Introducción a la minería de datos

Datos

Atributo, observaciones, características de algo, comúnmente sin ningún significado

Información

Datos procesados

¿Qué es la minería de datos?

Extracción no trivial de información implícita previamente desconocida y potencialmente útil de los datos

Exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos para descubrir patrones significativos

Métodos de predicción

Utilice algunas variables para predecir valores desconocidos o futuros de otras variables.

Métodos de descripción

Encontrar patrones interpretables por humanos que describan los datos

Definición de agrupamiento

Dado un conjunto de puntos de datos, cada uno con un conjunto de atributos y una medida de similitud entre ellos, encuentre conglomerados tales que:

- Los puntos de datos de un conglomerado sean más similares entre sí.
- Los puntos de datos de conglomerados separados sean menos similares entre sí.

Medidas de similitud:

- Distancia euclidiana si los atributos son continuos.
- Otras medidas específicas del problema.

Regresión

Predecir un valor de una variable continua dada basándose en los valores de otras variables, asumiendo un modelo de dependencia lineal o no lineal.

Desafíos de la minería de datos

- Escalabilidad
- Dimensionalidad
- Datos complejos y heterogéneos
- Calidad de los datos
- Propiedad y distribución de los datos
- Preservación de la privacidad
- Transmisión de datos

Análisis descriptivo de datos

¿Qué son los datos?

Recopilación de objetos de datos y sus atributos

Un atributo es una propiedad o característica de un objeto

Se le conoce también como variable, campo, característica o rasgo

Valores de los atributos

Los valores de atributo son números o símbolos asignados a un atributo

Tipos de atributos

- **Nominal** Ejemplos: Números de identificación, color de ojos, códigos postales
- **Ordinal** Ejemplos: clasificaciones (p. ej., sabor de las papas fritas en una escala del 1 al 10), grados, estatura en {alto, mediano, bajo}
- **Intervalo** Ejemplos: temperaturas en Celsius o Fahrenheit.
- **Relación** Ejemplos: temperatura en Kelvin, longitud, tiempo, recuentos

Atributo discreto

Tiene solo un conjunto finito o infinito contable de valores

Ejemplos Codigos postales, recuentos o el conjunto de palabras en una colección de documentos

A menudo se representan como variables enteras.

Nota: los atributos **binarios** son un caso especial de atributos discretos

Atributo continuo

Tiene números reales como valores de atributos

Los atributos continuos normalmente se representan como variables de punto flotante.

Datos de registro

Datos que consisten en una colección de registros, cada uno de los cuales consta de un conjunto fijo de atributos

Matriz de datos

Datos de la transacción

Un tipo especial de datos de registro, donde
cada registro (transacción) implica un conjunto de artículos.

Ruido

El ruido se refiere a la modificación de los valores originales

Valores atípicos

Los valores atípicos son objetos de datos con características que son considerablemente diferentes a las de la mayoría de los demás objetos de datos en el conjunto de datos

Valores faltantes

No se recopila la información

Datos duplicados

▪ **Problema importante al fusionar datos de fuentes heterogéneas.**

La misma persona con varias direcciones de correo electrónico.

Limpieza de datos.

▪ Proceso para gestionar problemas de datos duplicados.

¿Qué es la exploración de datos?

Una exploración preliminar de los datos para comprender mejor sus características.

• Relacionado con el área del **Análisis Exploratorio de Datos (EDA)**

Creado por el estadístico John Tukey

EDA

▪ El enfoque se centró en la visualización

▪ La agrupación y la detección de anomalías se consideraron técnicas exploratorias

Resumen estadístico

Las estadísticas de resumen **son números que resumen las propiedades de los datos**

Las propiedades resumidas incluyen frecuencia, ubicación y propagación.

Frecuencia

La frecuencia de un valor de atributo es el **porcentaje de tiempo que el valor aparece en el conjunto de datos**

Medidas de ubicación: media y mediana

la media es muy sensible a los valores atípicos.

Medidas de dispersión: rango y varianza

El rango es la diferencia entre el máximo y el min

La **varianza o desviación estándar** es la medida más común de la dispersión de un conjunto de puntos.

Visualización

La visualización es la conversión de datos a un formato visual o tabular para que se puedan analizar o informar las características de los datos y las relaciones entre los elementos o atributos de los datos

es una de las técnicas más poderosas y atractivas para la exploración de datos.

Representación

Es el mapeo de la información a un formato visual

Acuerdo

Es la colocación de elementos visuales dentro de una exhibición.

Selección

Es la eliminación o la desestimación de ciertos objetos y atributos.

Técnicas de visualización:

histogramas

▪ Generalmente muestra la distribución de valores de una sola variable.

▪ Divide los valores en intervalos y muestra un gráfico de barras con el número de objetos en cada intervalo.

Histogramas bidimensionales

Mostrar la distribución conjunta de los valores de dos atributos

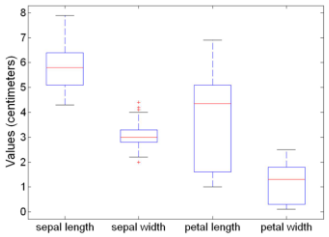
Ejemplo: ancho del pétalo y largo del pétalo

Box Plots

▪ Inventados por J. Tukey

▪ Otra forma de mostrar la distribución de datos

se pueden utilizar para comparar atributos



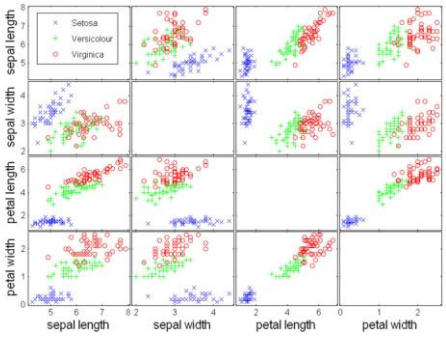
Diagramas de dispersión

▪ Los valores de los atributos determinan la posición.

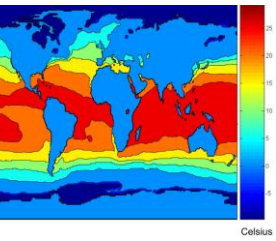
▪ Los diagramas de dispersión bidimensionales son los más comunes, pero pueden incluir diagramas de dispersión tridimensionales.

▪ A menudo, se pueden mostrar atributos adicionales utilizando el tamaño, la forma y el color de los marcadores que representan los objetos.

▪ Es útil tener matrices de diagramas de dispersión que puedan resumir de forma compacta las relaciones de varios pares de atributos.



• Gráficos de contornos



▪ **Útiles cuando se mide un atributo continuo en una cuadrícula espacial.**

▪ Dividen el plano en regiones con valores similares.

▪ Las curvas de nivel que delimitan estas regiones conectan puntos con valores iguales.

▪ El ejemplo más común son los mapas de curvas de nivel de elevación.

▪ También pueden mostrar temperatura, precipitaciones, presión atmosférica, etc.

Matrix Plots

Se puede graficar la matriz de datos.

▪ Esto puede ser útil cuando los objetos se ordenan por clase.

Coordenadas paralelas

▪ Se utilizan para representar gráficamente los valores de los atributos de datos de alta dimensión.

▪ En lugar de usar ejes perpendiculares, se utiliza un conjunto de ejes paralelos.

▪ Los valores de los atributos de cada objeto se representan como un punto en cada eje de coordenadas correspondiente, y los puntos se conectan mediante una línea.

▪ Por lo tanto, cada objeto se representa como una línea.

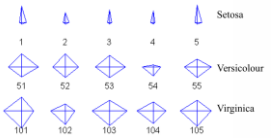
▪ A menudo, las líneas que representan una clase distinta de objetos se agrupan, al menos para algunos atributos.

▪ El orden de los atributos es importante para visualizar dichas agrupaciones.

Gráficos de estrellas

▪ Enfoque similar al de coordenadas paralelas, pero los ejes irradian desde un punto central.

▪ La línea que une los valores de un objeto es un polígono.

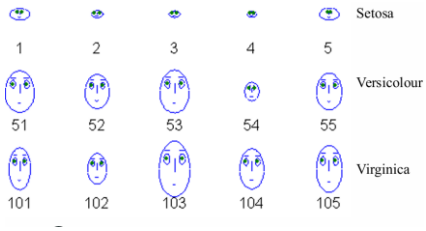


Chernoff Faces

▪ Enfoque creado por Herman Chernoff.

▪ Este enfoque asocia cada atributo con una característica de un rostro.

- Los valores de cada atributo determinan la apariencia de la característica facial correspondiente.
- Cada objeto se convierte en un rostro independiente.
- Se basa en la capacidad humana para distinguir rostros.



Aprendizaje Máquina o Automático (Machine learning)

Subcampo de las Ciencias de la computación y una rama de la Inteligencia Artificial , cuyo **objetivo es desarrollar técnicas que permitan que las computadoras aprendan.**

¿Qué es Python?

Es un **lenguaje de programación** con el cual se escriben programas de cómputo.

Es un programa de cómputo (término técnico "intérprete") el cual ejecuta programas en lenguaje Python

Esos programas se almacenan en archivos de texto con terminación .py

Módulo Numpy

Proporciona estructuras de datos y algoritmos necesarios para aplicaciones científicas que **involucren datos numéricos**

Módulo Pandas

Proporciona un estructuras y funciones de alto nivel diseñadas para hacer el trabajo con datos **estructurados o tabulares más fácil, rápido, y expresivo.**

Módulo Matplotlib

Biblioteca más popular para producir gráficos y visualizaciones de datos en dos dimensiones.

Módulo SciPy

Colección de paquetes para manejar un conjunto de problemas de dominios estándar en el cómputo científico

Módulo Scikit-learn

Conjunto de herramientas para aprendizaje máquina

DataFrames

Un DataFrame representa una **tabla rectangular de datos y contiene una colección ordenada de columnas**, cada una puede tener diferente tipo (numérico, cadena de texto, booleano, etc.).

Inducción del Árbol

Estrategia Greedy

Dividir los registros basados en una prueba de atributos que optimice cierto criterio.

Issues

- Determinar cómo dividir los registros.
  - ¿Cómo especificar la condición de prueba del atributo?
  - ¿Cómo determinar la major división (split)?
- Determinar cuándo parar de dividir (stop splitting).

¿Cómo determinar la mejor división o separación (split)?

- Enfoque Greedy:
  - Son preferidos nodos con distribución de clases homogéneas
- Es necesaria una medida de **Impuridad**:

C0: 5 C1: 5	C0: 9 C1: 1
No homogénea	Homogénea
Alto grado de impuridad	Bajo grado de imuridad

¿Me pueden dar un ejemplo?

Medidas de impuridad de un Nodo

• Índice de Gini

• **Entropía**

Examplos para el cálculo de la Entropía

$$Entropy(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

• Error de malclasificación

Criterio de paro para la inducción del Árbol

• Dejar de expandir un nodo cuando todos los registros pertenecen a la misma clase.

• Dejar de expandir un nodo cuando todos los registros tengan valores de atributos similares

MATRIZ DE CONFUSION

Accuracy

		CLASE PREDICHA	
CLASE REAL		Clase=SI	Clase=NO
	Clase=SI	a (TP)	b (FN)
	Clase=NO	c (FP)	d (TN)

• Métrica más ampliamente usada:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Medidas alternativas

Dr. Angel Ju
ar

		CLASE PREDICHA	
		Clase=SI	Clase=NO
CLASE REAL	Clase=SI	a	b
	Clase=NO	c	d

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

Validación cruzada (cross validation)

Es una manera de predecir el ajuste de un modelo a un hipotético conjunto de datos de prueba cuando no disponemos del conjunto explícito de datos de prueba.

Clasificador Random Forest

• Algoritmo de clasificación supervisado.

• Puede ser usado para clasificación o para Regresión

• Es el más flexible y fácil de usar.

Clasificación con KNN

Calcular la distancia entre dos puntos:

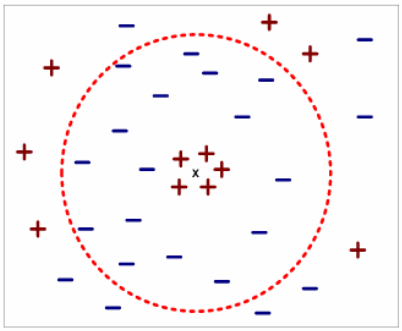
• Distancia Euclideana

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determinar la clase a partir de la lista de vecinos más cercanos.
  - Tomar la mayoría de votos de las etiquetas clase entre los k-vecinos más cercanos.
  - Ponderar el voto acorde a la distancia.
  - Factor de peso: $w = 1/d^2$

Clasificación con KNN

- Escogiendo el valor de k:
  - Si k es muy pequeño, KNN es sensible a valores ruidosos.
  - Si k is muy grande, La vecindad puede incluir puntos de otras clases.



Clasificador Bayesiano

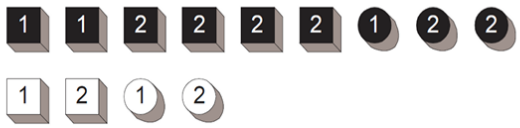
Una manera probabilística para resolver problemas de clasificación

- Teorema de Bayes:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

angelsanchez@uv.mx

Ejemplo



- A cada objeto de la Figura le asignamos una probabilidad de 1/13.
- UNO es el conjunto de los elementos que tienen un uno y DOS el que tienen un dos, NEGRO todos objetos negros y BLANCO todos los blancos.
- ¿Cuál sería la probabilidad de obtener un 1 dado que la figura sea negra?. Mediante el teorema de Bayes tenenemos:
- $P(UNO|NEGRO) = \frac{P(NEGRO|UNO)P(UNO)}{P(NEGRO)}$
- $P(UNO|NEGRO) = \frac{\left(\frac{3}{5}\right)\left(\frac{5}{13}\right)}{\left(\frac{9}{13}\right)} = \left(\frac{1}{3}\right)$

angelsanche

Ejemplo del teorema de Bayes

- Dado:
  - Un doctor sabe que la meningitis causa rigidez en el cuello el 50% de las veces
  - La probabilidad a priori de que un paciente tenga le de meningitis (M) es 1/50,000
  - La probabilidad a priori de que algún paciente tenga el cuello rígido (S) es 1/20
- Si un paciente tiene rigidez en el cuello, ¿Cuál es la probabilidad de que él/ella tenga meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

angesanchez@uv.mx

Usando el Teorema de Bayes para Clasificar

Enfoque:

- Calcular la probabilidad a posteriori P(Y | X_1, X_2, …, X_d) usando el teorema de Bayes

$$P(Y \mid X_1 X_2 \ldots X_n) = \frac{P(X_1 X_2 \ldots X_d \mid Y) P(Y)}{P(X_1 X_2 \ldots X_d)}$$

- *Maximum a-posteriori*: Escoger Y que maximice:
  P(Y | X_1, X_2, …, X_d)

- Equivalente a escoger el valor de Y que maximice
  P(X_1, X_2, …, X_d|Y) P(Y)

¿Cómo estimar P(X_1, X_2, …, X_d | Y)?

Ejemplo Naive Bayes (Han y Kamber, 2006)

P(compra computadora="sí") = 9/14=0.643
P(edad="<=30"|compra computadora="sí") = 2/9 = 0.222
P(ingreso="medio"|compra computadora="sí") = 4/9 = 0.444
P(estudiante="sí"|compra computadora="sí") = 6/9 = 0.667
P(crédito="suficiente"|compra computadora="sí") = 6/9 = 0.667

P(compra computadora="no") = 5/14=0.357
P(edad="<=30"|compra computadora="no") = 3/5 = 0.600
P(ingreso="medio"|compra computadora="no") = 2/5 = 0.400
P(estudiante="sí"|compra computadora="no") = 1/5 = 0.200
P(crédito="suficiente"|compra computadora="no") = 2/5 = 0.400

Ahora se deben computar dos operaciones:
P(compra computadora="sí" | edad="<=30", ingreso="medio", estudiante="sí", crédito="suficiente")
P(compra computadora="no" | edad="<=30", ingreso="medio", estudiante="sí", crédito="suficiente")

Dr. Angel Juan Sánchez García
angesanchez@uv.mx

Ejemplo Naive Bayes (Han y Kamber, 2006)

Utilizando el teorema de Bayes se tiene lo siguiente:

$$P(c|E) = \frac{P(E|c)P(c)}{P(E)}$$

P(E | compra computadora="sí") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(E | compra computadora="sí") P(compra computadora = "sí") = 0.044 x 0.643 = 0.028

P(E | compra computadora="no") = 0.600 x 0.400 x 0.200 x 0.400 = 0.019
P(E | compra computadora="no") P(compra computadora = "no") = 0.019 x 0.357 = 0.007

Dado estos resultados, la respuesta es "sí", pues maximizó la probabilidad.