Minería de Datos

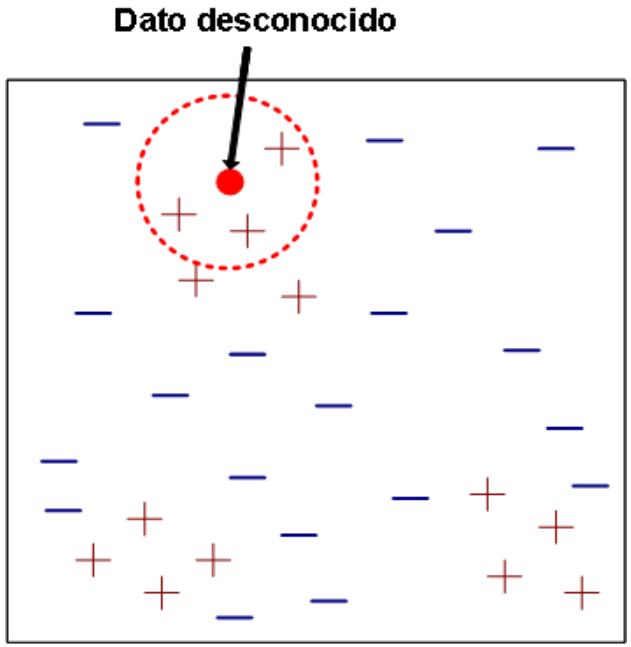
Técnicas de Minería de datos – K-vecinos más cercanos

Agenda:

- **Algoritmo K-vecinos**
- **Ejemplos de aplicación**

Algoritmo k-vecinos

- **Idea básica: Encontrar los K ejemplos de entrenamiento que son más similares al ejemplo test**
- **Estos ejemplos = vecinos más próximos**
- **Se utilizan para determinar la clase del ejemplo test**



Requiere:

- **Un conjunto de registros almacenados**
- **Una medida de distancia o similitud**
- **El valor de K, el número de vecinos a recuperar**

Algoritmo k-vecinos

Algunas variaciones de K-NN

- **Regla K-NN con rechazo:**

  La clasificación sólo se realiza en el caso de que alguna de las clases reciba un número de votos mayor a un umbral pre-establecido.

- **Regla K-NN por distancia media:**

  A partir de los K vecinos más próximos, a un nuevo caso a clasificar se le asigna la clase con distancia media menor.

- **Clasificador de distancia mínima:**

  Se selecciona un representante o prototipo por clase. Luego, para clasificar un dato, se le asigna la clase del representante más próximo.

Además, se puede utilizar el pesado o ponderación de atributos en las medidas de distancias o similitud; aquellas variables con pesos más grandes tendrán más influencia en el resultado final

Algoritmo k-vecinos

Para resumir:

- **Utiliza instancias de entrenamiento para hacer predicciones, sin tener que mantener un modelo derivado a partir de datos.**
- **Requiere de una medida de similitud o distancia y una función de clasificación que retorna la clase o valor predicho para una nueva instancia.**
- **La clasificación de un ejemplo test puede ser costoso computacionalmente, debido a la necesidad de calcular los valores de proximidad del ejemplo y cada instancia del conjunto de entrenamiento.**
- **Realizan sus predicciones basados en información local.**
- **Es importante seleccionar un buen valor de K**

Ejemplo: **Clasificación de documentos de TEG**

Objetivo: ***Construir un modelo que permita realizar una categorización de documentos de TEG según las Opciones Profesionales***

Tarea de minería de datos*: *Clasificación

Recolección de los datos:

Áreas	Cant. de docs. digitales	Cant. de docs. en físico	Cant. de docs. por área
Aplicaciones de Tecnología Internet (ATI)	24	26	50
Tecnología en Comunicaciones y Redes de Computadoras (Redes)	20	29	49
Bases de Datos (BD)	9	41	50
Inteligencia Artificial (IA)	7	42	49
Total	60	138	**N = 198**
N = cantidad de documentos de la colección (D)			

Aplicaciones

Preparación de los datos:

- **Eliminación de signos de puntuación y demás caracteres especiales. Los acentos también fueron removidos, para facilitar el análisis de los textos.**
- **Construcción un diccionario de palabras frecuentes en el español que no aportan información para la tarea de clasificación de textos (artículos, adjetivos, pronombres, entre otros).**
- **Estas palabras fueron eliminadas de los documentos aplicando un proceso de comparación con el diccionario.**
- **Luego se realizó el proceso de lematización (*stemming*) sobre el resto, mediante la aplicación del algoritmo *Porter Stemming* para el español.**
- **Como resultado se obtuvo un total de 3.747 raíces informativas a partir de los documentos recopilados**

Aplicaciones

a) Indexación:

- **Se utilizó la representación mediante el modelo de espacio vectorial, calculando el peso a_{ij} del término j en el documento i**

	t_1	…	t_j	…	T_{3747}
d_1	a_{11}	…	a_{1j}	…	$a_{1\ 3747}$
…	…	…	…	…	…
d_i	a_{i1}	…	a_{ij}	…	$a_{i\ 3747}$
…	…	…	…	…	…
d_{198}	$a_{198\ 1}$	…	$a_{198\ j}$	…	$a_{198\ 3747}$

Algunas opciones para obtener a_{ij}

- **Frecuencia del término** → $a_{ij} = f_{ij}$ ***Frecuencia del término j en el documento i***

- **Frecuencia relativa** → $a_{ij} = f_{ij} * \log(N/n_j)$ ***n_j = número total de veces que el término j aparece en la colección (D).***

b) Selección de variables (reducción de la dimensionalidad):

- **Se utilizaron diferentes técnicas de selección de atributos con el fin de identificar las variables más informativas para la tarea de clasificación**
- **Para seleccionar los atributos se utilizó un esquema de votación simple, por mayoría**

Resultado de la fase de preparación de datos:

➡ **Vista minable =** ***Tabla atributo-valor con 198 registros (filas) y 42 variables: 41 términos y la clase***

Aplicaciones

Extracto de la tabla de características:

Criterio	Ltc	Entropía	Tfc	Ranking
Mayoría	desarroll	web	desarroll	1
	web	siti	web	2
	siti	javascript	aplic	3
	clasificacion	clasificacion	siti	4
	dat	dat	clasificacion	5
	red	red	red	6
	cre	ruby	servici	7
	servici	administr	ruby	8
	ruby	usabil	prototip	9
	prototip	prototip	manej	10
	manej	manej	metod	11
	metod	metod	disposit	12
	control	disposit	protocol	13
	disposit	distribu	human	14
	protocol	protocol	conoc	15
	human	human	segur	16
	conoc	conoc	fundament	17
	profesional	profesional	inteligent	18
	segur	segur	asoci	19
	oos	expert	inalambr	20
	fundament	domini	movil	21
	domini	inteligent	agent	22
	inteligent	inalambr	ontologi	23
	ipv6	regl	profesional	24

Aplicaciones

Minería de datos:

- **Tarea de minería de datos: Clasificación**
- **Lenguaje de representación: no es un requerimiento**
- **Algoritmo: k-vecinos más cercanos**
- **Medida de rendimiento: exactitud predictiva**
- **Técnica de evaluación: validación cruzada de 10 particiones**

Aplicaciones

- **Se realizaron varios experimentos aplicando el algoritmo k-NN, configurando sus parámetros a los siguientes valores:**

  - Valor de k: 1, 3, 5 y 7.

  - Medida de distancia: Chebyshev y Euclídea

Resultados:

kNN con K=5 y distancia Euclídea						
Clasificaciones correctas	163		82,3232%			
Clasificaciones Incorrectas	35		17,6768%			
Nro. Total Instancias	198		100,0000%			
Medidas de Rendimiento	Precisión		*Recall*			
Clase a = ATI	0,729		0,860			
Clase b = Redes	1,000		0,816			
Clase c = BD	0,695		0,820			
Clase d = IA	0,975		0,796			
Promedio	0,848		0,823			
Matriz de Confusión			Clasificador			
			a	b	c	d
	Reales	a	43	0	7	0
		b	4	40	5	0
		c	8	0	41	1
		d	4	0	6	39