. Usar algoritmos que toleren datos de poca calidad
- i.e., limpieza de datos

¿Por qué se producen errores?

Tipos de errores:

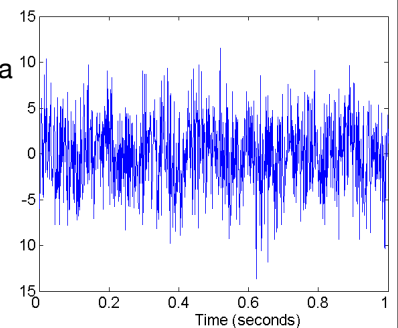
- Ruido y outliers
- Valores faltantes
- Datos duplicados
- Sesgo (Bias)



Artist
in rights obtainable from
inStock.com

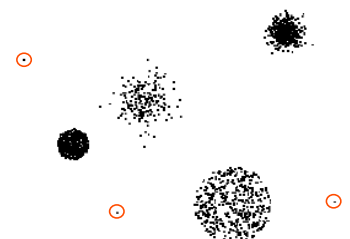
¿Qué es el ruido?

- Componente aleatoria en la medición (distorsión de voz en un teléfono malo)
- Datos espaciales, temporales



Outliers

- Objetos con características considerablemente diferentes a la mayoría



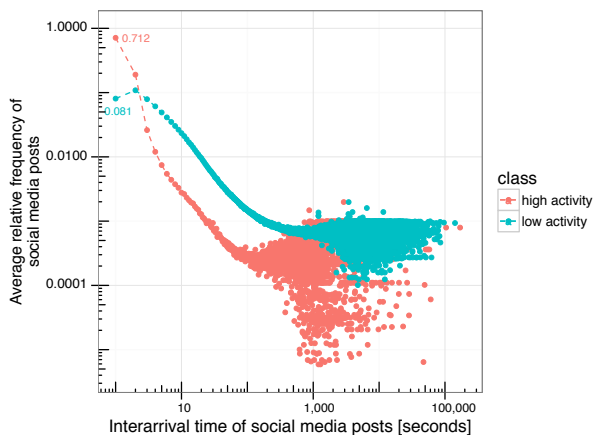
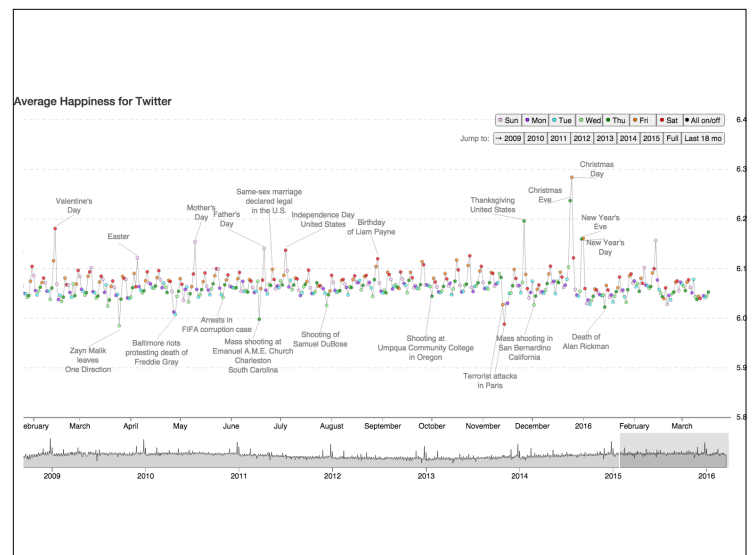


Pre-procesamiento de datos

- Creación de atributos
- Selección de un subconjunto de atributos
- Agregación
- Normalización
- Muestreo
- Reducción de dimensionalidad
- Discretización y binarización
- Transformación

Agregación de datos

- Combinar 2 o más atributos (o objetos) en un único atributo (o objeto)
- ¿Propósito?
 - Reducción de datos
 - Cambio de escala
 - Datos más estables



Agregación

Muestreo

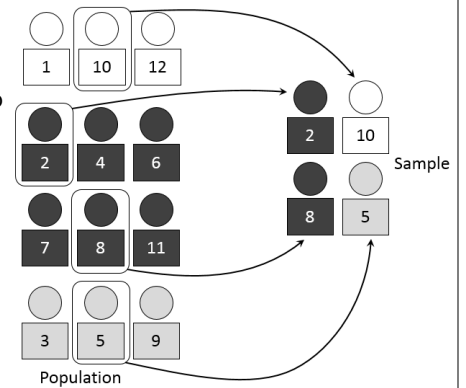
- Principal técnica de selección de datos (investigación preliminar o final)
- Usado en Estadística y DM
- ¿Cuándo es efectivo?

Tipos de Muestreo

- Muestreo aleatorio
- ¿Ventajas?
- ¿Desventajas?

Tipos de Muestreo

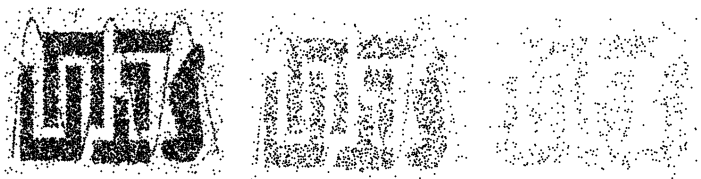
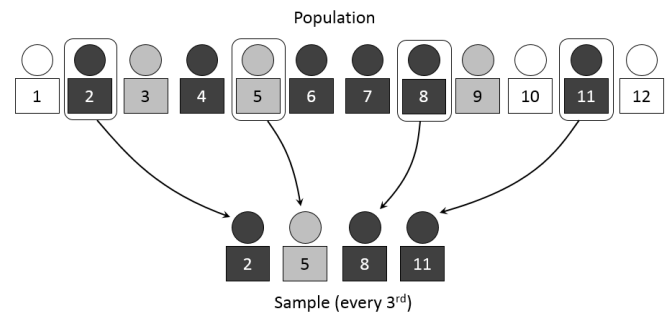
- Muestreo estratificado
- ¿Ventajas?
- ¿Desventajas?



Tipos de Muestreo

- Muestreo aleatorio
- Muestreo estratificado
- Muchos otros tipos de muestreo

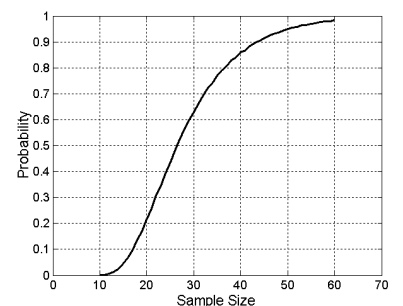
Systematic random sample



8000 points

2000 Points

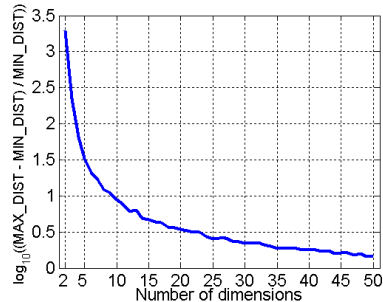
500 Points



- ¿Cómo obtener al menos un objeto de cada uno de los 10 grupos?

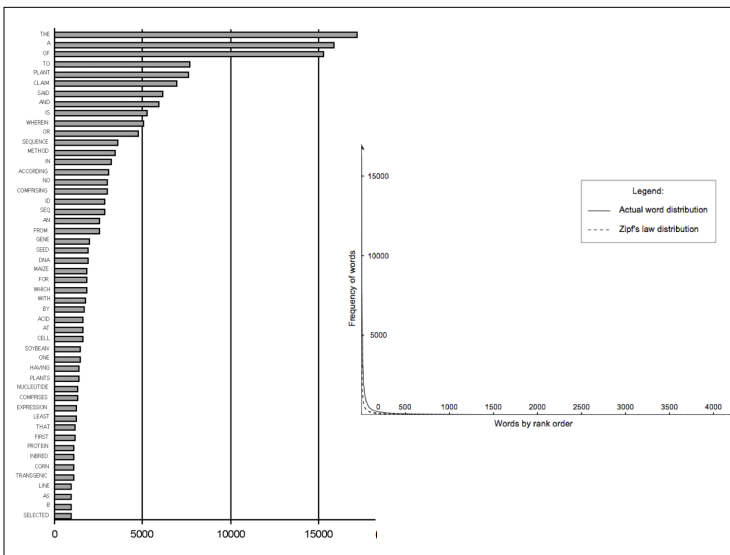
CURSE OF DIMENSIONALITY

- Al aumentar dimensionalidad, los datos se vuelven más dispersos en el espacio
- Pierden significado las medidas, i.e. densidad y distancia entre puntos (clustering y detección de outliers)



Reducción de Dimensionalidad

- ¿Propósito?
- Evitar curse of dimensionality
- Reducir costos asociados a aplicar el algoritmos (tiempo, memoria)
- Mejor visualización de los datos
- Ayuda a quitar atributos irrelevantes o ruidosos



Reducción de Dimensionalidad y Selección de atributos

- Sirven para lo mismo:
 - Selección de atributos
 - Reducción de dimensionalidad

Selección de atributos

- Fuerza bruta: trial and error muchas veces
- Missing Values Ratio
- Low Variance Filter
- High Correlation Filter
- <http://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html>

Reducción de Dimensionalidad

- Random Forest/Ensemble Trees
- Backwards/Forward Feature Elimination/Construction (pocas columnas)
- Técnicas de álgebra lineal: PCA, SVD, ISOMAP (nuevo espacio, pierde interpretabilidad)

Dimensionality Reduction	Reduction Rate	Accuracy on validation set	Best Threshold	AuC	Notes
Baseline	0%	73%	-	81%	Baseline models are using all input features
Missing Values Ratio	71%	76%	0.4	82%	-
Low Variance Filter	73%	82%	0.03	82%	Only for numerical columns
High Correlation Filter	74%	79%	0.2	82%	No correlation available between numerical and nominal columns
PCA	62%	74%	-	72%	Only for numerical columns
Random Forrest / Ensemble Trees	86%	76%	-	82%	-
Backward Feature Elimination + missing values ratio	99%	94%	-	78%	Backward Feature Elimination and Forward Feature Construction are prohibitively slow on high dimensional data sets. It becomes practical to use them, only if following other dimensionality reduction techniques, like here the one based on the number of missing values.
Forward Feature Construction + missing values ratio	91%	83%	-	63%	

<http://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html>

PCA

- Principal Component Analysis
- Para atributos continuos
- Busca un nuevo set de atributos (componentes principales) que
 1. Son combinaciones lineales de los atributos originales
 2. Son ortogonales (perpendiculares) entre sí
 3. Capturan la máxima variación de los datos

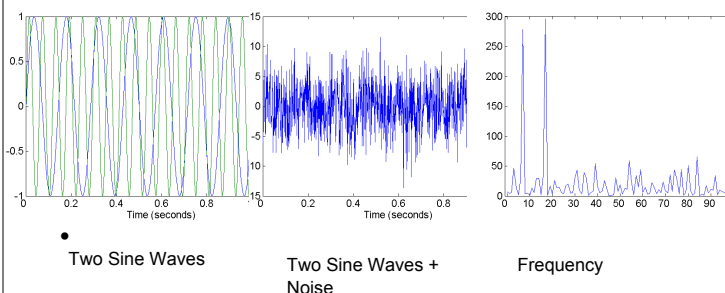
Dar pesos a los atributos

- Se asigna peso a los atributos según su importancia
 - SVM lo hace automáticamente
 - Normalización

Crear atributos

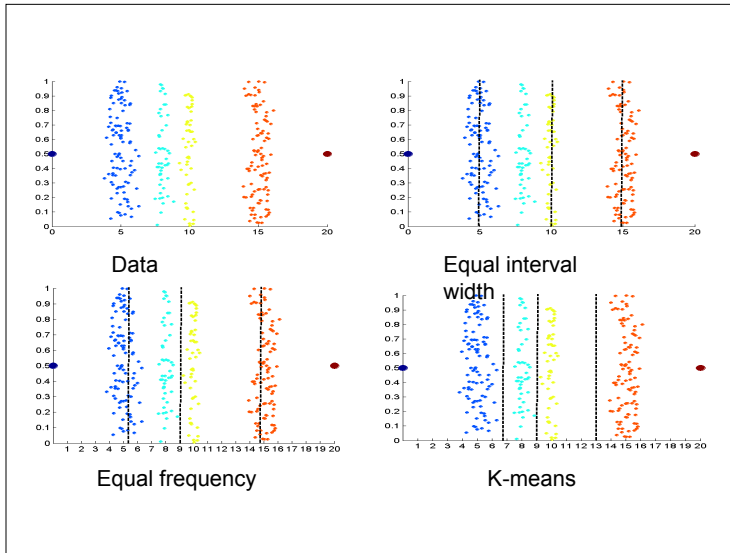
- Extraer atributos
- Mapear atributos a un nuevo espacio
- Construir atributos

Mapear a un nuevo espacio



Discretizar

- Decidir cuántas categorías tendremos
- Supervisado (con clases, considerando entropía y pureza)
- y no-supervisado (mismo intervalo, misma cantidad)



Transformación de atributos

- Una función que mapea el set de valores a otro set de datos.
- Funciones simples x^k , $\log(x)$, e^x , $|x|$
- Estandarización y Normalización

Próxima Clase

- Lab Exploración (2 sesiones: 1.1 y 1.2)
- Hito I (conformar grupos, elegir tema):
 - Exploración
 - (Pre-procesamiento)
 - Objetivos/Hipótesis iniciales (¿Por qué?)
 - primeros análisis

Proyectos de años anteriores...

- Análisis transporte en Santiago
- PIB por regiones
- Películas: Predecir ranking de las películas
- Música: Encontrar conjuntos de artistas similares para recomendación
- Recolección y análisis de datos de las elecciones presidenciales en Twitter
- Análisis de foro de u-cursos
- Análisis de datos de galaxias
- Proyecto Redes
- etc, etc...