

Clasificación

(Parte 1: Introducción, Framework, Evaluación)

Bárbara Poblete

Sobre la Clasificación

- Técnica utilizada en minería de datos
- Viene del área de Machine Learning
- Método de “aprendizaje supervisado”

¿Qué es la Clasificación?

- Técnica que “aprende” automáticamente cómo clasificar objetos en dos o más clases determinadas
- Este aprendizaje se basa en datos previamente etiquetados (clásificados)
- Se aplica en caso en que “etiquetar” tiene un alto costo (por ej: trabajo humano experto)

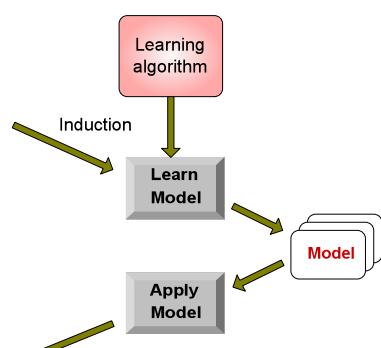
Componentes principales

- Conjunto de entrenamiento
- Algoritmo de clasificación
- Conjunto de validación
- Producen un “Modelo de Clasificación”

Proceso de Clasificación

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

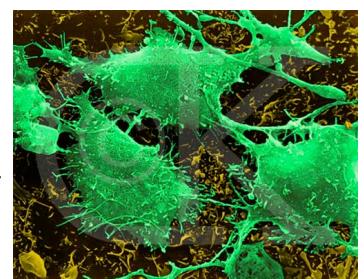


Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Ej. de tareas de clasificación

- Predecir si células de tumores son malignas o benignas
- Predecir respuesta de células cancerígenas a ciertas drogas



Ej. de tareas de clasificación

- Clasificar transacciones de tarjetas de crédito como legítimas o fraudulentas
- Predecir clientes en alto riesgo de “fuga”



Ej. de tareas de clasificación

- Categorizar tweets (sentimiento, geolocalización, y mucho más)
- Detección de caídas de Netflix usando clasificación de tweets



En resumen

- Dada una colección de objetos (set de entrenamiento)
 - Cada record contiene un set de atributos, uno de los cuales es su clase
- Encontrar un modelo para el atributo clase, en base a los otros atributos
- Meta: records nuevos deben ser asignados correctamente su clase
 - Un set de evaluación se utiliza para medir la exactitud del modelo.

Tarea de mapear set x a una clase y



Usos de los modelos

- **Descriptivo:** el modelo se utiliza como una herramienta descriptiva
- **Predictivo:** se utiliza para predecir la clase de objetos nuevos

Uso descriptivo

Table 4.1. The vertebrate data set.

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo	cold-blooded	scales	no	no	no	yes	no	reptile
dragon								
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard	cold-blooded	scales	yes	yes	no	no	no	fish
shark								
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Nota sobre la clasificación

- es mejor para **datos binarios y nominales**,
- **no es tan bueno para ordinales**, ya que **no consideran relación de orden** entre clases (ej. alto, mediano, bajo), también ignora información de subclases-superclases (mamíferos -> primates -> {humanos, monos}).
- Nos enfocamos en clases binarias y nominales.

Técnicas de clasificación

- Basados en Árboles de Decisión
- Métodos basados en Reglas
- Razonamiento en base a memoria
- Redes Neuronales
- Naïve Bayes y Redes de Soporte Bayesianas
- Support Vector Machines

La Clave del Éxito

- El modelo construido debe ser “generalizable”, es decir, debe aprender bien con muchos tipos de datos nuevos

¿Cómo saber si un modelo es bueno o no?

- Enfoque en la capacidad predictiva del modelo
- Más que en su rapidez para clasificar, construir modelos y escalar...

¿Cómo saber si un modelo es bueno o no?

1. Utilizando métricas de desempeño (performance metrics), y
2. Comparándolo con el desempeño de otros modelos posibles (bajo error de entrenamiento)
3. Probando con datos de prueba y otros datasets (generalizable, i.e. bajo error de generalización)

Performance Metrics (métricas de desempeño)

- Basadas en **contar** datos **correcta** e **incorrectamente** clasificados
- Accuracy (Exactitud): métrica más usada, o
- Error rate (Tasa de error)

Matriz de Confusión

		Clase predicha	
		clase = +	clase = -
Clase real	clase = +	a	b
	clase = -	c	d

Accuracy (Exactitud)

		Clase predicha	
Clase real		clase = +	clase = -
	clase = +	a (TP)	b (FN)
	clase = -	c (FP)	d (TN)

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



Ejemplo: Rumores en Twitter

(presentación externa <https://prezi.com/r6xefyatuywg/information-credibility-on-twitter/>)