

Agenda:

- **Clasificación con métodos basados en vecindad**
- **Algoritmo K-vecinos**
- **Ejemplos de aplicaciones**

Clasificación con métodos basados en vecindad

- La predicción se basa en la utilización de instancias o ejemplos “vecinos” al dato que hay que procesar.
- Idea: ante una nueva situación, se podría actuar como se hizo en situaciones anteriores similares, si éstas fueron exitosas.
- La similitud o distancia entre cada ejemplo y el dato a procesar es esencial en el proceso

Ejemplo:

En clasificación:

Asignar una clase a un nuevo dato, observando la clase de datos similares

En regresión:

El valor numérico predicho para un nuevo dato, se obtiene de los valores obtenidos para ejemplos similares

Clasificación con métodos basados en vecindad

Dos aspectos importantes:

- a) ¿Qué se entiende por similitud?
- b) ¿Cuándo se explota dicha similitud?

a) ¿Qué se entiende por similitud?

Similitud

- Es una medida numérica que indica el grado al cual dos objetos se parecen
- A más alto este valor más parecidos los objetos
- Es no negativa y generalmente entre 1 (similitud máxima) y 0 (no hay similitud)

Sin embargo, es común utilizar la **distancia** (inverso de la similitud), también conocida como **disimilitud**:

Clasificación con métodos basados en vecindad

Distancia

- Es una medida numérica que indica el grado al cual dos objetos son diferentes
- Mientras más bajo este valor, más parecidos
- Puede asumir valores entre $[0, 1]$ o entre $[0, \infty]$

Las medidas que satisfacen las siguientes tres propiedades

1. Positividad: $d(X, Y) \geq 0 \quad \forall X, Y$

$d(X, Y) = 0 \longrightarrow X = Y$



**Métricas o
distancias**

2. Simetría: $d(X, Y) = d(Y, X) \quad \forall X, Y$

3. Desigualdad triangular: $d(X, Z) \leq d(X, Y) + d(Y, Z) \quad \forall X, Y, Z$

Sin embargo, muchas medidas no satisfacen estas 3 propiedades pero son muy útiles
Para las medidas de similitud la desigualdad triangular generalmente no se cumple

Clasificación con métodos basados en vecindad

Algunas medidas de similitud:

- Para datos numéricos

$$\text{Cos}(X, Y) = \frac{X \cdot Y}{||X|| \cdot ||Y||}$$

Ejemplo:

Si $X = (2,1)$, $Y = (3,2)$, $Z = (5,1)$

$$\text{Cos}(X, Y) = \frac{(2 \times 3) + (1 \times 2)}{(2.23) \cdot (3.60)} = 0.99 \quad \leftarrow \quad \text{X, Y son más parecidos}$$

$$\text{Cos}(X, Z) = \frac{(2 \times 5) + (1 \times 1)}{(2.23) \cdot (5.09)} = 0.96$$

Clasificación con métodos basados en vecindad

- Para datos categóricos

$$\text{Similitud}(X, Y) = \frac{\text{comunes}}{\text{comunes} + \text{no_comunes}}$$

Ejemplo:

Si $X = (\text{Rojo}, \text{Alto}, \text{Maracay}, \text{Pequeño}, \text{Redondo})$

$Y = (\text{Rojo}, \text{Bajo}, \text{Maracay}, \text{Grande}, \text{Redondo})$

$Z = (\text{Verde}, \text{Bajo}, \text{Caracas}, \text{Grande}, \text{Cuadrado})$

$$\text{Similitud}(X, Y) = \frac{3}{3 + 2} = 0.6 \quad \leftarrow \text{X, Y son más parecidos}$$

$$\text{Similitud}(Z, Y) = \frac{2}{3 + 2} = 0.4$$

Clasificación con métodos basados en vecindad

Algunas medidas de distancia:

- Para datos numéricos

Sea $X = (x_1, x_2, x_3, \dots, x_d)$

$Y = (y_1, y_2, y_3, \dots, y_d)$

$$\text{Distancia Euclídea} = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$$

$$\text{Distancia Manhattan} = \sum_{i=1}^d |x_i - y_i|$$

$$\text{Distancia Canberra} = \sum_{i=1}^d \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

También se han definido medidas de similitud o de distancia para datos binarios, datos complejos como cadenas de caracteres, grafos, árboles, datos difusos, entre otros

Clasificación con métodos basados en vecindad

b) ¿Cuándo se explota esta similitud?

En el marco de clasificación mostrado hasta el momento

- ➡**
 - **Se realiza un paso inductivo para construir un modelo a partir de los datos.**
 - **Luego se aplica un paso deductivo para aplicar el modelo a los datos de test.**

—————→ ***Métodos anticipados***

Otro esquema:

- **Esperar a que se plantee una predicción sobre un nuevo dato.**
- **En este momento, determinar las instancias o casos más parecidos (similares) y utilizar estos datos para obtener una respuesta (clase)**

—————→ ***Métodos retardados o perezosos***

Clasificación con métodos basados en vecindad

☞ Métodos anticipados:

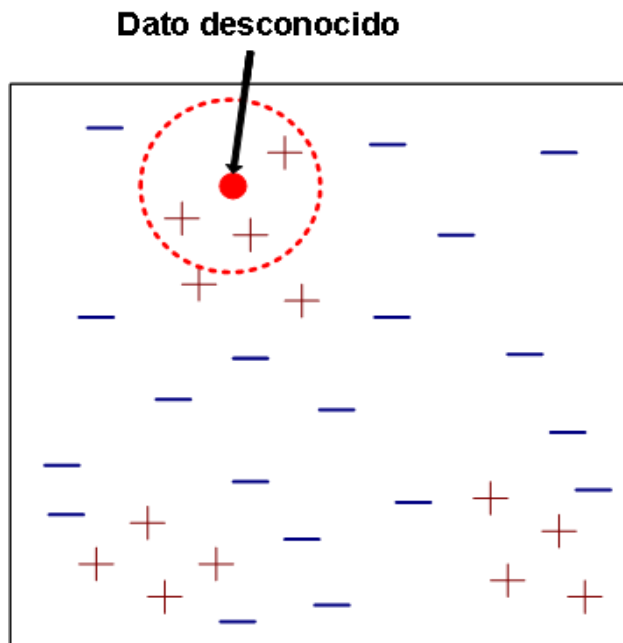
- **Construyen un modelo antes de realizar una tarea de predicción o generalización**
- **Se construye una aproximación global utilizando la totalidad del conjunto de datos**
- **Ejemplo: algoritmos de árboles de decisión**

☞ Métodos retardados o perezosos (lazy)

- **No construyen un modelo y retrasan la decisión de predicción hasta el instante en que se recibe un nuevo dato a procesar**
- **Realizan una aproximación local al dato a generalizar (hace predicciones basado en información local).**
- **Ejemplo: algoritmo k-vecinos**

Algoritmo K-vecinos

- **Idea básica:** Encontrar los K ejemplos o instancias del conjunto de aprendizaje que son más similares al nuevo dato a clasificar
- **Estos ejemplos = vecinos más próximos**
- **Se utilizan para determinar la clase de la nueva instancia**



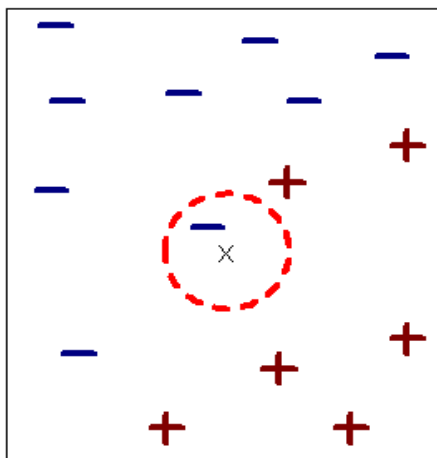
Requiere:

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

Para clasificar un nuevo dato:

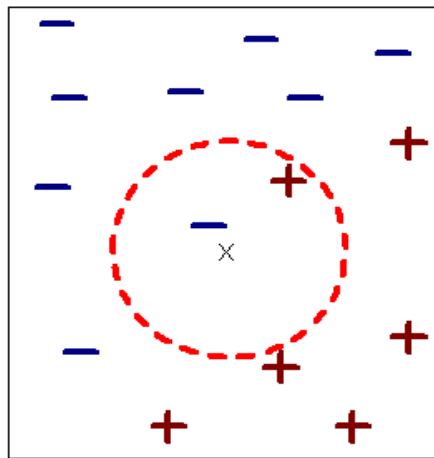
- Calcular la distancia a todos los datos del conjunto de aprendizaje
- Determinar los k vecinos (más parecidos)
- Utilizar la etiqueta de clase de los vecinos para determinar la clase del nuevo dato (por ejemplo, por mayoría)

Ejemplo:



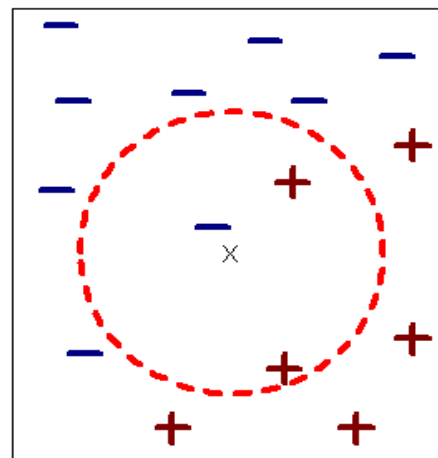
a) 1- vecino más próximo

Clase = —



b) 2- vecinos más próximos

Clase = ?



c) 3- vecinos más próximos

Clase = +

Algoritmo:

{Entrada: D (conjunto de entrenamiento), K (número de vecinos)}

Para cada ejemplo z_i

Para $j = 1$ hasta N

$d_i(z_i, X_j)$ = distancia entre z_i y el ejemplo de entrenamiento X_j

Fin_Para

D_z = conjunto de los K ejemplos más cercanos a z_i (lista de vecinos)

$y_i = \operatorname{argmax}_v \sum_{(x_k, y_k) \in D_z} I(v = y_k)$ % Clasifica el ejemplo de acuerdo a la clase mayoritaria de sus vecinos

Fin_Para

{Salida: conjunto de ejemplos test clasificados}

Donde,

$$y_i = \operatorname{argmax}_v \sum_{(x_k, y_k) \in D_z} I(v = y_k)$$

$$\text{Función indicadora} = \begin{cases} 1 & \text{si } v = y \\ 0 & \text{si } v \neq y \end{cases}$$

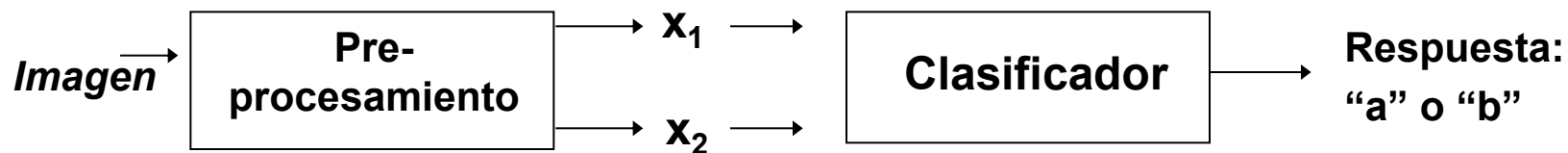
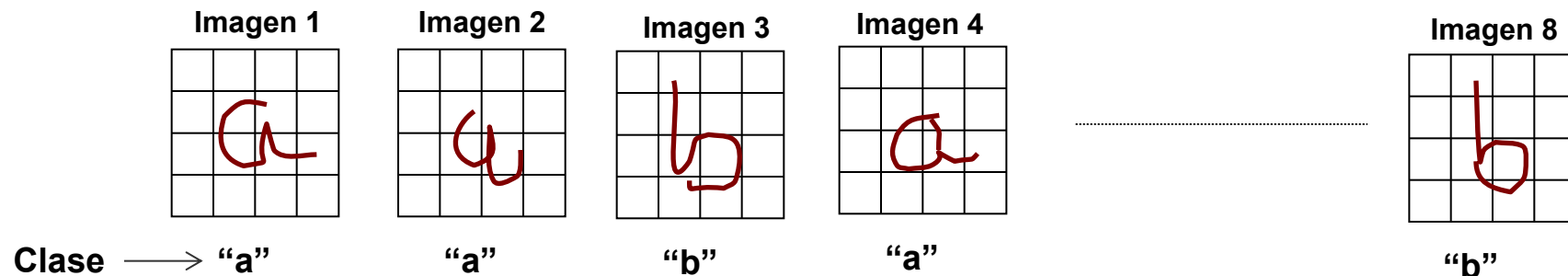
Etiqueta de clase

Algoritmo k-vecinos

Ejemplo: Reconocimiento de caracteres manuscritos a partir de imágenes

- Tarea: Clasificación
- Algoritmo: K-vecinos

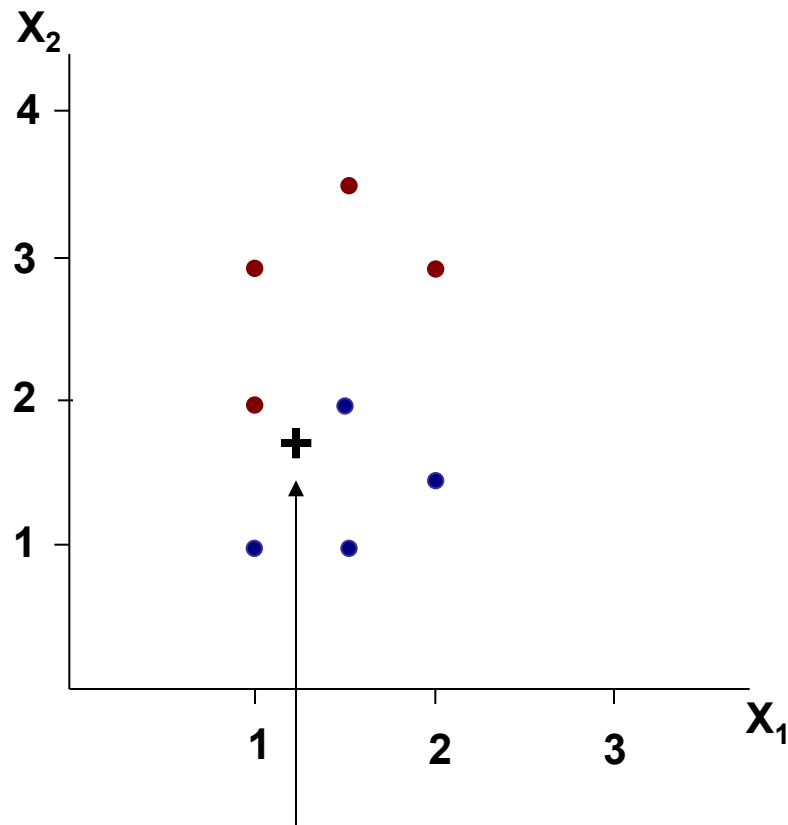
Conjunto de datos:



Donde: x_1 = ancho del caracter
 x_2 = alto del caracter

Vista minable:

X_1	X_2	CLASE
1.0	1.0	a
1.5	1.0	a
1.5	2.0	a
2.0	1.5	a
1.0	2.0	b
1.0	3.0	b
1.5	3.5	b
2.0	3.0	b



Nuevo dato: $z = (1.2, 1.8)$

¿Cómo se clasifica?

Si $K = 3$ y distancia Euclídea

1. Calcular la distancia de z a cada uno de instancias en el conjunto de datos D :

$$\begin{array}{ll} d(x^1, z) = 0.82 & d(x^5, z) = 0.28 \\ d(x^2, z) = 0.85 & d(x^6, z) = 1.21 \\ d(x^3, z) = 0.36 & d(x^7, z) = 1.72 \\ d(x^4, z) = 0.85 & d(x^8, z) = 1.44 \end{array}$$

2. Determinar la lista de los 3 vecinos más próximos:

$$\text{Vecinos}_z = D_z = \{(x^1, y_1), (x^3, y_3), (x^5, y_5)\}$$

3. Clasificar z de acuerdo a la clase mayoritaria:

$$\left\{ \begin{array}{l} y_1 = a \\ y_3 = a \\ y_5 = b \end{array} \right. \Rightarrow \begin{array}{l} \text{Votos para "a"} = 2 \\ \text{Votos para "b"} = 1 \end{array} \Rightarrow \text{Se asigna a } z \text{ la clase "a"}$$

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 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

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

Clasificación de documentos de TEG

- **Motivación principal = posibilidad de realizar una asignación automática de jurados o evaluadores, en función de los objetivos o temas abordados en los TEG.**
- **En muchos casos, debido a la interdisciplinariedad presente en estos documentos, esta tarea no es fácil de realizar por las comisiones designadas para tal fin, debido a que no se dispone de la experticia necesaria para decidir a cuál área u opción pertenece un TEG, lo cual es necesario para llevar a cabo una asignación adecuada de jurados**

Objetivo: Realizar, de manera automática, una categorización de documentos de TEG según las Opciones Profesionales

Tarea de minería de datos: Clasificación

Recolección de los datos:

El corpus o conjunto de datos está constituido por documentos de Trabajos Especiales de Grado de la Licenciatura de Computación de la Universidad Central de Venezuela, para un período de dos años

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)			

De los TEG sólo se consideraron el título, resumen (o en su defecto la introducción) y palabras claves.

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.**

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 $\rightarrow a_{ij} = f_{ij}$ *Frecuencia del término j en el documento i*
- Frecuencia relativa $\rightarrow a_{ij} = f_{ij} * \log(N/n_j)$ *n_j = número total de veces que el término j aparece en la colección (D).*
- Otras: entropía, tfc, ltc,

Ejemplos de aplicaciones

➔ *Tabla atributo – valor:*

	Inalambr	...	neuronal	...	manej	siti	Opción
d₁	0.295	...	0.000	...	0.001	0.000	Redes
d₂	0.000		0.000		0.002	0.135	ATI
d₃	0.000	...	0.318	...	0.000	0.002	IA
⋮	⋮		⋮		⋮	⋮	⋮
d₁₉₈	0.000	...	0.000	...	0.115	0.000	BD

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	No. de atributos seleccionados
Pesado por <i>tfc</i>	41
Pesado por <i>ltc</i>	41
Pesado por <i>entropía</i>	35

Minería de datos y evaluación:

- **Tarea de minería de datos: Clasificación**
- **Algoritmo: k-vecinos más cercanos**
- **Medida de rendimiento: exactitud predictiva**
- **Técnica de evaluación: validación cruzada de 10 particiones**

Ejemplos de 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: Euclídea

Distancia	K	tfc	ltc	entropía
Euclídea	1	78.78	78.28	80.81
	3	80.80	82.32	80.81
	5	77.27	81.31	76.76
	7	75.75	80.30	74.75
Euclídea ponderada	1	78.78	78.28	80.81
	3	80.80	81.81	79.79
	5	79.79	81.81	82.32
	7	78.78	80.80	80.81

$$\text{Distancia Euclídea} = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$$

$$\text{Euclídea ponderada} = \sqrt{\sum_{i=1}^d w_i (x_i - y_i)^2}$$

Ejemplos de aplicaciones

Donde:

$$a_{ij}(tfc) = \frac{f_{ij} \times \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{j=1..M} \left[f_{ij} \times \log\left(\frac{N}{n_j}\right) \right]^2}}$$

f_{ij} = frecuencia del término j en el documento i

n_j = número total de veces que el término j aparece en la colección (D)

N = cantidad de documentos en la colección (cardinalidad de D)

M = cantidad de términos en la colección (cardinalidad de T);

$$a_{ij}(ltc) = \frac{\log(f_{ij} + 1.0) \times \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{i=1..M} \left[\log(f_{ij} + 1.0) \times \log\left(\frac{N}{n_j}\right) \right]^2}}$$

$$a_{ij}(entropia) = \log(f_{ij} + 1.0) \times \left(1 + \frac{1}{\log(N)} \sum_{i=1..M} \frac{f_{ij}}{n_j} \times \log\left(\frac{f_{ij}}{n_j}\right) \right)$$

Ejemplos de aplicaciones

Resultados:

kNN con K=5 y distancia Euclídea ponderada como 1/distancia							
Clasificaciones correctas		163	82,3232%				
Clasificaciones Incorrectas		35	17,6768%				
Medidas de Rendimiento		Precisión		Sensibilidad			
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		

Sensibilidad o recall:

$$\frac{VP}{VP + FN}$$

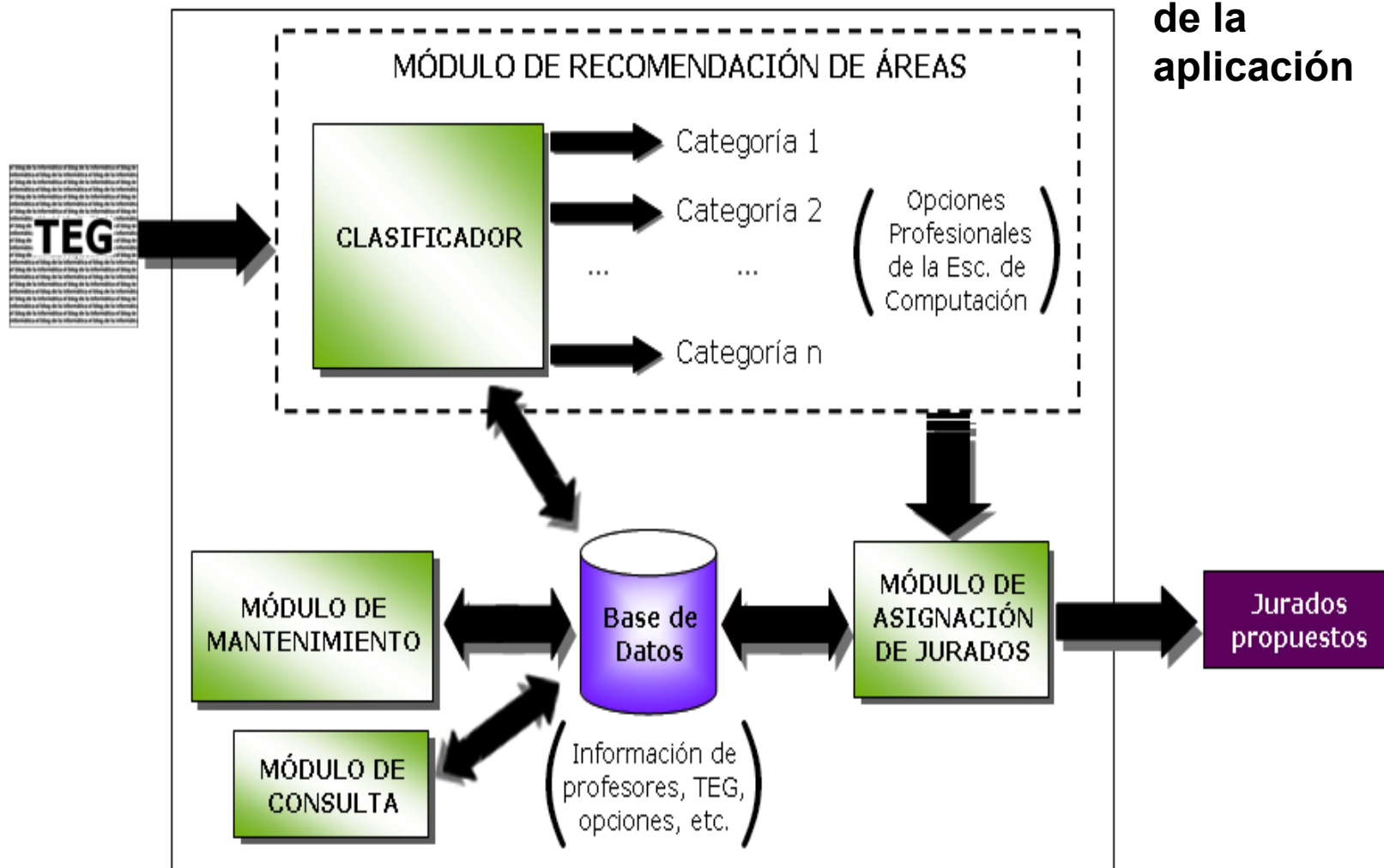
Precisión:

$$\frac{VP}{VP + FP}$$

Ejemplos de aplicaciones


Difusión y uso:

Arquitectura de la aplicación



Ejemplos de aplicaciones



Interfaz de la aplicación:



Sistema de Asignación de Jurados a TEG

[Inicio](#)
[Contáctenos](#)
[Créditos](#)

Caracas, 01 de mayo de 2011, 09:26 am


 Usuario: admin
 
 Cerrar Sesión

Menú de Opciones

- Asignar Jurado
- Consultar Jurado
- Cambiar Clave
- Cerrar Sesión

Administración

- Configuración General
- Opciones Usuarios
- Opciones Áreas
- Opciones Profesores

Inicio > Asignar Jurados - Paso 1 > Asignar Jurados - Paso 2

Áreas recomendadas

Área 1

 Área 2

¿Desea aceptar el área recomendada o desea editarla?

Universidad Central de Venezuela | Fac. de Ciencias | Esc. de Corte Asignación de Jurados a TEG

 Optimizado para los navegadores IE Explorer 8 y Firenada: 1024x768

Ejemplos de aplicaciones

Una vez desarrollada la aplicación, se clasificaron 23 documentos de TEG, pertenecientes a un corpus que fue coleccionado después de la construcción del clasificador.

Cada documento en este corpus estaba asociado a una opción profesional; sin embargo, con el apoyo de expertos, en algunos casos se asignó una segunda área considerando la posible interdisciplinariedad presente en estos trabajos.

Nro. TEG	Categorización proporcionada por los expertos	Recomendación realizada por la aplicación
1	ATI	ATI
2	REDES	REDES y BD
3	ATI	ATI
4	REDES	REDES
5	ATI y BD	BD y ATI
6	REDES	REDES
7	ATI y BD	ATI
8	ATI	ATI
9	REDES	REDES e IA
10	REDES	REDES
11	ATI	ATI
12	ATI	ATI y BD
13	ATI	ATI
14	IA	REDES
15	REDES	REDES
16	ATI	BD y ATI
17	BD	BD
18	BD y ATI	ATI y BD
19	ATI	ATI
20	ATI y BD	BD y ATI
21	ATI	ATI
22	ATI y BD	BD
23	IA	BD e IA

Ejemplos de aplicaciones

Se quiere construir un sistema para apoyar el diagnóstico de cáncer de mamá basado en una plataforma de telemedicina. El usuario se podrá comunicar con un módulo experto a través de una interfaz Web, mediante la cual podrá introducir las características que observa en las células del tejido. El módulo experto le enviará como respuesta si la muestra es benigna o maligna. Como se dispone de registros médicos asociados a pacientes, se quiere utilizar la minería de datos para construir el modelo de diagnóstico.

Aplique el proceso de minería de datos para construir este modelo de clasificación utilizando el conjunto de datos BREAST-CANCER-WISCONSIN, del repositorio UCI. Utilice tres técnicas: C4.5, RIPPER y K-NN. Compare el rendimiento de los clasificadores utilizando validación cruzada y la matriz de confusión. ¿Cuál escogería tomando en cuenta que los expertos consideran que los errores cometidos al diagnosticar una muestra maligna tienen un peso mayor que los cometidos sobre una muestra benigna?

Ejemplos de aplicaciones

Conjunto de datos *BREAST-CANCER-WISCONSIN*:

Número de instancias: 699

Clases: BENIGNO (65.5%) = 2

Número de variables: 10

MALIGNANT (34.5%) = 4

Nombre	Tipo	Ausencias	Media	StdDev	Mínimo	Máximo	Moda
ID	Numérico	0	1071704.099	617095.73	61634	13454352	
CLUMP	Numérico	0	4.418	2.816	1	10	
USIZ	Numérico	0	3.134	3.051	1	10	
USHA	Numérico	0	3.207	2.972	1	10	
MAR	Numérico	0	2.807	2.855	1	10	
EPI	Numérico	0	3.216	2.214	1	10	
BARE	Numérico	16 (2%)	3.545	3.644	1	10	
BLAN	Numérico	0	3.438	2.438	1	10	
NORM	Numérico	0	2.867	3.054	1	10	
MIT	Numérico	0	1.589	1.715	1	10	
CLASE	Nominal						4

1) Preparación de los datos:

a) Limpieza de los datos

- Se eliminaron los registros con ausencias (16)

b) Eliminación/Selección de variables

- Se eliminó la variable ID (identificador)

2) Minería de datos

- Tarea de minería de datos: Clasificación
- Algoritmos: C4.5, RIPPER y K-NN
- Medida de rendimiento: exactitud predictiva
- Técnica de evaluación: validación cruzada de 10 particiones