

MINERÍA DE DATOS

CLASE 4

Datos, Pre-procesamiento

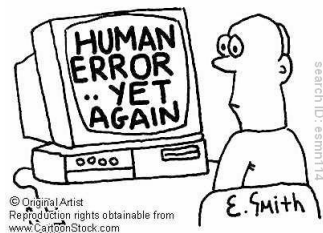
Bárbara Poblete

CALIDAD DE LOS DATOS

- No poseen la calidad deseada apriori
 1. Detección y corrección de problemas de calidad
 2. Usar algoritmos que toleren datos de poca calidad
- i.e., limpieza de datos

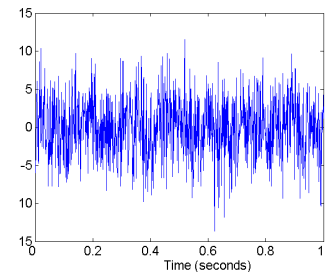
¿POR QUÉ SE PRODUCEN ERRORES?

- Ruido y outliers
- Valores faltantes
- Datos duplicados



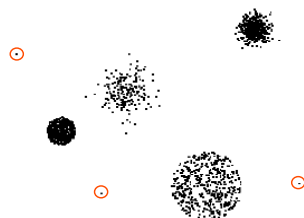
¿QUÉ ES EL RUIDO?

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



OUTLIERS

- Objetos con características considerablemente diferentes a la mayoría



VALORES FALTANTES

- ¿Motivos?
 - Información no recolectada (e.j: no quieren dar edad y/o peso)
 - Atributos no aplicables a todos (e.j: impuesto en niños)

VALORES FALTANTES...

- ¿Cómo los manejo?
- Eliminando el objeto
- Estimando (interpolando) valores
- Ignorar

VALORES INCONSISTENTES

- Datos mal ingresados

DATOS DUPLICADOS

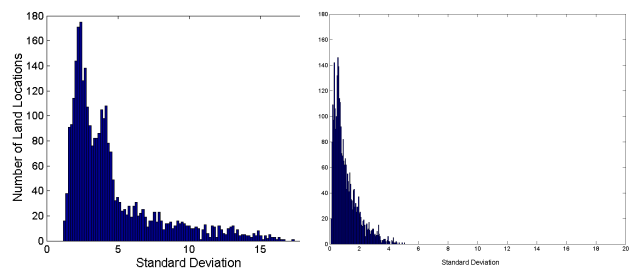
- El dataset incluye datos duplicados o cuasi-duplicados
- Gran problema al juntar datos de fuentes múltiples
- e.j: RT (casos deseados, no deseados)

PRE-PROCESAMIENTO

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

AGREGACIÓN

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



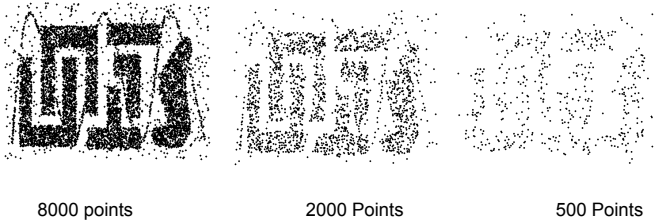
AGREGACIÓN

MUESTREO

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

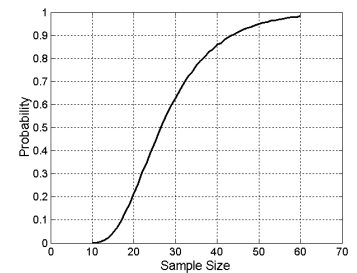
TIPOS DE MUESTREO

- Muestreo aleatorio
- Muestreo sin reposición
- Muestreo con reposición
- Muestreo estratificado



TAMAÑO DE LA MUESTRA

¿Cuál es el mejor?

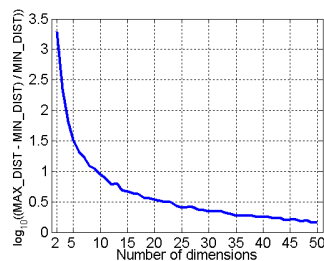


TAMAÑO DE LA MUESTRA

¿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 que ocupan
- Pierden significatividad 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 or ruidosos
- Técnicas de álgebra lineal: PCA, SVD, ISOMAP

SELECCIÓN DE SUBCONJUNTO DE ATRIBUTOS

- Eliminar atributos redundantes o irrelevantes
 - Enfoque de fuerza bruta
 - Enfoques embebidos
 - Enfoques de filtrado
 - Enfoques encapsulados

DAR PESO A LOS ATRIBUTOS

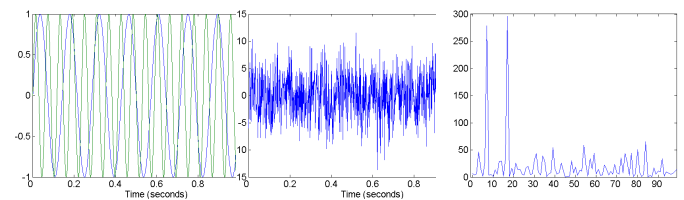
- Se asigna peso a los atributos según su importancia
 - SVM lo hace automáticamente
 - Normalización

CREAR ATRIBUTOS

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

MAPPEAR A UN NUEVO ESPACIO

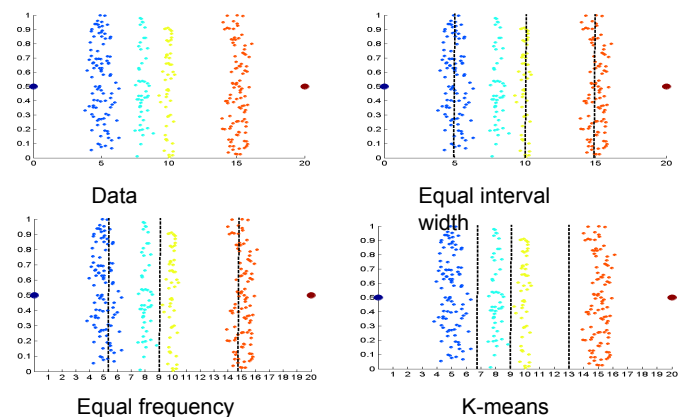
- Aplicar transformada de Fourier



• Two Sine Waves Two Sine Waves + Noise Frequency

DISCRETIZACIÓN

- 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 SEMANA

- Habrá repaso para Pregrado de lo visto hasta ahora (2 Clases).
- Postgrado: Estudiar capítulo 3 del libro (Exploración de Datos: repaso Estadística y Visualización)
- Reporte 1 página (mínimo) sobre alguna(s) aplicación(es) interesante(s) de visualización que tenga(n) que ver con DM (por u-cursos, zip del fuente html para posterior publicación en el blog)