



ESCUELA TÉCNICA SUPERIOR DE INGENIEROS  
INFORMÁTICOS

UNIVERSIDAD POLITÉCNICA DE MADRID

---

# Aprendizaje automático para flujos de datos

---

TRABAJO FIN DE MÁSTER  
MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL

AUTOR: Javier Ramos Fernández  
TUTOR/ES: María Concepción Bielza Lozoya y  
Pedro María Larrañaga Múgica



# Índice general

<b>1. Introducción</b>	<b>1</b>
1.1. Motivación . . . . .	1
1.2. Objetivos . . . . .	2
1.3. Metodología . . . . .	3
1.4. Organización de la memoria . . . . .	4
<b>2. Aprendizaje automático</b>	<b>5</b>
2.1. Notación . . . . .	6
2.2. Algoritmos de aprendizaje supervisado . . . . .	7
2.2.1. Clasificadores bayesianos . . . . .	7
2.2.2. Árboles de decisión . . . . .	11
2.2.3. Inducción de reglas . . . . .	13
2.2.4. Redes neuronales . . . . .	14
2.2.5. $k$ -vecinos más cercanos . . . . .	17
2.2.6. Máquinas de vector soporte . . . . .	18
2.2.7. Regresión logística . . . . .	19
2.2.8. Métodos combinados de aprendizaje . . . . .	21
2.3. Algoritmos de aprendizaje no supervisado . . . . .	23
2.3.1. Agrupamiento . . . . .	24
2.4. Series temporales . . . . .	27
2.5. Redes bayesianas para el descubrimiento de conocimiento . . . . .	27
<b>3. Estado del arte: Aprendizaje automático para flujos de datos</b>	<b>29</b>
3.1. Introducción . . . . .	29
3.1.1. Conceptos . . . . .	31
3.2. Algoritmos de aprendizaje supervisado . . . . .	35
3.2.1. Clasificadores bayesianos . . . . .	36
3.2.2. Árboles de decisión . . . . .	36
3.2.3. Inducción de reglas . . . . .	39
3.2.4. Redes neuronales . . . . .	40
3.2.5. $k$ -Vecinos más cercanos . . . . .	40
3.2.6. Máquinas de vector soporte . . . . .	40
3.2.7. Regresión logística . . . . .	41
3.2.8. Métodos combinados de aprendizaje . . . . .	41
3.3. Algoritmos de aprendizaje no supervisado . . . . .	41

3.3.1. Agrupamiento . . . . .	41
3.4. Series temporales . . . . .	42
3.5. Redes bayesianas para el descubrimiento de conocimiento . . . . .	42
<b>4. Conclusiones y líneas futuras de trabajo</b>	<b>43</b>
<b>5. Sobre las referencias</b>	<b>45</b>
<b>A. Anexos</b>	<b>49</b>
<b>Bibliografía</b>	<b>51</b>

# Índice de figuras

2.1.	Estructura del manto de Markov . . . . .	9
2.2.	Naive Bayes. Fuente: <a href="#">Bielza and Larranaga [2014]</a> . . . . .	9
2.3.	TAN. Fuente: <a href="#">Bielza and Larranaga [2014]</a> . . . . .	9
2.4.	$k$ -dependence Bayesian classifier. Fuente: <a href="#">Bielza and Larranaga [2014]</a> . . . . .	9
2.5.	Semi-naive Bayes. Fuente: <a href="#">Bielza and Larranaga [2014]</a> . . . . .	9
2.6.	Bayesian multinet. Fuente: <a href="#">Bielza and Larranaga [2014]</a> . . . . .	10
2.7.	Estructura de una red neuronal <a href="#">phuongphuong 4112716 and 405k4213231585</a> . . . . .	15
2.8.	SVM. Fuente: <a href="#">Unagar and Unagar [2017]</a> . . . . .	19
2.9.	Función logística o sigmoide. Fuente: <a href="#">mis [a]</a> . . . . .	20
2.10.	Problema de agrupamiento. Fuente: <a href="#">mis [b]</a> . . . . .	24
2.11.	Representación de un dendograma. Fuente: <a href="#">mis [c]</a> . . . . .	26
2.12.	Iteración del algoritmo EM. Fuente: <a href="#">Larrañaga and c. Bielza</a> . . . . .	27
3.1.	Real vs Virtual concept drift. Fuente: <a href="#">Pesaranghader et al. [2018]</a> . . . . .	32
3.2.	Tipos de concept drift según el ritmo de cambio. Fuente: <a href="#">Zliobaite [2010]</a> . . . . .	33
3.3.	Modelo <i>landmark window</i> . Fuente: <a href="#">E. Ntoutsi and Zimek [2015]</a> . . . . .	34
3.4.	Modelo <i>sliding window</i> . Fuente: <a href="#">E. Ntoutsi and Zimek [2015]</a> . . . . .	34
3.5.	Modelo <i>damped window</i> . Efecto del valor del factor de desvanecimiento $\lambda$ . Fuente: <a href="#">E. Ntoutsi and Zimek [2015]</a> . . . . .	35



# Índice de tablas

2.1. Notación utilizada en el trabajo . . . . .	6
2.2. Problema de clasificación supervisada . . . . .	7
2.3. Estructura de los datos en un problema de clasificación no supervisado	23
3.1. Algoritmos de aprendizaje para flujos de datos basados en árboles de decisión . . . . .	38





# Índice de algoritmos

1.	Pseudocódigo del algoritmo KNN. Fuente: <a href="#">Larranaga et al.</a> . . . . .	17
2.	Pseudocódigo del algoritmo k-medias estándar . . . . .	24



# Capítulo 1

## Introducción

### 1.1. Motivación

Actualmente vivimos en la era de la información, una época en la que se está generando una cantidad ingente de información. Esta abundancia de datos se debe principalmente a la aparición de ordenadores y de otros dispositivos que son capaces de recoger información de lo que nos rodea, procesarla y transmitirla, como son los teléfonos móviles. Además, su capacidad de conexión a Internet hace que haya una generación de tal cantidad de información que es imposible tener un control absoluto de todo lo que circula por esta red informática de nivel mundial.

Este aumento exponencial del volumen y variedad de información ha generado la necesidad de llevar a cabo un almacenamiento masivo de los datos, y con ello el interés por analizar, interpretar y extraer información útil de los mismos con el objetivo de obtener conocimiento. Para manejar toda esta información, los sistemas tradicionales de almacenamiento de datos no son convenientes puesto que no tienen la capacidad necesaria para su correcto procesamiento. El volumen, la variedad y la velocidad de los grandes datos causan inconvenientes de rendimiento cuando se utilizan técnicas tradicionales de procesamiento de datos [Mohanty et al. \[2015\]](#). Por ello surge lo que se denomina **Big Data**, un concepto relativamente nuevo que se refiere a conjuntos de datos cuyo tamaño va más allá de la capacidad de las herramientas típicas de software de bases de datos para almacenar, gestionar y analizar datos [Oguntimilehin and Ademola \[2014\]](#).

Los datos son la materia prima para conseguir información provechosa, que se puede utilizar para llevar a cabo una toma de decisiones y la realización de conclusiones, por lo que se han desarrollado nuevas herramientas que sobrepasan las herramientas disponibles anteriormente para tratar este tipo de volúmenes de datos. De esta manera, surge el concepto de **minería de datos**, que se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea primordial de la minería de datos es encontrar modelos inteligibles a partir de los datos que posibiliten el hallazgo de aspectos previamente desconocidos de los mismos.

Para ejecutar el proceso de extracción del conocimiento, una de las posibilidades más populares es la aplicación de una rama de la inteligencia artificial denominada **aprendizaje automático**, la ciencia (y el arte) de programar computadoras para que puedan aprender de los datos [Gron \[2017\]](#). En este caso, el objetivo es que los ordenadores aprendan automáticamente sin intervención humana. Este proceso de aprendizaje se realiza proporcionándoles a los algoritmos pertinentes una serie de datos sobre los que se entrenan con el objetivo de buscar patrones en los mismos y llevar a cabo mejores decisiones en el futuro.

En general, los algoritmos de aprendizaje automático asumen que los datos están disponibles a la hora de llevar a cabo el entrenamiento y que se generan a partir de una distribución estática. No obstante, la información presente hoy en día circula en un entorno que cambia de manera *continua y rápida*, de manera que la distribución que genera los datos puede sufrir transformaciones. Todo esto ha propiciado la aparición de lo que se denominan **flujos de datos** (*data streams*), que son secuencias continuas y ordenadas de datos en tiempo real.

A la hora de tratar con flujos de datos, los algoritmos de aprendizaje automático tradicionales no son capaces de funcionar correctamente puesto que el volumen de los mismos puede llegar a ser *infinito*. Para que puedan realizar el entrenamiento, el conjunto de datos debe estar almacenado en memoria, y solo pueden llevar a cabo tareas de predicción cuando la fase de entrenamiento haya finalizado. Sin embargo, el almacenamiento de datos generados continuamente se hace inviable. Además, para que los algoritmos puedan manejar flujos de datos cuya distribución que los subyace puede cambiar es necesario un *procesamiento en tiempo real*, característica que los algoritmos tradicionales no poseen.

En este trabajo pretendemos realizar una revisión de literatura sobre métodos propuestos para aplicar aprendizaje automático en flujos de datos con el objetivo de tener una visión global de las diferentes posibilidades existentes para resolver el problema planteado anteriormente. Para ello, vamos a tener en cuenta diferentes algoritmos de aprendizaje automático existentes en el mercado actual.

## 1.2. Objetivos

El objetivo principal de este trabajo es realizar una **comparativa amplia de las diferentes aproximaciones propuestas para resolver el problema de aprendizaje automático para flujos de datos, principalmente tanto supervisado como no supervisado**. Para abordar esta meta, nos centraremos en los algoritmos de aprendizaje automático más populares. En el caso de aprendizaje supervisado, nos enfocaremos en propuestas relacionadas con *árboles de decisión*, *redes bayesianas*, *redes neuronales*, *inducción de reglas*, *vecinos más cercanos*, *máquinas de soporte vectorial* y *métodos combinados de aprendizaje* (ensemble). Con respecto al aprendizaje no supervisado, consideraremos primordialmente los algoritmos de *agrupamiento* (clustering), uno de los principales de este tipo de aprendizaje.

También se hará hincapié en algunas propuestas relacionadas con aprendizaje

### semisupervisado

Para alcanzar este objetivo, se han contemplado una serie de metas específicas dentro del proyecto:

- *Búsqueda de artículos relacionados con cada uno de los algoritmos de aprendizaje automático contemplados.* Con el objetivo de llevar a cabo una comparativa de los diferentes métodos propuestos de aprendizaje automático para flujos de datos, es de vital importancia realizar una búsqueda de diferentes artículos propuestos para cada uno de los algoritmos de aprendizaje automático mencionados anteriormente. Para ello hemos utilizado diferentes revistas que aborden esta rama de la inteligencia artificial, así como otros recursos como *Google Académico*.
- *Anotación de aspectos claves de los diferentes artículos encontrados.* Con ello se pretende tener disponible información resumida de las diferentes propuestas que se encuentran a nuestra disposición y utilizarla para compararlas con otros artículos con la finalidad de dar una perspectiva general de la utilidad de las diferentes aproximaciones existentes.
- *Estructuración de la revisión de la literatura de algoritmos de aprendizaje automático para flujos de datos en función de los algoritmos en los que se centren las diferentes propuestas encontradas.* Se persigue comparar los diferentes artículos hallados según el algoritmo de aprendizaje automático que aborden para que funcionen correctamente con flujos de datos. Para ello hemos realizado una división de los mismos en diferentes apartados.
- *Exploración de las diferentes revisiones halladas sobre algoritmos de aprendizaje automático para flujos de datos.* El objetivo de esto es conocer qué propuestas de las que hemos encontrado se encuentran referenciadas en esas revisiones para tener en cuenta qué artículos son novedosos con respecto a dichas revisiones. Además, se lleva a cabo esta exploración para tener conocimiento de que algoritmos de aprendizaje automático tratan con el fin de aportar nuevos algoritmos. Todo esto se realiza con la finalidad de establecer qué características nos diferencian de las revisiones encontradas.

## 1.3. Metodología

Para desarrollar este proyecto se ha seguido una metodología que permitiera sobrellevar las dificultades del mismo de la mejor forma posible. A continuación se enumeran los pasos ejecutados durante este proceso:

- En primer lugar, hemos procedido a realizar una búsqueda **exhaustiva de artículos relacionados con aprendizaje automático para flujos de datos**. Antes de realizar una comparación entre las diferentes propuestas, es necesario recabar la mayor cantidad de aproximaciones desarrolladas con el fin de

ampliar nuestra visión genérica del estado del arte del tema abordado. Para lograr esto, hemos indagado en numerosas revistas que incluyen dentro de su temática el aprendizaje automático para flujos de datos. Algunas que se han consultado y que son relevantes en el mundo académico son *Journal of Machine Learning Research*, *IEEE Transactions on Knowledge and Data Engineering* y *Machine Learning Journal*. También se han explorado diferentes editoriales como *Elsevier* y *Springer* en las que, aparte de artículos publicados en revistas, también aparecen propuestas de conferencias. De forma complementaria, hemos buscado artículos en *Google Académico*, una herramienta de búsqueda de Google que permite hallar literatura del mundo científico de diferentes recursos (bibliotecas, editoriales, etcétera).

- Tras aglomerar una cantidad aceptable de artículos, hemos iniciado la **lectura de los mismos**. Durante este proceso, hemos ido apuntando características relevantes de las propuestas con el fin de poder tener la información fundamental para realizar la comparativa entre diferentes artículos. Para guardar dicha información, hemos creado un documento en el que, por cada propuesta, se apunta el *título*, la *fecha de publicación* y *contenido de interés* de las mismas, así como *comparaciones con otras aproximaciones* que se encontraba dentro de los artículos para utilizarlas en el estado del arte. Asimismo, hemos añadido información adicional con respecto a la *presencia o no de los artículos de las diferentes revisiones* encontradas que abordan el tema de aprendizaje automático para flujos de datos con el objetivo de tener conocimiento sobre qué artículos de los que hemos encontrado aportan nueva información con respecto a dichas revisiones para establecer características que nos diferencian de las revisiones halladas. Estas propuestas se han ordenado por *fecha de publicación* para contemplar la evolución cronológica de las mismas.
- De forma paralela a la lectura de artículos, hemos procedido a **buscar más propuestas**, enfocándonos en encontrar aquellas que son más recientes para aportar más información que nos diferencie de las revisiones encontradas. También hemos aprovechado las referencias a otros papers que hemos hallado en los artículos revisados para añadirlos al estado del arte.
- Por otra parte, se ha creado un documento en el que se ha ido **almacenando información útil durante la lectura de los artículos para relacionar diferentes propuestas de la mejor manera posible**. Por ejemplo, hemos anotado diferentes categorías en las que se podrían clasificar las distintas propuestas.

## 1.4. Organización de la memoria

Pendiente

## Capítulo 2

# Aprendizaje automático

Arthur Samuel, uno de los pioneros del aprendizaje automático, estableció en 1959 una definición general de esta rama de la inteligencia artificial:

*El aprendizaje automático es el campo de estudio que da a las computadoras la capacidad de aprender sin estar programadas explícitamente.*

Tom Mitchell, otro investigador de aprendizaje automático reputado, propuso en 1997 una definición más precisa y más orientado a la ingeniería:

*Se dice que un programa de ordenador aprende de la experiencia  $E$  con respecto a alguna tarea  $T$  y alguna medida de rendimiento  $P$ , si su rendimiento en  $T$ , medido por  $P$ , mejora con la experiencia  $E$ .*

Por lo tanto, el aprendizaje automático se centra en aplicar sistemáticamente algoritmos para sintetizar las relaciones subyacentes de forma automática en un conjunto de datos proporcionados en forma de ejemplos a través de una fase de entrenamiento, de tal forma que en el futuro se utilice esta información para la ejecución de predicciones de eventos desconocidos y una mejor toma de decisiones. Los campos de aplicación que existen de esta rama de la inteligencia artificial son muy variados. Algunos de ellos son la predicción bursátil, predicción meteorológica, detección de correos spam, construcción de sistemas de recomendación y detección de fraude en el uso de tarjetas de crédito.

Según el propósito que persigan los algoritmos de aprendizaje automático, éstos se clasifican en dos categorías principales: **aprendizaje supervisado** y **aprendizaje no supervisado**. En el aprendizaje supervisado se engloban aquellos algoritmos que buscan aprender una *función de mapeo* entre una serie de características de entrada (variables predictoras) y una variable de salida (variable clase) mediante la ejecución de una fase de entrenamiento en la que se utilizan *datos de entrenamiento etiquetados* (valor de la variable clase conocida), de tal forma que se utiliza esta función para predecir el valor de la variable de salida a partir de los valores de las variables predictoras. En cambio, el aprendizaje no supervisado la finalidad es aprender una *función que modele la estructura o distribución subyacente en los datos*

a partir de datos de entrenamiento no etiquetados, de manera que se lleva a cabo una exploración de los datos sólo conociendo los valores de las variables predictoras.

Existe además otra categoría en la que se pueden incluir las técnicas de aprendizaje automático denominada **aprendizaje semisupervisado**. Los algoritmos que llevan a cabo este tipo de aprendizaje entrenan sobre un conjunto de datos parcialmente etiquetados, generalmente muchos datos sin etiquetar y una pequeña parte de datos etiquetados. Debido a que para la fase de entrenamiento se utilizan tanto datos etiquetados como no etiquetados, se considera que esta categoría se sitúa entre el aprendizaje supervisado y el no supervisado. En este trabajo se van a abordar algunas propuestas relacionadas con el aprendizaje semisupervisado para flujos de datos; no obstante, la contribución predominante de algoritmos de aprendizaje automático para flujos de datos que se va a realizar en este proyecto proviene de las categorías de aprendizaje supervisado y no supervisado.

## 2.1. Notación

A continuación se expone las notaciones que más se van a utilizar en este trabajo:

Tabla 2.1: Notación utilizada en el trabajo

Símbolo	Explicación
$X_i$	Variable predictora $i$
$\mathbf{X}$	Conjunto de variables predictoras
$x_i$	Valor de la variable predictora $i$
$\Omega_{X_i}$	Dominio de valores de la variable $i$
$\Omega_C$	Dominio de valores de la variable clase (finito)
$x_{ij}$	Valor $j$ de la variable predictora discreta $i$
$p(x_{ij})$	Probabilidad del valor $j$ de una variable $i$
$\mathbf{x}$	Instancia de las variables predictoras
$x_i^{(j)}$	Valor de la variable predictora $i$ de la instancia $j$
$\mathbf{x}^{(j)}$	Valores de las variables predictoras de la instancia $j$
$C$	Variable clase
$c_j$	Posible valor $i$ de la variable clase
$c^{(j)}$	Valor de la variable clase en la instancia $j$
$(\mathbf{x}^{(j)}, c^{(j)})$	Estructura de una instancia $j$ de clasificación supervisada
$N$	Número de instancias de un conjunto de datos
$n$	Número de variables predictoras
$m$	Número de valores que puede tomar una variable predictora
$k$	Número de valores que puede tomar la variable clase
$D$	Fichero de casos



## 2.2. Algoritmos de aprendizaje supervisado

La mayor parte de las propuestas que se van a abordar en este trabajo relacionadas con el aprendizaje automático para flujos de datos se engloban dentro de la categoría de **aprendizaje supervisado**. Los algoritmos de clasificación supervisada son aquellos en los que, a partir de un conjunto de ejemplos clasificados (conjunto de entrenamiento)  $D = ((\mathbf{x}^{(1)}, c^{(1)}), \dots, (\mathbf{x}^{(N)}, c^{(N)}))$ , se intenta clasificar un segundo conjunto de instancias. Formalmente, en el aprendizaje supervisado el objetivo es encontrar una función  $f$  que permita mapear una instancia  $\mathbf{x} = (x_1, \dots, x_n)$  a una determinada clase  $c$ :

$$\begin{aligned} f : \Omega_{X_1} \times \dots \times \Omega_{X_n} &\longrightarrow \Omega_c \\ \mathbf{x} = (x_1, \dots, x_n) &\mapsto c \end{aligned} \quad (2.1)$$

A continuación se expone la estructura típica que presenta un problema de aprendizaje supervisado:

Tabla 2.2: Problema de clasificación supervisada

	$X_1$	$\dots$	$X_n$	C
$(\mathbf{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\dots$	$x_n^{(1)}$	$c^{(1)}$
$(\mathbf{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\dots$	$x_n^{(2)}$	$c^{(2)}$
$\dots$		$\dots$		$\dots$
$(\mathbf{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\dots$	$x_n^{(N)}$	$c^{(N)}$
$\mathbf{x}^{(N+1)}$	$x_1^{(N+1)}$	$\dots$	$x_n^{(N+1)}$	???

Los algoritmos de aprendizaje automático que se van a tratar en este documento gozan de una gran popularidad. Concretamente, nos enfocaremos en **clasificadores bayesianos**, **árboles de decisión**, **inducción de reglas**, **redes neuronales**, **k-vecinos más cercanos**, **máquinas de vector soporte**, **regresión logística** y **métodos combinados de aprendizaje**.

### 2.2.1. Clasificadores bayesianos

Un tipo de modelo que se utiliza ampliamente para la clasificación supervisada son las **redes bayesianas**. Las *redes bayesianas* son modelos gráficos que permiten representar de manera sencilla, compacta, precisa y comprensible la distribución de probabilidad conjunta de un conjunto de variables aleatorias. Este modelo gráfico está compuesto por *nodos*, que representan a las variables aleatorias; *arcos*, que representan las relaciones de dependencia entre nodos; y *tablas de probabilidad condicional*, que representan la distribución de probabilidad condicional de cada uno de los nodos.

Formalmente, la estructura de una red bayesiana sobre un conjunto de variables aleatorias  $X_1, \dots, X_n, C$  es un grafo acíclico dirigido cuyos vértices corresponden a las variables cuyos arcos codifican las dependencias e independencias probabilísticas entre tripletas de variables y, en cada uno de los vértices, se representa una distribución categórica local  $p(x_i|\mathbf{pa}(x_i))$  o  $p(c|\mathbf{pa}(c))$ , donde  $\mathbf{pa}(x_i)$  es un conjunto de valores para el conjunto de variables  $\mathbf{Pa}(X_i)$ , que son los padres de la variable  $X_i$  en el modelo gráfico. Lo mismo se aplica para  $\mathbf{pa}(c)$  (Bielza and Larranaga [2014]). Por lo tanto, la factorización que permite llevar a cabo la red bayesiana de la probabilidad conjunta de todas las variables aleatorias y que evita estimar un número exponencial de parámetros es la siguiente:

$$p(\mathbf{x}, c) = p(c|\mathbf{pa}(c)) \prod_{i=1}^n p(x_i|\mathbf{pa}(x_i)) \quad (2.2)$$

Las redes bayesianas, cuando se utilizan con propósitos de realizar tareas de clasificación, reciben el nombre de **clasificadores bayesianos**. En los clasificadores bayesianos, el objetivo es asignar la clase más probable a una instancia determinada, definida por un conjunto de valores de las variables predictoras. En términos probabilísticos, se asigna a una instancia de prueba la etiqueta de clase con la *mayor probabilidad a posteriori* (MAP). Es decir:

$$\operatorname{argmax}_c p(c|\mathbf{x}) \quad (2.3)$$

Utilizando la regla de Bayes, podemos relacionar los términos de las ecuaciones 2.2 y 2.3 y además, puesto que el objetivo es calcular el valor de  $C$  con mayor probabilidad a posteriori, no es necesario tener en cuenta el denominador en la regla de Bayes (el factor de normalización). De esta manera, obtenemos la siguiente expresión (Bielza and Larranaga [2014]):

$$\operatorname{argmax}_c p(c|\mathbf{x}) = \operatorname{argmax}_c p(\mathbf{x}, c) \quad (2.4)$$

De esta manera, podemos utilizar la ecuación 2.2 para hallar la clase con la mayor probabilidad a posteriori. Esta ecuación establece el caso general de los clasificadores bayesianos, en el que  $p(\mathbf{x}, c)$  se puede factorizar de diferentes maneras, por lo que tenemos que buscar lo que se denomina el **manto de Markov** (*Markov blanket*) de la variable  $C$  para encontrar la solución de la ecuación 2.2. El *manto de Markov* se define como el conjunto de variables  $MB_c$  que hacen que, dado dichas variables, la variable  $C$  sea condicionalmente independiente de las demás variables de la red bayesiana. El *manto de Markov* está formado, cogiendo a la variable  $C$  de referencia, por los **padres**, los **hijos** y los **padres de los hijos**. De esta forma (Bielza and Larranaga [2014]):.

$$p(c|\mathbf{x}) = p(c|\mathbf{x}_{MB_c}) \quad (2.5)$$

A continuación se expone la estructura del manto de Markov:

Figura 2.1: Estructura del manto de Markov



Para el caso específico en el que la variable  $C$  no tenga padres y, utilizando la regla de la cadena, la probabilidad conjunta de las variables predictoras y de la variable clase se puede expresar de la siguiente manera (Bielza and Larranaga [2014]):

$$p(\mathbf{x}, c) = p(c)p(\mathbf{x}|c) \quad (2.6)$$

de tal forma que el objetivo es maximizar en  $c$

Con respecto a lo comentado previamente, los distintos clasificadores bayesianos que existen se basan en establecer que la variable clase  $C$  no tenga padres, y se diferencian en la forma en la que factorizan  $p(\mathbf{x}|c)$ . Los clasificadores bayesianos más conocidos son **naive Bayes**, **Tree Agumented Naive Bayes (TAN)**, **k-dependence Bayesian classifier ( $k$ -DB)**, **Semi-naive Bayes** y **Bayesian multinet**. A continuación se expone la estructura de cada uno de ellos:

Figura 2.2: Naive Bayes. Fuente: Bielza and Larranaga [2014]



Figura 2.3: TAN. Fuente: Bielza and Larranaga [2014]

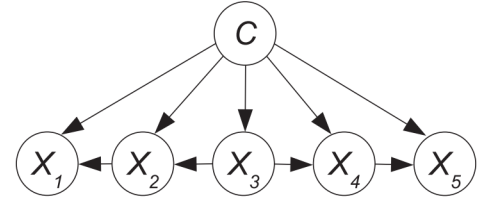
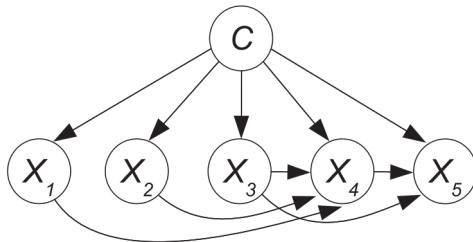
Figura 2.4:  $k$ -dependence Bayesian classifier. Fuente: Bielza and Larranaga [2014]

Figura 2.5: Semi-naive Bayes. Fuente: Bielza and Larranaga [2014]

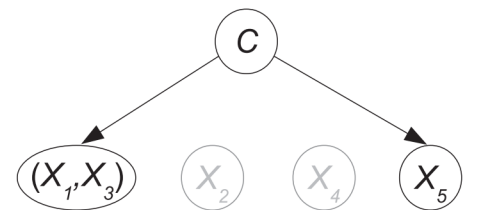
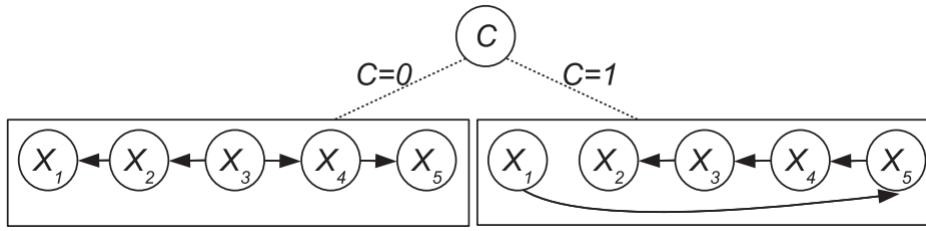


Figura 2.6: Bayesian multinet. Fuente: [Bielza and Larranaga \[2014\]](#)

El clasificador *Naive Bayes* es uno de los clasificadores bayesianos más simples, cuya estructura se basa en establecer la variable clase sin padre, las variables predictoras como vértices hijo de la variable clase y sin ningún tipo de dependencias entre las variables predictoras; asume que, dada la variable clase, las variables predictoras son condicionalmente independientes. Se ha demostrado que este tipo de clasificador bayesiano presenta un desempeño bastante aceptable, incluso realizando suposiciones de independencias tan firmes y que en la realidad no se suelen cumplir; no obstante, se ha intentado mejorar este clasificador relajando estas suposiciones. En este caso, el clasificador *TAN* se basa en realizar un aumento de la estructura de la red bayesiana **añadiendo arcos entre los diferentes atributos**, de tal forma que cada variable predictora tenga como padres la variable clase y una variable predictora como máximo. En este sentido, el clasificador *k-DB* permite a las variables predictoras tener como padres a la variable clase y como máximo un número  $k$  de variables predictoras, de tal manera que se pueden representar más posibles dependencias entre atributos. El clasificador *k-DB* generaliza tanto al *Naive Bayes* como al *TAN* puesto que el *Naive Bayes* se puede ver como un *k-DB* con  $k = 0$  y el *TAN* como un *k-DB* con  $k = 1$ .

Otro clasificador bayesiano utilizado con frecuencia es el *Semi-naive Bayes*, que particiona el conjunto de atributos; es decir, considera **nuevas variables teniendo en cuenta productos cartesianos de las mismas**, de tal forma que modela dependencias entre las variables predictoras originales. Por otra parte, existe el clasificador *Bayesian multinet*, cuya estructura está compuesta por varias redes bayesianas locales, donde cada una de ellas representa la **probabilidad conjunta de las variables predictoras condicionada a un subconjunto de valores de la variable clase**. De esta manera, en función de los posibles valores de la variable clase, las dependencias entre las variables predictoras pueden ser distintas, dando lugar a la noción de *asimetría en las declaraciones de independencia*.

(Meter si es necesario lo del aprendizaje generativo y discriminativo de las redes bayesianas)

Aprendizaje discriminativo de clasificadores Bayesianos Discrete Bayesian Network Classifiers: A Survey Learning Augmented Bayesian Classifiers: A Comparison of Distribution-based and Classification-based Approaches Efficient Heuristics for Discriminative Structure Learning of Bayesian Network Classifiers Bayesian classifiers based on kernel density estimation: Flexible classifiers

### 2.2.2. Árboles de decisión

El aprendizaje mediante **árboles de decisión** es un método de aproximación de una función objetivo en el cual la función objetivo es representada mediante un *árbol de decisión*, un clasificador expresado como una partición recursiva del espacio de instancias que consta de *nodos*, *ramas* y *hojas*.

- Los *nodos* representan atributos utilizados para particionar un conjunto de datos en subconjuntos de los mismos de acuerdo con una determinada función discreta de los valores de los atributos de entrada. Es decir, representa un test del valor de un atributo.
- Las *ramas* son los distintos valores de los atributos (nodos), que dan lugar a los diferentes hijos de los nodos (diferentes particiones). Las particiones que se realizan en los distintos nodos del árbol de decisión varían en función de si los atributos son *discretos* o *numéricos*. En el caso de los atributos *discretos*, se realizan las particiones en función de cada uno de los posibles valores de ese atributo. En el caso de los atributos *numéricos*, se particiona teniendo en cuenta diferentes rangos de valores.
- Las *hojas* representan las posibles etiquetas de clase. Los árboles de decisión también se conocen como *árboles de clasificación*, de tal forma que este paradigma clasifica una instancia a una determinada clase. No obstante, también existen los *árboles de regresión*, donde la variable de destino puede tomar valores continuos.

De esta manera, a la hora de clasificar una instancia, se comienza desde el nodo *raíz* y, en función de los valores de las variables predictoras de esa instancia, el ejemplo va recorriendo las ramas pertinentes asociadas a esos valores hasta que llega a una hoja, que tiene asignada una clase que se va a utilizar para clasificar la instancia. Por lo tanto, las *ramas* representan las conjunciones de características que conducen a esas etiquetas de clase.

A la hora de llevar a cabo la construcción de ese paradigma clasificatorio, es necesario que las particiones se realizan de tal forma que los subconjuntos resultantes sean lo más *puros* posible, es decir, que en cada subconjunto de instancias de las hojas todos los ejemplos pertenezcan a la misma clase, lo que conlleva a una correcta partición de los datos proporcionados. La construcción del árbol se basa principalmente escoger, en cada caso, el atributo que mejor particiona en cada nodo, y para tomar esta decisión se utiliza una *función de medida* del grado de impureza de la partición realizada por cada uno de los atributos considerados. Algunas de las funciones más populares son la **información mutua** (o *ganancia de información*) y el **índice Gini**. La *información mutua* se define como la cantidad que mide una relación entre dos variables aleatorias; concretamente mide cuánto se reduce la incertidumbre (*entropía*, medida de impureza) de una variable al conocer el valor de la otra variable. En el caso de la elección del mejor atributo en los árboles de decisión, interesa la cantidad de información mutua entre una variable predictora y la variable

clase, de tal forma que se elige aquel atributo que más reduzca la incertidumbre. La fórmula matemática es la siguiente:

$$I(X_i, C) = H(C) - H(C|X_i) \quad (2.7)$$

donde  $H(C)$  representa la entropía de una variable, en este caso de la variable clase, que se define de la siguiente manera:

$$H(C) = - \sum_{j=1}^m p(c_j) \log_2 p(c_j) \quad (2.8)$$

y  $H(C_i)$  representa la entropía de una variable sabiendo el valor de otra variable, en este caso la entropía de la variable clase sabiendo el valor de una variable predictora, que se define de la siguiente manera:

$$H(C|X_i) = - \sum_{j=1}^{m_C} \sum_{l=1}^{m_{X_i}} p(c_j, x_{il}) \log_2 p(c_j|x_{il}) \quad (2.9)$$

Por otra parte, el *índice Gini* es un criterio basado en impurezas que mide las divergencias entre la distribución de probabilidad de los valores de los atributos objetivo (Gulati et al. [2016]). Es una alternativa a la *ganancia de información*. Se define con la siguiente fórmula:

$$Gini = 1 - \sum_{j=1}^m p(c_j)^2 \quad (2.10)$$

Existen dos maneras llevar a cabo el proceso de construcción del árbol de decisión, desde **arriba hacia abajo** (*top-down*) y desde **abajo hasta arriba** (*bottom-up*). No obstante, la aproximación *top-down* es la que goza de mayor popularidad. Algunos de los algoritmos más conocidos que llevan a la práctica esta aproximación son el **ID3** (Quinlan [1986]), el **C4.5** (Quinlan [1993]) y **CART** (Breiman et al. [1984]). Estos algoritmos pertenecen a la familia de los *Top Down Induction of Decision Trees* (TDIDT), que inducen el modelo del árbol de decisión a partir de datos preclasificados. El algoritmo de construcción en el que se basan los árboles de esta familia es el *método de Hunt*, que es el siguiente:

(Meter algoritmo del documento "Tema 8. Árboles de decisión")

El algoritmo *ID3* construye el árbol de decisión mediante la aproximación *top-down* sin realizar *backtracking*, es decir, lleva a cabo una estrategia de búsqueda voraz a través del espacio de todos los árboles de clasificación posibles. Para tomar la decisión de elegir la variable que aporta mayor información a la hora de realizar las diferentes particiones el algoritmo utiliza la **ganancia de información**. El *ID3* tiene algunos inconvenientes: se sobreajusta a los datos de entrenamiento, no es capaz de manejar atributos numéricos ni valores faltantes y no soporta un podado del árbol.

El algoritmo *C4.5* se desarrolló como una mejora del algoritmo *ID3*. En primer lugar, en lugar de utilizar la *ganancia de información* para elegir la variable

predictora más informativa en cada momento, emplea lo que se denomina la **proporción de ganancia** (*gain ratio*), que se calcula dividiendo la información mútua entre una variable predictora y la variable clase por la entropía de la variable predictora ( $I(X_i, C)/H(X_i)$ ), que evita a aquellas variables predictoras que tengan un mayor rango de valores no tengan tanta probabilidad de ser elegidas, como ocurría en el *ID3*. Además, permite trabajar con **atributos continuos** definiendo un umbral, de tal forma que particiona las instancias en función de si el valor del atributo es mayor o menor que el umbral. Por otra parte, con respecto al manejo de **datos faltantes**, se estiman los mismos imputación, y también **asigna pesos** diferentes a los atributos en función del coste asociado a los mismos. Asimismo, trata de lidiar el sobreajuste del modelo realizando una **poda**, que puede hacerse parando la construcción del árbol antes de que clasifique perfectamente los datos (*pre-prunning*) o permitiendo que el modelo se sobreajuste y luego sustituir subárboles por hojas (*post-prunning*) y es capaz de manejar el ruido. La limitación principal del *C4.5* es que produce modelos que no dan buenos resultados cuando los atributos tienen un amplio rango de valores.

Con respecto a *CART* (*Classification and Regression Tree*, Árboles de Clasificación y Regresión), produce árboles binarios. Una característica importante de este algoritmo es que es capaz de generar *árboles de regresión*; en este caso, *CART* busca aquellas particiones que minimicen el *error cuadrático medio* y la predicción en cada hoja se realiza mediante una *media ponderada*. Por otra parte, para elegir el atributo que va a particionar en cada nodo del árbol *CART* utiliza el *índice Gini* y, al igual que el *C4.5*, es capaz de manejar atributos categóricos y numéricos, el ruido. También tiene la habilidad de tratar con valores atípicos. No obstante, este algoritmo puede producir árboles inestables y realiza particiones mediante una sola variable.

### 2.2.3. Inducción de reglas

Otro de los paradigmas utilizados frecuentemente para tareas de clasificación es la **inducción de reglas**. El objetivo de ese modelo es encontrar *asociaciones* o *correlaciones* entre las variables que describen las instancias de un conjunto de datos mediante la inducción de *reglas de asociación*, que tienen la siguiente forma:

$$Y \implies Z \quad (2.11)$$

donde  $Y$  y  $Z$  son conjuntos de literales (o atributos) que tienen asociado un determinado valor y  $Y \cap Z = \emptyset$ . El significado de esta representación es que las instancias del conjunto de datos que contienen a  $Y$  tienden a contener a  $Z$ . Los conjuntos  $Y$  y  $Z$  se denominan **antecedente** y **consecuente** de la regla, respectivamente. Estas reglas también se pueden expresar en el formato *IF antecedente THEN consecuente*.

En los métodos de clasificación basados en este tipo de estructuras, las reglas se utilizan con finalidad de llevar a cabo una tarea de clasificación, recibiendo el nombre de *reglas de clasificación*. En las *reglas de asociación*, tanto en la parte del antecedente como la del consecuente puede aparecer cualquier variable del conjunto

de datos y *más de un par variable-valor* ; no obstante, en las *reglas de clasificación* la parte del antecedente contiene pares *variable predictora-valor* que se combinan para definir la parte del consecuente, que va a ser **la clase a la que se va a proceder a clasificar la instancia con dichos valores en los atributos**. En el caso de los atributos cuyo rango de valores sea continuo, se utilizan particiones de ese rango para discretizarlos. De esta manera, las reglas de clasificación se expresan de la forma *IF combinación\_valores\_variables\_predictoras THEN valor\_clase*.

La inducción de reglas, aparte de ser un modelo **transparente** y fácilmente **interpretable**, es parecida al paradigma de *árboles de decisión* puesto que un árbol de decisión se puede descomponer en un conjunto de reglas de clasificación (la inducción de reglas es más **genérica**). Además, la estructura que presentan las reglas es más **flexible** que la forma jerárquica que tiene el árbol de decisión puesto que las reglas son componentes separadas que pueden evaluarse de forma aislada y ser eliminadas del modelo sin dificultades, al contrario que los árboles de decisión que habría que reestructuralo al realizar alguna eliminación en el modelo. No obstante, las reglas **no garantizan que puedan cubrir toda la región del espacio de entrada**, de manera que puede ocurrir que llegue una nueva instancia a clasificar y las reglas no cubran dicha instancia; a diferencia de la inducción de reglas, las aproximaciones utilizadas para clasificación que se basan en árboles de decisión cubren todo el espacio de valores de los atributos de entrada.

Con respecto a la región de entrada de datos que cubren los *árboles de decisión*, éstos lo particionan en *regiones mutuamente exclusivas*, y en algunas ocasiones no es conveniente este tipo de divisiones del espacio de entrada; en su lugar, sería adecuado que esas **regiones de decisión se solapen**, y con el modelo de inducción de reglas se puede conseguir.

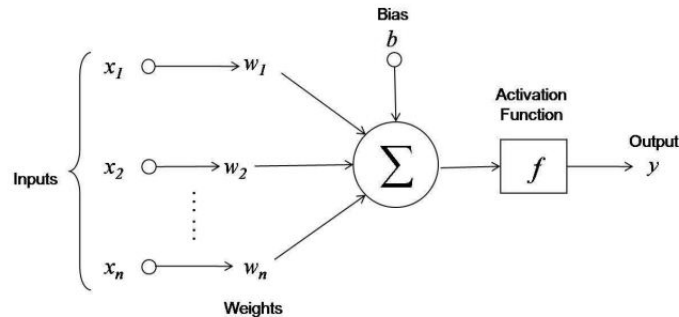
A la hora de llevar a cabo la predicción de una nueva instancia mediante el paradigma de *inducción de reglas*, se comprueban los antecedentes de las reglas para ver si los valores de las variables predictoras de la instancia coinciden con la parte izquierda de las reglas. Una vez realizado esto, se comprueba el valor de la variable clase del consecuente del conjunto de reglas obtenido del paso anterior. Si todas ellas tienen asignada la misma clase, el nuevo ejemplo se clasifica a dicha clase; en caso contrario, es necesario resolver el conflicto mediante la utilización de alguna métrica.

## 2.2.4. Redes neuronales

Las **redes neuronales** son un modelo computacional de interconexión de neuronas en una red que colabora para producir un estímulo de salida. El bloque de construcción de este sistema son las *neuronas artificiales*, que son unidades computacionales simples que ponderan las señales de entrada y producen una señal de salida usando una función de activación:



Figura 2.7: Estructura de una red neuronal [phuongphuong 4112716](#) and [405k4213231585](#)



En primer lugar en el perceptrón multicapa se recibe una serie de **entradas**, que pueden ser características de un conjunto de entrenamiento o salidas de otras neuronas. A continuación, se aplican unos **pesos** a las entradas, que se suelen inicializar a valores aleatorios pequeños, como valores en el rango de 0 a 0.3. Después, las entradas ponderadas se suman junto con un *sesgo* o *bias* que tiene la neurona (se interpreta como una entrada que permite desplazar la función de activación a la izquierda o a la derecha, que siempre tiene el valor 1.0 y que también debe ser ponderada) que, a su vez, pasan a través de una **función de activación**, obteniendo así las **salidas**. Esta función de activación es un simple mapeo de la entrada ponderada sumada a la salida de la neurona; es decir, se utiliza para determinar la salida de la red neuronal como si o no: mapea los valores resultantes de 0 a 1 o de -1 a 1, etc. (dependiendo de la función). Las distintas funciones de activación se engloban en dos tipos, *funciones de activación lineales* y *funciones de activación no lineales* y existen una gran variedad de ellas: *función sigmoide*, *Tanh*, *ReLU*, etc.

La estructura de una red neuronal se compone de las siguientes partes:

- **Capa de entrada:** La capa que toma la entrada del conjunto de datos se denomina *capa de entrada*, puesto que es la parte que se expone en la red de la red. Éstas no son neuronas como se describió anteriormente, sino que simplemente pasan los valores de las entradas a la siguiente capa.
- **Capas ocultas:** Las capas posteriores a la capa de entrada se denominan *capas ocultas* porque no se exponen directamente a la entrada y están formadas por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores.
- **Capa de salida:** La última capa de la red neuronal se denomina *capa de salida* y es responsable de producir un valor o un vector de valores que dependerán del problema a resolver. La elección de la función de activación en la capa de salida está fuertemente limitada por el tipo problema que se está modelando.

## Entrenamiento de una red neuronal

El primer paso para entrenar una red neuronal es *preparar los datos*, de tal forma que éstos deben ser numéricos y tienen que estar escalados de manera consistente. Con la *normalización* (reescalar al rango entre 0 y 1) y la *estandarización* (para que la distribución de cada columna tenga la media cero y la desviación estándar de 1, de modo que todas las entradas se expresan en rangos similares), el proceso de entrenamiento de la red neuronal se realiza con mucha mayor velocidad.

Uno de los algoritmos de entrenamiento para redes neuronales más populares se denomina **descenso por gradiente** (*gradient descent*). En este algoritmo la red procesa la entrada hacia delante activando las neuronas a medida que se va avanzando a través de las capas ocultas hasta que finalmente se obtiene un valor de salida; esto se denomina *propagación hacia delante* en la red neuronal. La salida del grafo se compara con la salida esperada y se calcula el error. Este error es entonces propagado de nuevo hacia atrás a través de la red neuronal, una capa a la vez, y los pesos son actualizados de acuerdo a su grado de contribución al error calculado (*algoritmo de retropropagación*).

A la hora de llevar a cabo el entrenamiento de la red neuronal, existen diferentes posibilidades de realizarlo según el número de instancias que se procesen antes de actualizar los pesos de la red neuronal, que está definido por el hiperparámetro denominado **tamaño del lote** (*batch size*). Una de ellas es procesar todo el conjunto de datos (el tamaño del lote es el número de instancias del conjunto de datos), guardar los errores de todos los ejemplos de entrenamiento y actualizar los parámetros de la red (*batch gradient descent*). Otra posibilidad es definir el tamaño del lote a una instancia, de tal manera que los pesos en la red neuronal se pueden actualizar al procesar un solo ejemplo de entrenamiento (*stochastic gradient descent*). También existe la opción de definir un tamaño de lote que se encuentre entre una sola instancia y todo el conjunto de entrenamiento, del tal forma que se actualizan los pesos tras un número de instancias establecido (*mini-batch gradient descent*).

Por otra parte, existe un hiperparámetro denominado **época** (*epoch*), que establece el número de iteraciones que va a realizar el algoritmo de aprendizaje a través de todo el conjunto de datos de entrenamiento. Una época comprende uno o más lotes. En el caso del *batch gradient descent*, una época tiene un solo lote puesto que se actualizan los pesos tras procesar todo el conjunto de datos, y en el caso del *stochastic gradient descent* una época contiene tantos lotes como ejemplos haya en el conjunto de datos puesto que actualiza los parámetros de la red al procesar una instancia.

Por otra parte, el grado en el que se actualizan los pesos es controlado por un parámetro de configuración denominado **velocidad de aprendizaje** (*learning rate*). Este parámetro controla el cambio realizado en el peso de la red neuronal para un error determinado. A menudo se utilizan tamaños de peso pequeños tales como 0.1, 0.01 o más pequeños.

Una vez que una red neuronal ha sido entrenada puede ser usada para realizar predicciones. La topología de la red neuronal y el conjunto final de pesos es todo lo que se necesita para implantar el modelo. Las predicciones se realizan proporcio-

nando la entrada a la red y ejecutando una propagación hacia delante que genera una salida que se utiliza como predicción.

### 2.2.5. $k$ -vecinos más cercanos

El paradigma clasificatorio denominado  **$k$ -Vecinos más Cercanos** (*K-Nearest Neighbours*, K-NN) se fundamenta en la idea de identificar el grupo de  $k$  objetos en el conjunto de datos de entrenamiento que más cerca está de un nuevo objeto a clasificar, de manera que a este nuevo caso se le asigna la etiqueta de clase más frecuente en ese grupo de  $k$  objetos. Este método de clasificación supervisada se basa en tres componentes principales: un *conjunto de datos de entrenamiento etiquetados*, una *métrica de distancia* para calcular las distancias entre distintos objetos y el número  $k$  de vecinos más cercanos. De esta forma, utilizando la métrica de distancia correspondiente, se calcula la distancia entre el nuevo caso a clasificar y los casos etiquetados para averiguar cuáles son las  $k$  instancias etiquetas que más próximas están al nuevo objeto y, una vez ejecutado este paso, se calcula la clase más repetida en el grupo de  $k$  casos más cercanos y se le asigna al nuevo objeto.

En la siguiente figura se presenta un pseudocódigo para el clasificador KNN básico:

---

**Algoritmo 1** Pseudocódigo del algoritmo KNN. Fuente: [Larranaga et al.](#)

---

COMIENZO

Entrada:  $D = \{(\mathbf{x}^{(1)}, c^{(1)}), \dots, (\mathbf{x}^{(N)}, c^{(N)})\}$

$\mathbf{x} = (x_1, \dots, x_n)$  nuevo caso a clasificar

PARA todo objeto ya clasificado  $(\mathbf{x}^{(i)}, c^{(i)})$

calcular  $d^{(i)} = d(\mathbf{x}^{(i)}, \mathbf{x})$

Ordenar  $d^{(i)} (i = 1, \dots, N)$  en orden ascendente

Quedarnos con los  $K$  casos  $D_{\mathbf{x}}^K$  ya clasificados ms cercanos a  $\mathbf{x}$

Asignar a  $\mathbf{x}$  la clase ms frecuente en  $D_{\mathbf{x}}^K$

FIN

---

donde:

- $D_{\mathbf{x}}^K$  es el conjunto de casos de tamaño  $K$  más cercano a la instancia  $\mathbf{x}$
- $d^{(i)}$  es la distancia entre el caso clasificado  $i$  y la nueva instancia a clasificar  $\mathbf{x}$

En este algoritmo puede ocurrir que, a la hora de averiguar la clase a la que se va a clasificar una nueva instancia, dos o más clases obtengan el mismo número de votos, por lo que hay que establecer *reglas de desempate*. Algunas de ellas pueden ser elegir aquella clase que tenga menor distancia media, aquella a la que pertenezca el vecino más cercano, etcétera ([Mohanty et al. \[2015\]](#)). Además, todos los datos deben estar *normalizados* para evitar que las características en el conjunto de entrada con valores más altos dominen el cálculo de la distancia.

El algoritmo KNN lleva a cabo un **aprendizaje vago** (*lazy learning*) puesto que no aprende un modelo discriminativo de los datos de entrenamiento (*eager learning*), sino que memoriza el conjunto de datos de entrenamiento y la labor de predicción se realiza cuando llega un nuevo caso a clasificar. Por lo tanto, en contraste con otros paradigmas clasificatorios, donde hay un proceso de construcción del modelo de predicción y posteriormente su aplicación sobre nuevas instancias, el algoritmo KNN engloba estos dos pasos en uno. Además, es un algoritmo **no paramétrico** puesto que no hace ninguna suposición sobre la distribución de datos subyacente en el conjunto de datos de entrenamiento. Un aspecto importante de este paradigma clasificatorio es el número  $K$  de vecinos que se va a utilizar para decidir la clase a la que pertenece una nueva instancia. Una de las opciones para establecer este número es hacerlo de forma *fija*, y se ha constatado empíricamente que la proporción de casos clasificados correctamente es no monótono con respecto a  $K$ , de manera que el rendimiento del clasificador no aumenta siempre al incrementar  $K$ ; un valor adecuado sería entre 3 y 7 vecinos (Larranaga et al.). No obstante, no es práctico asignar un valor fijo a este número para todos los nuevos casos a clasificar, sino modificarlo en función de las características de cada uno de ellos (Wang et al. [2006], Cheng et al. [2014], Zhang et al. [2018]).

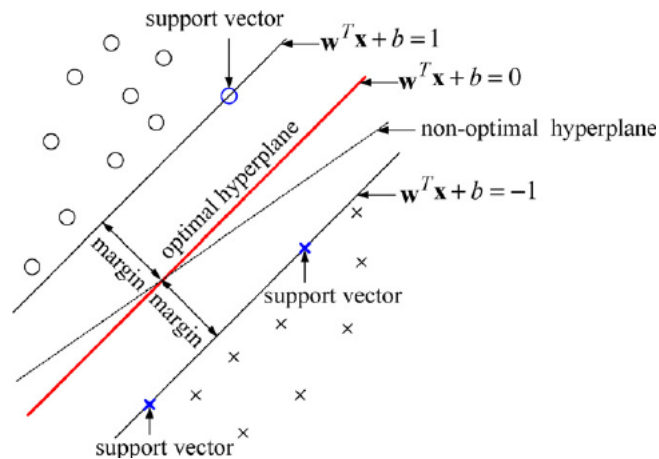
Con respecto al algoritmo básico del KNN, se han propuesto diferentes variantes con el objetivo de mejorar su rendimiento. Uno de ellos es el **KNN con rechazo**, en el que se tienen que cumplir una serie de condiciones que garanticen la clasificación del nuevo caso, como puede ser una mayoría absoluta de una determinada clase, o que supere un determinado umbral de votos. Otras variantes del KNN básico son **KNN con pesado de casos seleccionados**, en el que se le da más importancia a unas instancias que a otras a la hora de realizar la clasificación según la cercanía que tengan con el nuevo caso a predecir; **KNN con pesados de variables**, en el que se le da más relevancia a ciertas variables predictoras que a otras a la hora de calcular las distancias entre los casos clasificados y la nueva instancia a clasificar; etcétera.

### 2.2.6. Máquinas de vector soporte

Las **máquinas de vector soporte** (*Support Vector Machines*, SVMs) son un algoritmo de aprendizaje supervisado que puede utilizarse tanto para desafíos de clasificación como de regresión; no obstante, se utiliza principalmente en problemas de clasificación. En este algoritmo trazamos cada elemento de datos (cada muestra de entrenamiento) como un punto en el espacio  $n$ -dimensional (donde  $n$  es el número de características del problema a resolver), con el valor de cada característica mapeándose al valor de una determinada coordenada. A continuación, dado este conjunto de ejemplos de entrenamiento, cada uno perteneciente a una clase, entrenamos una SVM para construir un modelo que prediga la clase de una nueva muestra mediante la construcción de un **hiperplano** que separe las clases de los datos de entrenamiento y maximice el margen entre esas clases (maximice la distancia entre los puntos más cercanos de cada clase al hiperplano de separación óptimo, llamados

vectores soporte) en el espacio  $n$ -dimensional:

Figura 2.8: SVM. Fuente: Unagar and Unagar [2017]

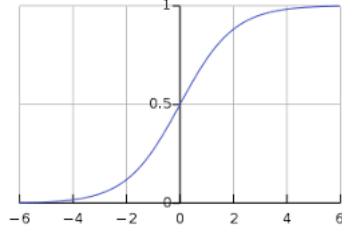


Las máquinas de vector soporte, cuando los datos de entrada no son *linealmente separables*, convierte esos datos a un **espacio de mayor dimensión**, de tal forma que en ese espacio el algoritmo puede que encuentre un hiperplano que sea capaz de separar los datos (*separación lineal*). A la hora de realizar este proceso de mapeo de los datos de entrada a un espacio de mayor dimensión, no es necesario calcular la transformación de cada uno de los puntos originales a dicho espacio, sino que solo es necesario calcular **el producto escalar de los vectores soporte en el espacio de mayor dimensión** puesto que es este cálculo lo que se necesita para encontrar el hiperplano que maximice el margen entre las distintas clases. El cálculo de este producto escalar de los puntos en una dimensión mayor es mucho más sencillo que convertir los puntos originales a dicho espacio a través de lo que se denomina el **kernel trick**. El *kernel trick* consiste en utilizar lo que se denomina una *función kernel*, que permite obtener el producto escalar entre puntos en un espacio de mayor dimensión sin necesitar la función de mapeo de un punto original a un punto de dimensión mayor.

### 2.2.7. Regresión logística

La **regresión logística** es un modelo estadístico que se utiliza en aprendizaje automático para describir las relaciones que existen entre un conjunto de variables predictoras y una variable clase con el objetivo de estimar la *probabilidad* de que una instancia pertenezca a una determinada clase. Se trata de un *clasificador binario* (predicción dicotómica).

El nombre del modelo procede de la función sobre la que se fundamenta, que recibe el nombre de **función logística** o **función sigmoide**:

Figura 2.9: Función logística o sigmoide. Fuente: [mis \[a\]](#)

Esta función tiene la característica de que mapea cualquier valor real a un número **comprendido entre 0 y 1** (sin llegar a dar como salida estos dos valores puesto que hay asíntotas horizontales en esos puntos), lo que permite obtener valores de probabilidad. De esta manera, si la probabilidad que estima el modelo es mayor que 0.5, entonces se predice que la instancia si pertenece a esa clase (se le asigna un 1, refiriéndose a la **clase positiva**) o, en caso contrario, que no pertenece (se le asigna un 0, refiriéndose a la **clase negativa**). La fórmula matemática de esta función es la siguiente:

$$f_{\theta}(\mathbf{x}) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}} \quad (2.12)$$

donde  $\theta^T \mathbf{x}$  es una función lineal de una instancia de entrada  $\mathbf{x}$ :

$$\theta^T \mathbf{x} = \theta_0 + \sum_{i=1}^n \theta_i X_i \quad (2.13)$$

donde  $\theta^T$  es el vector de parámetros que define el modelo de *regresión lineal*. Es decir, la *regresión logística* se fundamenta en la idea de que el modelo de *regresión lineal* no se puede utilizar para tareas de clasificación puesto que produce como salida un rango infinito de valores. Esto es adecuado para resolver un problema de regresión, pero en este caso el interés reside en solventar un problema de clasificación, donde la variable clase a predecir debe tomar **valores discretos**, por lo que surge el modelo de regresión logística con el objetivo de *mapear la salida de un modelo de regresión lineal a una probabilidad de pertenencia a una determinada clase*.

Cuando la clase a predecir puede tomar más de dos valores discretos, se utiliza un modelo denominado *regresión logística multinomial*, que es una extensión de la regresión logística binaria. La fórmula matemática es la que se detalla a continuación:

$$f_{\theta_i}(\mathbf{x}) = \frac{e^{-\theta^{(i)T} \mathbf{x}}}{\sum_{j=1}^k e^{-\theta^{(j)T} \mathbf{x}}} \quad (2.14)$$

donde  $f_{\theta_i}(\mathbf{x})$  en este caso es la probabilidad de que la instancia pertenezca a la clase  $i$  y  $\theta^{(i)T}$  son los parámetros de la clase  $i$ . Esta fórmula recibe el nombre de **función softmax**, por lo que la regresión logística multinomial también recibe el nombre de *regresión softmax*. Este clasificador predice la clase que reciba el mayor valor de probabilidad.

La clasificación multinomial con el modelo de regresión logística también se puede abordar mediante la aproximación **one vs all**. En este método, si hay  $k$  clases posibles para predecir, se crean  **$k-1$  modelos de regresión logística binaria**, de tal forma que, a la hora de aprender cada uno de ellos, se elige una determinada clase (*clase positiva*) y las instancias que no pertenezcan a esa clase se agrupan en una segunda clase (*clase negativa*) con el objetivo de construir el modelo de regresión logística binaria pertinente. La predicción se realiza de la misma forma que en la *regresión logística multinomial* y, con respecto a la utilización de la aproximación *one vs all* o de la *regresión logística multinomial*, no existe una superioridad notoria de uno de ellos sobre el otro y depende de la cuestión que se quiera afrontar.

Algunas de las ventajas de la utilización del modelo de regresión logística como paradigma clasificatorio es que es **eficiente**, es capaz de interpretar los parámetros del modelo como **indicadores de la importancia de las características**, **no es necesario que se escalen los atributos de entrada** y se puede **regularizar** para evitar su sobreajuste sobre los datos. Además, a la hora de clasificar proporciona probabilidades, por lo que obtenemos **más información acerca de cuánta es la confianza con la que el clasificador predice una determinada clase**. No obstante, una desventaja importante de la regresión logística es que **no es capaz de resolver problemas no lineales** puesto que la región de decisión que da como salida es lineal.

### 2.2.8. Métodos combinados de aprendizaje

Con el objetivo de obtener un mejor rendimiento a la hora de realizar labores de predicción, existen lo que se denominan **métodos combinados de aprendizaje** (*ensemble methods*). Los *métodos combinados de aprendizaje* son una técnica de aprendizaje automático que están constituidos por un *conjunto de paradigmas clasificatorios*, como los clasificadores bayesianos, de tal forma que las predicciones que lleve a cabo cada uno de ellos se combinen para clasificar una nueva instancia (se agregan para forman un solo clasificador); por ello reciben el nombre de *meta-algoritmos*. De esta forma, no se lleva a cabo un aprendizaje de un solo clasificador, sino de un conjunto de ellos. En general, esta técnica tiene un mejor desempeño que los modelos de clasificación por separado y, para que esto suceda, los clasificadores de los que se compone este meta-algoritmo tienen que ser *precisos* (que tengan una mejor tasa de error que la realización de una predicción aleatoria) y *variados* (que cometan diferentes errores al clasificar nuevas instancias) (Dietterich [2000]).

Algunas de las ventajas que presenta la utilización de *métodos combinados de aprendizaje* es que reducen las probabilidad de que haya un **sobreajuste a los datos durante la fase de entrenamiento** y disminuyen tanto el error de **varianza** (los resultados que proporcione el meta-algoritmo dependerán menos de las peculiaridades de los datos de entrenamiento) y de **bias** (al combinar clasificadores, se aprenden mejor particularidades del conjunto de datos de entrenamiento). Todo esto se debe a que, si se constituye una combinación de clasificadores *variados* a partir de un conjunto de instancias de entrenamiento, éstos pueden proporcionar *información*



*complementaria* con respecto a los patrones que subyacen a los datos y, por tanto, una mayor precisión a la hora de clasificar nuevos ejemplos. No obstante, debido a la complejidad de construcción de esta técnica de aprendizaje, los (métodos combinados de aprendizaje) tienen el inconveniente de que **aumenta su tiempo de procesamiento** puesto que no entrenan un solo clasificador, sino muchos.

De los diferentes métodos de *ensemble* que existen, los más conocidos son la **agregación Bootstrap** (*Bagging*, Breiman [1996]), **Boosting** (Freund and Schapire [1996]) y **Stacking** (Wolpert [1992]). El método *Bagging* se basa en entrenar cada uno de los clasificadores que componen el meta-clasificador sobre un conjunto de datos del **mismo tamaño que el conjunto de datos original de entrenamiento**, que se obtiene en cada caso escogiendo  $N$  instancias del conjunto de instancias original mediante un **muestreo uniforme y con reemplazamiento de  $D$** . De esta manera, cada clasificador individual se entrena sobre una muestra de datos *distinta* y habrá instancias del conjunto original que estarán *repetidas en dichas muestras*. La efectividad de este método se fundamenta en clasificadores individuales que sean *inestables*, es decir, aquellos cuyo aprendizaje sobre conjuntos de datos de entrenamiento ligeramente distintos realicen predicciones con grandes diferencias, como son los árboles de decisión.

Por otra parte, el método *Boosting* se basa en construir un meta-clasificador de forma **incremental**. En este sentido, en este método de *ensemble* se crea una sucesión de clasificadores individuales, donde cada uno de ellos se va a entrenar sobre un conjunto de datos de entrenamiento que va a estar determinado por **los ejemplos mal clasificados por los clasificadores individuales previos**. Es decir, al utilizar un clasificador individual para clasificar nuevas instancias, aquellas cuya clase se prediga erróneamente, a la hora de construir el conjunto de ejemplos para entrenar un nuevo clasificador individual, serán elegidas más frecuentemente con el objetivo de centrarse en clasificarlas bien en el siguiente paso. El algoritmo de *boosting* más conocido y exitoso es el denominado **AdaBoost** (*Adaptive Boosting*).

En comparación con el método *Bagging*, se asemejan en que combinan clasificadores del *mismo tipo* (métodos de *ensemble* homogéneos), utilizan un *sistema de votos* para realizar las predicciones y utilizan el mismo método para *muestrear los datos de entrenamiento de los clasificadores individuales*. No obstante, una de las diferencias del método *Boosting* con respecto al *Bagging* es que el primero muestrea del conjunto de datos original teniendo en cuenta el **rendimiento del clasificador individual previo** (las instancias tienen distinta probabilidad de ser elegidas del conjunto de datos original), mientras que en el segundo no (las instancias tienen la misma probabilidad de ser elegidas), por lo que en el primero los clasificadores individuales son *dependientes entre ellos* y en el segundo no.

Por otra parte, en el *Bagging* la clasificación se realiza sumando las predicciones individuales y eligiendo la clase más votada, mientras que en el *Boosting* el sistema de votos es **ponderado**, por lo que la predicción de cada uno de los clasificadores individuales no pesa igual en la decisión final. También se distinguen en que el método *Boosting* se centra en reducir el **bias**, es decir, en intentar incrementar la complejidad de los modelos que no son capaces de adaptarse a los datos (*underfit-*



ting), mientras que el método *Bagging* se enfoca en reducir la **varianza**, es decir, en reducir la complejidad de los modelos que se ajustan demasiado a los datos de entrenamiento (*overfitting*).

Otro método de *ensemble* popular es el denominado *Stacking*. Esta técnica de *ensemble* combina diferentes tipos de clasificadores base en un primer nivel (es un método *heterogéneo*, a diferencia del *Bagging* y del *Boosting*), de tal forma que las predicciones que realizan cada uno de ellos se utilizan como **atributos de entrada para entrenar un meta-clasificador** en un segundo nivel con el objetivo de que éste lleve a cabo la decisión final. Al llevar a cabo el entrenamiento sobre las predicciones de varios tipos de clasificadores, el meta-algoritmo puede saber en cuáles de ellos puede **confiar mayoritariamente**, de manera que aprende cuáles son los patrones que subyacen los valores de sus predicciones con el objetivo de mejorar el desempeño del meta-modelo. Así, una característica importante que distingue este método del *Bagging* y del *Boosting* es que estos dos últimos utilizan un sistema de **votación** para saber qué clase ha sido la más predicha por los clasificadores base y no hay un aprendizaje en el meta-nivel, mientras que en el *Stacking* si se realiza ese meta-aprendizaje.

## 2.3. Algoritmos de aprendizaje no supervisado

Los algoritmos de aprendizaje automático vistos anteriormente asumen que, para cada una de las instancias de entrada que utilizan en la fase de entrenamiento, tienen a su disposición la etiqueta clase a la que pertenecen. Sin embargo, en muchas aplicaciones del mundo real sucede que no tenemos conocimiento de la categoría a la que pertenecen cada una de las instancias (*conjunto de datos no etiquetados*), ya sea porque la obtención de las etiquetas resulte caro, sea propenso a errores o directamente sea imposible su adquisición. En este caso, con el objetivo de aplicar aprendizaje automático en esos datos, es necesario recurrir a algoritmos pertenecientes a la categoría de **aprendizaje no supervisado**. La finalidad de este tipo de aprendizaje es *descubrir las diferentes categorías que describen las características de los datos no etiquetados*. A continuación se expone la estructura de datos con la que trabajan comúnmente los algoritmos de aprendizaje no supervisado

Tabla 2.3: Estructura de los datos en un problema de clasificación no supervisado

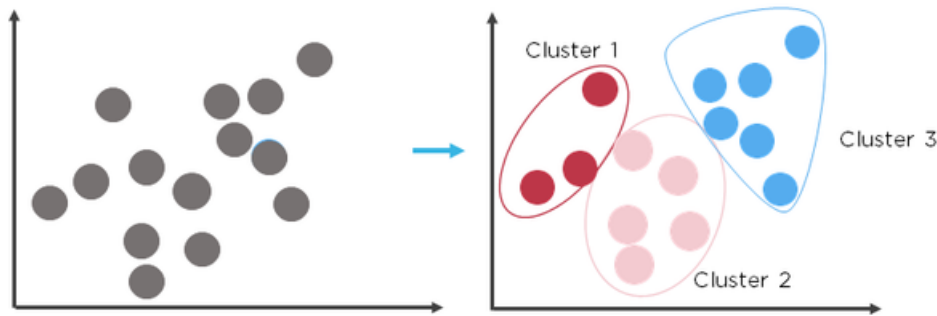
	$X_1$	$\dots$	$X_i$	$\dots$	$X_n$
$\mathbf{x}^{(1)}$	$x_1^{(1)}$	$\dots$	$x_i^{(1)}$	$\dots$	$x_n^{(1)}$
$\mathbf{x}^{(j)}$	$x_1^{(j)}$	$\dots$	$x_i^{(j)}$	$\dots$	$x_n^{(j)}$
$\mathbf{x}^{(N)}$	$x_1^{(N)}$	$\dots$	$x_i^{(N)}$	$\dots$	$x_n^{(N)}$

Uno de los algoritmos de clasificación no supervisada que más se utilizan es el denominado **agrupamiento** (*clustering*). Debido a la gran popularidad que presenta, en este nos centraremos en abordar una multitud de propuestas relacionadas con este algoritmo para flujos de datos.

### 2.3.1. Agrupamiento

El **agrupamiento** (*clustering*) es una técnica que consiste en particionar un conjunto de objetos de entrada en una serie de grupos o *clusters*, de tal forma que los objetos que se sitúan dentro de un grupo sean **muy similares** y que haya una **heterogeneidad alta** entre objetos de distintos grupos. Un ejemplo del resultado de esta técnica es el siguiente:

Figura 2.10: Problema de agrupamiento. Fuente: [mis](#) [b]



Existen diferentes tipos de métodos para abordar el problema del agrupamiento de objetos. No obstante, en este trabajo, nos vamos a centrar en aproximaciones de *clustering* para flujos de datos basadas en ***clustering* particional**, ***clustering* jerárquico** y ***clustering* probabilístico**. En el *clustering* particional el objetivo es dividir el conjunto de instancias de entrada en un número  $k$  de *clusters*, de tal forma que cada uno de los objetos pertenezca a un determinado grupo y los patrones que subyacen cada uno de los grupos sean similares dentro de los mismos y distintos entre diferentes *clusters*. El algoritmo de *clustering* particional más conocido es el denominado **k-medias** (*k-means*). La versión estándar de este algoritmo, propuesto por Stuart Lloyd en 1957 ([Lloyd \[1957\]](#)), aunque no publicado en una revista hasta 1982, es el siguiente ([Lloyd \[1982\]](#)):

---

**Algoritmo 2** Pseudocódigo del algoritmo k-medias estándar

---

- Paso 1: Crear un agrupamiento inicial de los objetos en  $k$ , cada uno representado mediante un *centroide*.
  - Paso 2: Calcular la distancia de cada uno de los objetos a los centroides para asignarlos al centroide ms cercano.
  - Paso 3: Calcular los  $k$  nuevos centroides de los nuevos grupos construidos tras la asignación.
  - Paso 4: Repetir desde el Paso 2 hasta que se cumpla una condición de parada.
- 

En primer lugar, este algoritmo elige un conjunto de  $k$  objetos inicial del conjunto de datos; esta elección puede realizarse de forma **aleatoria**, escogiendo los **primeros  $k$  objetos** del fichero, mediante una **heurística** que permita que los  $k$  objetos estén lo más alejados posibles, etcétera. A continuación, cada uno de los objetos se asigna

al *cluster* cuyo representante (*centroide*) se encuentre más cerca de esos objetos; para ello, es necesario utilizar una medida de distancia que, en la versión original del algoritmo, es la *distancia Euclídea*. Tras esto, se recalculan los centroides de los nuevos grupos construidos computando la *media* de los objetos incluidos en cada uno de los *clusters*. Todos estos pasos se vuelven a repetir hasta que se alcanza un criterio de convergencia que, comúnmente, suele ser cuando las asignaciones de los objetos a los distintos grupos no cambia de una iteración a otra. De esta manera, el objetivo de este proceso de agrupamiento es minimizar las distancias de los objetos de cada *cluster* al centroide del mismo (Alex Smola [2008]):

$$J(r, \mu) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^k r_{ij} \|x_i - \mu_j\|^2 \quad (2.15)$$

donde  $\mu_j$  representa el centroide del grupo  $k$  y  $r_{ij}$  es una variable que indica si el objeto  $i$  pertenece a la clase  $j$  (se le asigna el valor 1) o no (se le asigna el valor 0). El centroide  $\mu_j$ , que corresponde a la media de todos los objetos del grupo  $j$ , se calcula de la siguiente forma:

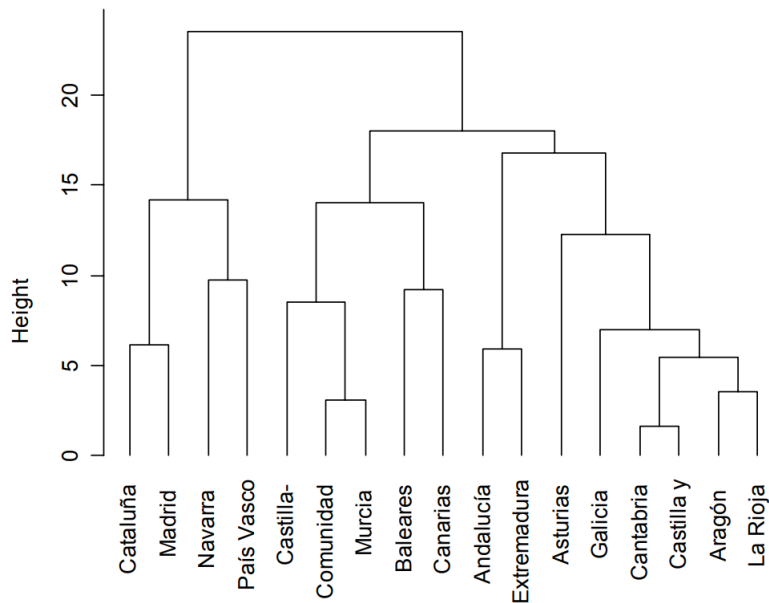
$$\mu_j = \frac{\sum_i r_{ij} x_i}{\sum_i r_{ij}} \quad (2.16)$$

donde el denominador corresponde al número de objetos asignados al *cluster*  $j$ .

Esta versión del algoritmo *k-medias* también fue publicado en 1965 por E.W. Forgy (Forgy [1965]), por lo que en ocasiones se denomina algoritmo de *Lloyd-Forgy* y una característica importante de esta versión es que los centroides se actualizan tras realizar **todas las asignaciones de los objetos a los diferentes grupos**. En este sentido, J. MacQueen propuso en 1967 (MacQueen [1967]) un algoritmo *k-means* en el que considera los primeros  $k$  objetos del fichero como los  $k$  grupos iniciales y la actualización de los centroides no se realiza tras llevar a cabo todas las asignaciones de los objetos a los distintos grupos, sino que, **cada vez que se asigna un objeto a un *cluster*, se recalcula el centroide de ese grupo**. El algoritmo de MacQueen es el método de *clustering* particional que más se utiliza.

Otro tipo de método utilizado en *clustering* es el ***clustering* jerárquico**, cuyo objetivo es agrupar los objetos de entrada en una estructura de árbol jerárquico denominada *dendograma* (Figura 2.11), de tal forma que los nuevos *clusters* que se construyan dependen de los creados previamente.

En función de como se genere esta estructura, el *clustering* jerárquico puede ser **aglomerativo** (*bottom-up*) y **divisivo** (*top-down*). En el *aglomerativo* se parte de tantos *clusters* como instancias haya en el conjunto de datos y se van agrupando por pares aquellos grupos que más cerca se encuentren. De forma contraria, el *divisivo* parte de un solo grupo con todos los objetos y se va dividiendo en grupos más pequeños hasta tener tantos *clusters* como instancias haya en el fichero de datos. De esta manera, el *aglomerativo* tiene un buen desempeño a la hora de identificar pequeños *clusters* y el *divisivo* en identificar grandes *clusters*.

Figura 2.11: Representación de un dendograma. Fuente: [mis](#) [c]

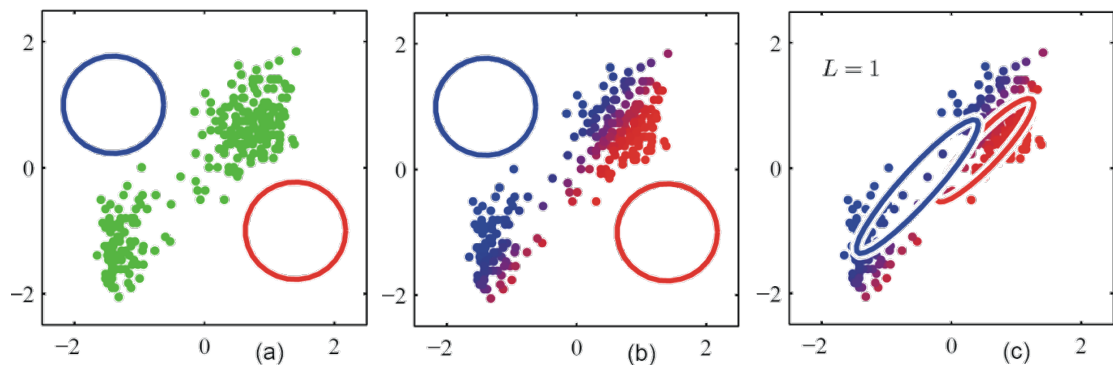
La estructura que genera ese tipo de *clustering* aporta más información que la partición realizada por el *clustering* particional puesto que el dendograma permite obtener diferentes números de grupos según a qué altura se desee cortar la estructura, donde la altura representa la **distancia entre los grupos formados**. De esta forma, con el *clustering* jerárquico **no es necesario especificar de antemano un número  $k$  de grupos**, a diferencia del *clustering* particional, aunque es **computacionalmente más costoso**. Además, en el *clustering* jerárquico **no se permiten reasignaciones de los objetos** a otros clusters, característica que sí posee el *clustering* particional.

Otro método de *clustering* conocido es el ***clustering* probabilístico**. En este tipo de *clustering* se asume que las instancias del conjunto de datos son generadas por una *mixtura de distribuciones de probabilidad*, donde cada una de las componentes que la forman representa un *cluster*, a diferencia de los otros tipos de *clustering* comentados anteriormente, que se basan en construir *clusters* a través de la optimización de criterios que se fundamentan en la distancia entre los objetos de entrada. De esta forma, las instancias no pertenecen a un solo *cluster* (*hard assignment*), como ocurre en el *clustering* particional y jerárquico, sino que considera la incertidumbre presente a la hora de asignar los objetos a los diferentes grupos mediante "**asignaciones suaves**" (*soft assignments*), de tal manera que estas asignaciones no se realizan de forma determinística, sino **probabilística**. Es decir, los objetos pertenecen a cada *cluster* con una determinada probabilidad, que viene establecida por la distribución de probabilidad que define a cada uno de los *clusters*.

De forma general, se supone que las distribuciones que representan a cada uno de los *clusters* son distribuciones **Gaussianas**. Para hallar los parámetros de cada una de las distribuciones, se utiliza el método de **estimación de máxima verosimilitud**, cuyo objetivo es maximizar la probabilidad de pertenencia de los distintos

objetos a cada uno de los *clusters* (maximizar la verosimilitud de los datos). A la hora de hallar las estimaciones de máxima verosimilitud, surge el problema de que este proceso de optimización no tiene una *solución analítica cerrada*. De esta manera, uno de los métodos más conocidos para hallar las estimaciones de máxima verosimilitud de los parámetros desconocidos de las distribuciones que representan a los distintos *clusters* se denomina **algoritmo EM** (Dempster et al. [1977]). Este método comienza estableciendo unos valores iniciales de los parámetros de las distribuciones (en el caso de las distribuciones Gaussianas la *media*, la *matriz de covarianzas* y la *proporción de puntos asignados a cada uno de los clusters*) e iterativamente alterna entre un paso de **Esperanza** (*E*) y otro paso de **Maximización** (*M*). El paso *E* consiste en, para cada uno de los objetos de entrada, utilizar los parámetros actuales para averiguar sus probabilidades de pertenencia a cada uno de los *clusters* (denominadas **responsabilidades**) y, una vez realizado esto, en el paso *M* se reestiman los parámetros de las distribuciones que maximizan la verosimilitud de los datos a partir de las probabilidades calculadas. A continuación se expone una ilustración de los pasos del algoritmo en una iteración:

Figura 2.12: Iteración del algoritmo EM. Fuente: Larrañaga and c. Bielza



Con respecto al algoritmo *k-means* visto con anterioridad, es un tipo particular de *clustering* probabilístico en el que las distribuciones de los *clusters* se suponen que siguen una distribución Gaussiana pero cuyas matrices de covarianzas son iguales y constantes (los *clusters* son círculos con un radio fijo). Entre las desventajas del *clustering* probabilístico, la principal es que **depende mucho de la distribución de probabilidad escogida**, por lo que la elección de un tipo u otro influye de forma notoria en el rendimiento del mismo.

## 2.4. Series temporales

## 2.5. Redes bayesianas para el descubrimiento de conocimiento



## Capítulo 3

# Estado del arte: Aprendizaje automático para flujos de datos

### 3.1. Introducción

En la actualidad, existen numerosas aplicaciones que constantemente están generando una cantidad inmensa de información. Entre los dominios donde se encuentran implantadas esas aplicaciones, se encuentran los *sistemas de monitorización de tráfico de red*, *redes de sensores para el control de los procesos de fabricación*, *gestión de redes de telecomunicaciones*, *modelado de usuarios en una red social*, *minería web*, *transacciones bancarias*, etcétera. Las técnicas tradicionales de minería de datos se enfocan en encontrar conocimiento en datos en **repositorios estáticos de datos**; no obstante, debido a las propiedades inherentes en los datos originados por las aplicaciones mencionadas anteriormente, este tipo de técnicas no se pueden aplicar en dichos datos.

En primer lugar, **no es factible ni tampoco práctico guardar tanta información en bases de datos** puesto que estos almacenes de datos utilizados por las técnicas comunes de aprendizaje automático suelen tener un *tamaño fijo*, pero la naturaleza de los datos generados continuamente implica que la cantidad de información originada puede llevar a ser *infinita*, característica inabordable por los repositorios de información tradicionales, sobre todo a la hora de entrenar los modelos cuyos datos de entrenamiento deben estar en la **memoria principal**, que posee poca capacidad de almacenamiento. Por otra parte, las aplicaciones comentadas previamente generan información a una **gran velocidad** y, a diferencia de los algoritmos de aprendizaje automático habituales que construyen *modelos estáticos* a partir de *datos fijos*, los patrones que subyacen dicha información puede **cambiar dinámicamente** durante el transcurso del tiempo debido al **entorno no estacionario** en el que se originan, por lo que es necesario que las técnicas de aprendizaje automático sean capaces de construir modelos de forma continua que se adapten a dichos cambios con el objetivo de que mantengan un buen rendimiento.

En adición a lo comentado previamente, los algoritmos de aprendizaje automático tradicionales tienen a su disposición la posibilidad de analizar *múltiples veces* el

conjunto de datos estático, pero debido a los problemas de almacenamiento de los datos generados actualmente por las aplicaciones del mundo real y a las propiedades de los mismos, **no es abordable realizar múltiples escaneos del conjunto de datos**. Los modelos generados por los algoritmos de aprendizaje automático deben estar actualizados a medida que se van originando nuevos datos para que ofrezcan un buen desempeño, y esto no se puede llevar a cabo si, a la hora de entrenarlos, realizamos varios ciclos de lectura de los datos.

En base a todo lo comentado con anterioridad, los sistemas modernos de aprendizaje automático deben tener en cuenta la rapidez y la continuidad con la que se generan los datos hoy en día. Dada las propiedades de estos datos, éstos reciben el nombre de **flujos de datos** y, en base a la importancia de extraer conocimiento de este tipo de datos, en los últimos años se han realizado una gran cantidad de investigaciones en el campo del **aprendizaje automático aplicada a los flujos de datos**. A la hora de desarrollar algoritmos de aprendizaje automático para manejar flujos de datos, teniendo en cuenta los problemas que ostentan los algoritmos tradicionales, deben asumir una serie de desafíos y restricciones:

- Las instancias de entrada del flujo de datos **deben ser procesadas una sola vez** (son descartadas después de ser procesadas), aunque el algoritmo puede recordar instancias pertenecientes a un corto plazo de tiempo.
- No hay un control sobre el **orden en el que los objetos de datos deben ser procesados**.
- El tamaño de un flujo de datos se debe suponer que es **ilimitado**.
- El proceso responsable de generar el flujo de datos puede ser **no estacionario**, es decir, la distribución de probabilidad que subyace los datos puede cambiar durante el transcurso del tiempo.
- La memoria utilizada por los algoritmos es **limitada**.
- El trabajo realizado por los algoritmos debe cumplir unas **restricciones estrictas de tiempo**.
- El modelo inducido por los algoritmos deben poder llevar a cabo tareas de predicción **en cualquier momento**.
- Pueden ocurrir, al igual que en las tareas de clasificación, problemas relacionados con *valores faltantes, sobreajuste del modelo, características irrelevantes, desbalanceo de las clases, aparición de nuevas clases*

En el campo de investigación relacionado con el aprendizaje para flujos de datos, al abordarse el problema de extracción del conocimiento desde una perspectiva diferente a las técnicas de aprendizaje automático tradicionales, surge una terminología característica del mismo. Para comprender las propuestas que se van abordar en este trabajo relacionadas con este campo, en el siguiente apartado se va a proceder a la descripción de diferentes conceptos englobados dentro de la terminología vinculada al aprendizaje automático para flujos de datos.



### 3.1.1. Conceptos

A la hora de tratar flujos de datos para realizar tareas de clasificación, existen una serie de desafíos, que han sido comentados previamente. Para hacer frente a estos problemas que pueden surgir de la aplicación de aprendizaje automático en flujos de datos, existen tres aproximaciones principales (Krawczyk and Wozniak [2015]):

- **Entrenar un clasificador cada vez que se disponga de nuevos datos.** Esta opción suele ser poco adoptada debido a que tiene altos costes computacionales.
- **Detectar cambios en los patrones de los datos** (*concept drifts*), de manera que si son relevantes, se vuelve a entrenar el modelo sobre los nuevos datos tras la ocurrencia del *concept drift*.
- **Llevar a cabo un aprendizaje incremental** con el objetivo de adaptar el modelo a los cambios en el concepto subyacente de los datos de forma gradual

Para entender como funcionan estos algoritmos, es imprescindible tener una idea general de las nociones sobre las que se basan.

#### Concept drift

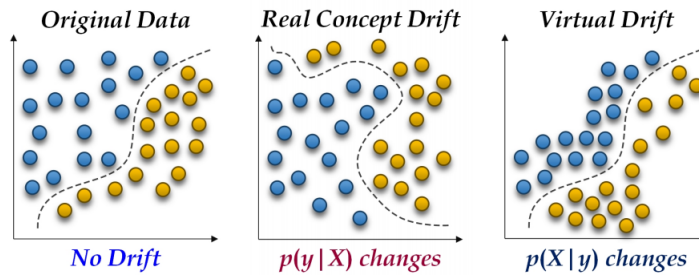
Uno de los desafíos comentados anteriormente al que deben hacer frente los paradigmas de aprendizaje automático para lidiar con flujos de datos es que la distribución que subyace los datos puede cambiar durante el transcurso del tiempo (la distribución de los datos es *no estacionaria*). Este fenómeno se denomina **concept drift**, de tal forma que la palabra *concept* se refiere al concepto que describe y está inherente en los datos.

Formalmente, el fenómeno de *concept drift* se presenta cuando se producen cambios en la **probabilidad conjunta de las variables predictoras y de la clase que se quiere predecir**, es decir,  $P(\mathbf{X}, C)$ . Para estimar esta probabilidad, se utiliza la probabilidad a priori de la clase,  $P(C)$ , y la probabilidad de las variables predictoras condicionada a la variable clase,  $P(\mathbf{X}|C)$ , de tal forma que  $P(\mathbf{X}, C) = P(C)P(\mathbf{X}|C)$ . A partir de esta estimación de la probabilidad conjunta y utilizando la regla de Bayes, podemos obtener la probabilidad de la clase condicionada a las variables predictoras,  $P(C|\mathbf{X})$ .

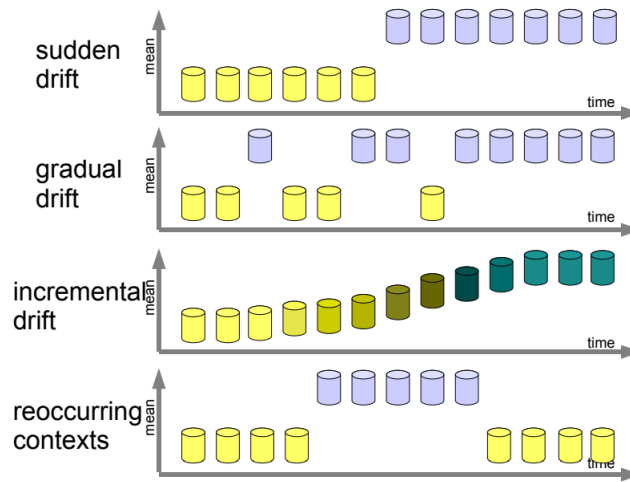
Teniendo en cuenta los términos probabilísticos mencionados anteriormente, existen dos tipos de *concept drift* según en cuál de ellos se produzca un cambio: **real concept drift**, **virtual concept drift** y **class prior concept drift** (Khamassi et al. [2016]). El primer tipo de *concept drift* se refiere a cambios que tienen lugar en la **probabilidad**  $P(C|\mathbf{X})$ , de manera que los límites de decisión para clasificar una instancia a una determinada clase. Con respecto al segundo tipo, sucede cuando se produce un cambio en la probabilidad conjunta de las variables predictoras  $P(\mathbf{X})$  y, por lo tanto, en la **probabilidad**  $P(\mathbf{X}|C)$ , pero no en la probabilidad a posteriori de la clase  $P(C|\mathbf{X})$ , de manera que esto implica que los límites de decisión de clasificación de una instancia a una clase en concreto no se ven afectadas. En cuanto

al tercer tipo de *concept drift*, se refiere a cambios que afectan a la **probabilidad a priori de la clase**  $P(C)$ , y comúnmente, según el comportamiento del cambio que se produce en dicha probabilidad, se ha clasificado este tipo como *real concept drift* o *virtual concept drift*, aunque es de gran relevancia tenerlo en cuenta como un tipo de *concept drift* aparte. Además, estos tipos de *concept drift* se pueden dar simultáneamente. A continuación se ilustran los dos tipos principales de *concept drift*:

Figura 3.1: Real vs Virtual concept drift. Fuente: [Pesaranghader et al. \[2018\]](#)



Otra categorización que se aplica con respecto a los *concept drifts* es en función del **ritmo** con el que ocurren. De esta forma, los *concept drifts* se pueden producir principalmente de manera **abrupta**, **gradual** o **recurrente**. Un *concept drift abrupto* se produce cuando, en cualquier momento, *ocurre* un *concept drift* de forma repentina, de manera que degrada el desempeño del modelo ya que el concepto subyacente de los datos cambia completamente. En cambio, un *concept drift gradual* tiene lugar cuando el fenómeno de *concept drift* va apareciendo de forma paulatina. El *concept drift* se puede presentar de dos maneras distintas; puede ocurrir que tanto el concepto antiguo como el nuevo estén activos, cada uno con una probabilidad de aparición asociada (los conceptos se alternan), predominando inicialmente el primero y, con el tiempo, desapareciendo con la presencia total del nuevo concepto (*gradual concept drift*, también denominado *gradual probabilistic drift*); por otro lado, el concepto antiguo puede ir sufriendo pequeñas modificaciones hasta la presencia completa del nuevo concepto, de tal manera que esos cambios son sutiles y solo se detectan en un intervalo de tiempo extenso (*incremental concept drift*, también denominado *gradual continuous drift*). En cuanto al *concept drift* recurrente, ocurre cuando conceptos que estuvieron presentes en el pasado vuelven a reaparecer, pudiendo ser *cíclico* si tienen lugar con cierta periodicidad, o *acíclico* si no posee la propiedad de periodicidad. La recurrencia del *concept drift* se puede dar de forma gradual o abrupta. En la figura siguiente se exponen dichos tipos de *concept drifts*:

Figura 3.2: Tipos de concept drift según el ritmo de cambio. Fuente: [Zliobaite \[2010\]](#)

Existen otras categorizaciones de *concept drifts*, siendo una si éstos ocurren de forma **local** (si los cambios del concepto de los datos ocurre en algunas región del espacio de entrada) o **global** (si los cambios del concepto de los datos ocurre en todo el espacio de entrada). Por otra parte, teniendo en cuenta la predictibilidad de los *concept drifts*, éstos se pueden clasificar en **predecibles** (si siguen un patrón) o **impredecibles** (si son totalmente aleatorios).

### Modelos de ventanas

Uno de los conceptos más recurrentes en la terminología de los algoritmos de aprendizaje automático aplicados a flujos de datos son las **ventanas deslizantes**. El objetivo de los algoritmos que se basan en *ventanas deslizantes* es **manejar los *concept drifts***, y se fundamentan en la idea de que las instancias más recientes del flujo de datos tienen *mayor relevancia a la hora de describir la distribución de probabilidad actual* que subyace los datos. Con respecto a este método, existen tres modelos utilizados frecuentemente: el modelo **landmark window**, el modelo **damping window** y el modelo **sliding window** ([Zhu and Shasha \[2002\]](#)).

El modelo *landmark window* se basa en utilizar **toda la historia del flujo de datos desde un punto de inicio en el pasado denominado *landmark* hasta el instante de tiempo actual**, de tal forma que los datos que se encuentren antes del *landmark* no se tienen en cuenta. De esta manera, el *landmark* se mantiene fijo, pero el punto que representa el instante de tiempo actual se va desplazando, por lo que el tamaño de la ventana va aumentando y se van teniendo en cuenta más datos. Un caso particular de este modelo es cuando el *landmark* se establece en el instante de tiempo del origen del flujo de datos, por lo que el modelo tiene en cuenta todo el flujo de datos generado hasta el momento actual. El problema que tiene el modelo *landmark window* es que es **difícil establecer el *landmark* idóneo** y todos los instantes de tiempo posteriores al punto inicial tienen la **misma importancia** a la

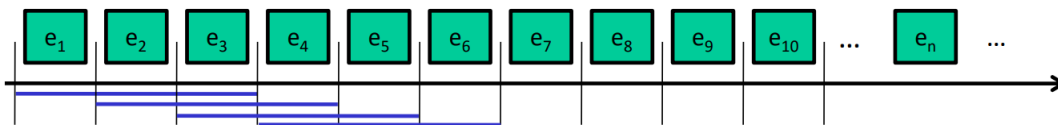
hora de construir el modelo.

Figura 3.3: Modelo *landmark window*. Fuente: E. Ntoutsis and Zimek [2015]



Por otra parte, en el modelo *sliding window* solo se tiene en cuenta la información **más reciente del flujo de datos** (desde el instante de tiempo actual hasta un instante del tiempo en el pasado). Esta información está definida por una *ventana temporal* cuyo tamaño define la cantidad de datos que van a ser relevantes para la construcción del modelo. De esta manera, la primera ventana del flujo de datos cubre el primer conjunto de datos que se van a utilizar para el entrenamiento del modelo y, cuando llega el siguiente instante de tiempo, la ventana se desplaza una unidad en el tiempo y se elimina de la misma la instancia de datos más antigua para mantener el tamaño de la ventana. Así, este proceso se repite a medida que va avanzando el tiempo. Esta ventana puede ser de tamaño *fijo* o *variable*, y la ventaja principal de ese modelo de ventana es que *evita que datos obsoletos* influyan en el proceso de generación del modelo de aprendizaje automático

Figura 3.4: Modelo *sliding window*. Fuente: E. Ntoutsis and Zimek [2015]



Con respecto al modelo *damped window*, al igual que el modelo *sliding window*, considera la información más reciente como relevante para la construcción del modelo de aprendizaje automático, pero en este caso se asignan una **serie de pesos a los datos en función del instante de tiempo en el que se han generado**. De esta forma, aquellas instancias que son más recientes van a recibir un *peso mayor* que aquellas que provienen de instantes de tiempo anteriores, por lo que van a influir más en la construcción del algoritmo de aprendizaje automático. Este modelo de ventana *no descarta instancias completamente*, sino que asigna pesos pequeños a los objetos antiguos, y para controlar el decrecimiento de los pesos a medida que se vuelve hacia atrás en el tiempo existe lo que se denomina un **factor de desvanecimiento** ( $\lambda$ ), de tal manera que, cuanto mayor es su valor, menor importancia tienen los datos del pasado.

Figura 3.5: Modelo *damped window*. Efecto del valor del factor de desvanecimiento  $\lambda$ . Fuente: E. Ntoutsis and Zimek [2015]



Con respecto a los tres modelos de ventana comentados anteriormente, el modelo *landmark window* se puede transformar en el modelo *damping window* añadiendo pesos de influencia de los datos sobre la construcción del modelo de aprendizaje automático, y el modelo *landmark window* se puede convertir al modelo *sliding window* realizando todo el proceso de construcción dentro de una ventana temporal. Por otra parte, existe también otro modelo de ventana denominado **tilted window**, que también se utiliza bastante y que consiste en guardar una síntesis del flujo de datos con memoria limitada en **diferentes niveles de granularidad temporal**. De esta forma, la información resumida de aquellos datos que son más recientes se almacena en un *nivel de granularidad temporal alto* (cada cuarto de hora, por ejemplo) y la de objetos más antiguos en un *nivel superior de granularidad* (cada día, por ejemplo). Al igual que el modelo *damped window*, se concentra en datos que son más recientes y no descarta plenamente objetos del pasado.

### Tipos de algoritmos de aprendizaje

Batch mode learning algorithms Incremental learning algorithms On-line learning algorithms Anytime learning algorithms

Metodos de evaluación de tareas de clasificación para flujos de datos

## 3.2. Algoritmos de aprendizaje supervisado

La mayor parte del esfuerzo dedicado para desarrollar algoritmos de aprendizaje automático para flujos de datos se ha enfocado en la realización de propuestas relacionadas con **aprendizaje supervisado**. Existen diversas revisiones dedicadas al aprendizaje automático para tareas de clasificación (Aggarwal [2014], Nguyen et al. [2014], Lemaire et al. [2015]); no obstante, en estas revisiones, por cada uno de los paradigmas clasificatorios, no se mencionan muchas propuestas, por lo que en este trabajo pretendemos **aportar más artículos que tratan el aprendizaje automático para flujos de datos**, enfocándonos en aquellas *propuestas más recientes*. Además, por cada uno de los algoritmos de aprendizaje automático, vamos a añadir

una **tabla comparativa entre las diferentes propuestas**, con el objetivo de que el lector adquiriera una idea global de las distintas características que poseen.

### 3.2.1. Clasificadores bayesianos

### 3.2.2. Árboles de decisión

Los **árboles de decisión** son uno de los algoritmos de aprendizaje automático más estudiados para clasificación de flujos de datos debido a su buen desempeño en los mismos, además de su simplicidad y de la interpretabilidad del modelo. Los desafíos que suponen realizar tareas de clasificación en flujos de datos provocan que los métodos de clasificación basados en *árboles de decisión* como el *ID3*, el *C4.5* y *CART* vistos con anterioridad no sean efectivos en tratar dicho tipo de datos. Estos métodos almacenan y procesan el conjunto de datos de entrenamiento enteros, pero en el caso de datos generados continuamente se necesita incrementar los requerimientos de procesamiento y solo se puede realizar hasta cierto punto, ya sea porque se sobrepasan los límites de memoria o porque el tiempo de computación es demasiado largo. Además, aunque estos métodos fueran capaces de manejar todas las instancias del flujo de datos, muchas de ellas pueden que no sean útiles para la construcción del modelo debido a cambios en el proceso de generación de datos, y estos métodos no son capaces de manejar dichos cambios.

Dadas estas características, los métodos mencionados previamente reciben el nombre de algoritmos **no incrementales de inducción de árboles**. Existen otros algoritmos basados en árboles de decisión que son **incrementales**, que permiten actualizar un árbol existente utilizando solo instancias recientes, sin tener que volver a procesar instancias pasadas. Entre las primeras versiones incrementales de construcción de árboles de decisión, se encuentran el **ID3'** (C. Schlimmer and Fisher [1986]), el **ID4** (C. Schlimmer and Fisher [1986]), el **ID5** (UTGOFF [1988]), el **ID5R** (Utgoff [1989]) y el **ITI** (Utgoff et al. [1997]).

El **ID3'** es una variante incremental del algoritmo *ID3* que se basa en un método de *fuerza bruta* para llevar a cabo la construcción del árbol de decisión, de manera que, cada vez que recibe una nueva instancia, vuelve a construir la estructura de árbol mediante el algoritmo *ID3*, por lo que, debido al costo computacional que conlleva, no es adecuado para la clasificación de flujos de datos. Por otra parte, el **ID4** es también una versión incremental del *ID3*, basado en, cada vez que recibe una nueva instancia de entrenamiento, actualizar el árbol; para ello se basa en almacenar el *número de instancias positivas y negativas* para cada posible valor de cada posible atributo no testeado en cada nodo del árbol de decisión actual, lo que permite testear diferentes atributos en los diferentes nodos del árbol. El algoritmo *ID4* no es eficiente puesto que *descarta subárboles cada vez que se realiza un cambio del atributo de testeo en un determinado nodo*, lo que provoca que el árbol no aprenda ciertos conceptos inherentes a los datos. Además, *no garantiza que el árbol construido sea similar al que produce el algoritmo ID3*.

Por otra parte, el *ID5*, a diferencia del *ID4*, no descarta subárboles a la hora de cambiar el nodo donde el atributo de test ya no es el mejor, sino que actualiza el árbol



de decisión realizando una serie de manipulaciones de la estructura relacionadas con la *subida del nuevo mejor atributo test en ese nodo hacia la parte superior del árbol*, de tal forma que el nuevo mejor atributo quede por encima del antiguo; no obstante no garantiza que se obtenga el mismo árbol de decisión que se obtendría con el *ID3*. Un algoritmo que si garantiza esto es el *ID5R*, que es una extensión del *ID5* y que, a diferencia de este último, tras llevar a cabo la manipulación de la estructura mencionada en el *ID5*, realiza una *reestructuración recursiva de los subárboles que están por debajo del nuevo mejor atributo* utilizada en la manipulación previa; este algoritmo puede llegar a ser más lento que *ID3* en algunos casos, dependiendo de las operaciones de reestructuración recursiva llevadas a cabo. Con respecto al algoritmo *ITI*, extiende el algoritmo *ID5R* de tal forma que es capaz de manejar *atributos numéricos, ruido y valores faltantes*, además de incorporar un *mecanismo de poda*; sin embargo, no es capaz de tratar con conjuntos de datos masivos.

En general los algoritmos incrementales basados en árboles de decisión comentados anteriormente *almacenan todos los ejemplos utilizados en memoria*. No obstante, los problemas de clasificación de flujos de datos requieren que se cumplan una serie de restricciones de recursos computacionales, entre ellas la memoria. De esta manera, el primer algoritmo que se propuso específicamente para resolver los problemas que plantea la clasificación de flujos de datos, entre ellas la *memoria limitada*, se denomina **Very Fast Decision Tree (VFDT)**, propuesto en [Domingos and Hulten \[2002\]](#).

Tener en cuenta para introducir el CVFDT -> Paper “Online Classification of Nonstationary Data Streams (2002)” : Continuous streams of data pose new problems to classification methods of data mining, like CART [4], ID3 [28], C4.5 [29], IFN [23], and many others. The common approach of these methods is to store and process the entire set of training examples. The growing amounts of training data increase the processing requirements of data mining systems up to a point, where they either run out of memory, or their computation time becomes prohibitively long. Furthermore, even if all the available examples can be handled by the system, the patterns discovered by an algorithm in the data from the past, may be hardly valid and useful for the new data obtained hours or even minutes later due to unexpected changes in the data-generating process (e.g., a political event which affects the stock prices). -> We could use it to start the comparatives of decision trees

—Since the Hoeffding Tree algorithm was proposed in literature, decision trees became one of the most popular tools for mining data streams. - Defeated by VFDT after addition of a framework, the results of another article [206]. Domingos states in the latest article that ITI was very slow after 100.000 examples (thus some tests are not made beyond 100.000 examples). He states also that ITI results in higher error rate, and higher sensibility to the target concept noise than VFDT.

-The earliest method specifically designed for decision-tree-based stream classification was proposed in [26]. This method is known as the Very Fast Decision Tree (VFDT) method. The very fast decision tree (VFDT) method [14] builds a Hoeffding tree using Hoeffding bounds [28]. It ensures an asymptotically identical hypothesis to the one outputted by batch learning. VFDT method does not store any data in

memory and keeps only statistics of data. Its major drawback is not being adapted to time-changing concepts. - Methods proposed prior to HT are either significantly less tight compared to HT in their approximation of a batch tree or unsuitable for noisy data streams. - The de facto standard for learning decision trees for streaming data is Hoeffding Tree, which is used as a base for many state-of-the-art drift learners. Domingos and Hulten presented one of the first algorithms for incrementally constructing a decision tree in their widely acclaimed work, "Mining High-Speed Data Streams".

-CVFDT [29] is an extension to VFDT handling concept drifts. The idea is to build a new sub-tree when the current one becomes questionable, and replace it only when it becomes less accurate than the new sub-tree.

-Many algorithms have been proposed to handle these difficulties, among them, the Very Fast Decision Tree (VFDT) algorithm. Although the VFDT has been widely used in data stream mining, in the last years, several authors have suggested modifications to increase its performance, putting aside memory concerns by proposing memory-costly solutions. Besides, most data stream mining solutions have been centred around ensembles, which combine the memory costs of their weak learners, usually VFDTs. To reduce the memory cost, keeping the predictive performance, this study proposes the Strict VFDT (SVFDT), a novel algorithm based on the VFDT. - Limitado inicialmente a streams de ejemplos con todos sus atributos simbólicos y provenientes de entornos estacionarios, VFDT [41] parte de la observación de Catlett [66] para sistemas off-line basados en árboles de decisión, según la cual, dado el esquema top-down seguido por estas técnicas para construir el modelo, no es necesario utilizar todos los ejemplos de entrenamiento para encontrar la mejor pareja (atributo,valor) como siguiente test-nodo del árbol, sino únicamente un subconjunto de aquellos no descritos previamente mediante cualquier predicado extraído de nodos anteriores.

Tabla 3.1: Algoritmos de aprendizaje para flujos de datos basados en árboles de decisión

Algoritmo	Detección de <i>concept drift</i>	Manejo de <i>outliers</i>	Manejo de datos faltantes	Manejo de variables continuas	Manejo de datos de alta dimensión	Manejo del ruido
ID3'						
ID4						
ID5			No	No		No
ID5R			No	No		No
ITI			Si	Si		Si

- Constructing an online decision tree is based on three main choices. .- In the first place, it is impossible in the data stream context, potentially of infinite size, to keep all the examples. The use of data summaries of limited size is necessary to be able to control the tree memory consumption. The fact that decisions are local to



the leaf justifies storing summaries in each leaf. .- Secondly, cut points are chosen by the evaluation in every leaf of a criterion (generally the Gini or the entropy criterion). This choice being a definitive action has to be robust and made with a certain confidence. .-Finally before a split occurs, the available information in leaves is not used. Using a local model in each leaf allows exploiting this information to improve the global prediction of the tree The quality of a decision tree depends on: (i) the summaries in the leaves, (ii) the split criterion, (iii) the local model. ->They dedicate sections to summaries in the leaves, split criterion and local model. Recently, Utgoff [18] proposes the algorithm ITI which maintains statistics within each leaf in order to reconstruct the tree with the arrival of new examples. The good interpretability of the decision trees and their ability to predict with a low latency make them a suitable choice for processing large amounts of data. However, decision trees are not adapted to manage concept drift. In this case, significant parts of the decision tree must be pruned and relearned.

-Paper Incremental Decision Tree based on order statistics (2013)>The difference in building an online tree and an offline tree comes from the fact that data arrive continuously for the first one. The choice of the attribute to cut is made according to the summary and not on all data. The choice of transforming a leaf into a node, is a definitive action. To make sure that this choice is realized with a certain confidence, Domingos and Hulten suggest in VFDT the use of the Hoeffding bound [16]. This bound brings a guarantee on the choice of the good attribute. The Hoeffding bound was afterwards often used to build online decision tree: VFDTc [12], CVFDT [17], IADEM [22], “ensemble of Hoeffding trees” [19]... The Hoeffding bound is also used in this article to construct the proposed online tree.

### 3.2.3. Inducción de reglas

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

### 3.2.4. Redes neuronales

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

### 3.2.5. $k$ -Vecinos más cercanos

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

### 3.2.6. Máquinas de vector soporte

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

### 3.2.7. Regresión logística

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

### 3.2.8. Métodos combinados de aprendizaje

[Gomes et al. \[2017\]](#), [Krawczyk et al. \[2017\]](#) -> Hablar sobre que son surveys bastante completo y que vamos a añadir algunas propuestas que no están en el mismo.

## 3.3. Algoritmos de aprendizaje no supervisado

### 3.3.1. Agrupamiento

[Nguyen et al. \[2014\]](#), [Silva et al. \[2014\]](#), [Aggarwal \[2013\]](#), [Sharma et al. \[2018\]](#), [Mansalis et al. \[2018\]](#)

### 3.4. Series temporales

A data stream is a flow of rapid data items that challenges the computing system's abilities to transmit, process, and store these massive amounts of incoming elements. Data streams have three models: • time series: data items come in ascending order without increments or decrements; • cash-register model: data items increment temporally; • turnstile model: data items increment and decrement temporally

### 3.5. Redes bayesianas para el descubrimiento de conocimiento

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

## Capítulo 4

### Conclusiones y líneas futuras de trabajo



# Capítulo 5

## Sobre las referencias

La bibliografía o referencias deben aparecer siempre al final de la tesis, incluso en aquellos casos donde se hayan utilizado notas finales. La bibliografía debe incluir los materiales utilizados, incluida la edición, para que la cita pueda ser fácilmente verificada.

*Citar dentro del texto:* Las fuentes consultadas se describen brevemente dentro del texto y estas citas cortas se amplían en una lista de referencias final, en la que se ofrece la información bibliográfica completa.

La cita dentro del texto es una referencia corta que permite identificar la publicación de dónde se ha extraído una frase o parafraseado una idea, e indica la localización precisa dentro de la publicación fuente. Esta cita informa del apellido del autor, la fecha de publicación y la página (o páginas) y se redacta de la forma que puede verse a través de los siguientes ejemplos:

Cuando se citan las palabras exactas del autor deben presentarse entre comillas e indicarse, tras el apellido del autor y, entre paréntesis, la fecha de publicación de la obra citada, seguida de la/s página/s.

Si lo que se reproduce es la idea de un autor (no sus palabras exactas) no se pondrán comillas y se indicará, entre paréntesis, el apellido del autor seguido de la fecha de publicación de la obra a la que se refiere.

No se puede eliminar una parte del texto citado sin señalarse; debe indicarse siempre con puntos suspensivos entre corchetes [...]

Ejemplos de como citar una referencia en el texto son los siguientes ? o ???.  
[Oguntimilehin and Ademola \[2014\]](#)

*Cómo ordenar las referencias:*

1. Las referencias bibliográficas deben presentarse ordenadas alfabéticamente por el apellido del autor, o del primer autor en caso de que sean varios.
2. Si un autor tiene varias obras se ordenarán por orden de aparición.
3. Si de un mismo autor existen varias referencias de un mismo año se especificarán los años seguidos de una letra minúscula y se ordenarán alfabéticamente.

4. Si son trabajos de un autor en colaboración con otros autores, el orden vendrá indicado por el apellido del segundo autor, independientemente del año de publicación. Las publicaciones individuales se colocan antes de las obras en colaboración.

*Cómo citar un artículo de revista:* Un artículo de revista, siguiendo las normas de la APA, se cita de acuerdo con el siguiente esquema general: Apellido(s), Iniciales del nombre o nombres. (Año de publicación). Título del artículo. Título de la revista en cursiva, volumen de la revista (número del fascículo entre paréntesis), primera página- última página del artículo.

*Cómo citar una monografía/libro:* Las monografías, siguiendo las normas de la APA, se citan de acuerdo con el siguiente esquema general: Apellido(s), Iniciales del nombre. (Año de publicación). Título del libro en cursiva. Lugar de publicación: Editorial. Opcionalmente podremos poner la mención de edición, que irá entre paréntesis a continuación del título; y, si fuera el caso el volumen que irá en cursiva.

*Cómo citar un capítulo de un libro:* Los capítulos de los libros se citan de acuerdo con el siguiente esquema general: Apellido(s), Iniciales del nombre o nombres. (Año). Título del capítulo. En A. A. Apellido(s) Editor A, B. B. Apellido(s) Editor B, y C. Apellido(s) Editor C (Eds. o Comps. etc.), Título del libro en cursiva (pp. xxx-xxx). Lugar de publicación: Editorial.

*Cómo citar un acta de un congreso:* Apellido(s), Iniciales del nombre o nombres. (Año). Título del trabajo. En A. A. Apellido(s) Editor A, B. B. Apellido(s) Editor B, y C. Apellido(s) Editor C (Eds. o Comps. etc.), Nombre de los proceedings en cursiva (pp. xxx-xxx). Lugar de publicación: Editorial.

*Cómo citar tesis doctorales, trabajos fin de máster o proyectos fin de carrera:* Apellido(s), Nombre. (Año). Título de la obra en cursiva. (Tesis doctoral). Institución académica en la que se presenta. Lugar.

*Cómo citar un recurso de Internet:* Los recursos disponibles en Internet pueden presentar una tipología muy variada: revistas, monografías, portales, bases de datos... Por ello, es muy difícil dar una pauta general que sirva para cualquier tipo de recurso. Como mínimo una referencia de Internet debe tener los siguientes datos:

1. Título y autores del documento.
2. Fecha en que se consultó el documento.
3. Dirección (URL “uniform resource locator”)

Veamos, a través de distintos ejemplos, cómo se citan específicamente algunos tipos de recursos electrónicos.



*Monografías:* Se emplea la misma forma de cita que para las monografías en versión impresa. Debe agregar la URL y la fecha en que se consultó el documento

*Artículos de revistas:* Se emplea la misma forma de cita que para los artículos de revista en versión impresa. Debe agregar la URL y la fecha en que se consultó el documento.

*Artículos de revistas electrónicas que se encuentran en una base de datos:* Se emplea la misma forma de cita que para los artículos de revista en versión impresa, pero debe añadirse el nombre de la base datos, la fecha en que se consultó el documento.



Apéndice A

Anexos



# Bibliografía

- Sigmoid function, a. URL [https://en.wikipedia.org/wiki/Sigmoid\\_function](https://en.wikipedia.org/wiki/Sigmoid_function). Accessed: 02-05-2019.
- What is the difference between k-means and hierarchical clustering?, b. URL <https://www.quora.com/What-is-the-difference-between-k-means-and-hierarchical-clustering>. Accessed: 03-05-2019.
- Ejemplos de análisis cluster, c. URL <https://www.ugr.es/~mvargas/3.DosEjesanalisisclusteryCCAA.pdf>. Accessed: 04-05-2019.
- C. C. Aggarwal. A survey of stream clustering algorithms. In *Data Clustering: Algorithms and Applications*, 2013.
- C. C. Aggarwal. A survey of stream classification algorithms. In *Data Classification: Algorithms and Applications*, 2014.
- S. V. Alex Smola. *Introduction to Machine Learning*. Cambridge University Press, 2008.
- C. Bielza and P. Larranaga. Discrete bayesian network classifiers: A survey. *ACM Computing Surveys*, 47:1–43, 07 2014. doi: 10.1145/2576868.
- L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996. ISSN 1573-0565. doi: 10.1007/BF00058655. URL <https://doi.org/10.1007/BF00058655>.
- L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. 1984.
- J. C. Schlimmer and D. Fisher. A case study of incremental concept induction. pages 496–501, 01 1986.
- D. Cheng, S. Zhang, Z. Deng, Y. Zhu, and M. Zong. knn algorithm with data-driven k value. pages 499–512, 12 2014. doi: 10.1007/978-3-319-14717-8\_39.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

- T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. ISBN 978-3-540-45014-6.
- P. Domingos and G. Hulten. Mining high-speed data streams. *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 11 2002. doi: 10.1145/347090.347107.
- M. S. E. Ntoutsis and A. Zimek. Lecture 8: Velocity: Data streams: Clustering, 2015.
- E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769, 1965.
- Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML’96, pages 148–156. Morgan Kaufmann Publishers Inc., 1996. ISBN 1-55860-419-7. URL <http://dl.acm.org/citation.cfm?id=3091696.3091715>.
- H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet. A survey on ensemble learning for data stream classification. *ACM Comput. Surv.*, 50:23:1–23:36, 03 2017. doi: 10.1145/3054925.
- A. Gron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O’Reilly Media, Inc., 1st edition, 2017. ISBN 1491962291, 9781491962299.
- P. Gulati, A. Sharma, and M. Gupta. Theoretical study of decision tree algorithms to identify pivotal factors for performance improvement: A review. *International Journal of Computer Applications*, 141:19–25, 05 2016. doi: 10.5120/ijca2016909926.
- I. Khamassi, M. Sayed Mouchaweh, M. Hammami, and K. Ghédira. Discussion and review on evolving data streams and concept drift adapting. *Evolving Systems*, 9, 10 2016. doi: 10.1007/s12530-016-9168-2.
- B. Krawczyk and M. Wozniak. Weighted naïve bayes classifier with forgetting for drifting data streams. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2147–2152, Oct 2015. doi: 10.1109/SMC.2015.375.
- B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132 – 156, 2017. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2017.02.004>. URL <http://www.sciencedirect.com/science/article/pii/S1566253516302329>.
- P. Larranaga, I. Inza, and A. Moujahid. Tema 5. clasificadores k-nn.
- P. Larrañaga and c. Bielza. C. unsupervised classification.

- V. Lemaire, C. Salperwyck, and A. Bondu. A survey on supervised classification on data streams. *Lecture Notes in Business Information Processing*, 03 2015. doi: 10.1007/978-3-319-17551-5\_4.
- S. P. Lloyd. Least squares quantization in pcm. Technical report, Bell Laboratories, 1957.
- S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982. ISSN 0018-9448. doi: 10.1109/TIT.1982.1056489.
- J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.
- S. Mansalis, E. Ntoutsis, N. Pelekis, and Y. Theodoridis. An evaluation of data stream clustering algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11, 06 2018. doi: 10.1002/sam.11380.
- S. Mohanty, K. Nathrout, S. Barik, S. Das, and A. Prof. A study on evolution of data in traditional rdbms to big data analytics. *International Journal of Advanced Research in Computer and Communication Engineering*, 4:230–232, 11 2015. doi: 10.17148/IJARCCCE.2015.41049.
- H.-L. Nguyen, Y.-K. Woon, and W. K. Ng. A survey on data stream clustering and classification. *Knowledge and Information Systems*, 45, 12 2014. doi: 10.1007/s10115-014-0808-1.
- A. Oguntimilehin and O. Ademola. A review of big data management, benefits and challenges. *Journal of Emerging Trends in Computing and Information Sciences*, 5:433–438, 07 2014.
- A. Pesaranghader, H. L. Viktor, and E. Paquet. Mediarimid drift detection methods for evolving data streams. *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2018.
- phuongphuong 4112716 and G. M. M. 405k4213231585. Diagram of an artificial neural network. URL <https://tex.stackexchange.com/questions/132444/diagram-of-an-artificial-neural-network>.
- J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, Mar 1986. ISSN 1573-0565. doi: 10.1007/BF00116251. URL <https://doi.org/10.1007/BF00116251>.
- J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN 1-55860-238-0.

- N. Sharma, S. Masih, and P. Makhija. A survey on clustering algorithms for data streams. *International Journal of Computer Applications*, 182:18–24, 10 2018. doi: 10.5120/ijca2018918014.
- J. Silva, E. Faria, R. Barros, E. Hruschka, A. de Carvalho, and J. Gama. Data stream clustering: A survey. *ACM Computing Surveys*, 46, 03 2014. doi: 10.1145/2522968.2522981.
- A. Unagar and A. Unagar. Support vector machines. unwinded., Mar 2017. URL <https://medium.com/data-science-group-iitr/support-vector-machinessvm-unraveled-e0e7e3ccd49b>.
- P. E. UTGOFF. Id5: An incremental id3. In J. Laird, editor, *Machine Learning Proceedings 1988*, pages 107 – 120. Morgan Kaufmann, San Francisco (CA), 1988. ISBN 978-0-934613-64-4. doi: <https://doi.org/10.1016/B978-0-934613-64-4.50017-7>. URL <http://www.sciencedirect.com/science/article/pii/B9780934613644500177>.
- P. E. Utgoff. Incremental induction of decision trees. *Machine Learning*, 4(2):161–186, Nov 1989. ISSN 1573-0565. doi: 10.1023/A:1022699900025. URL <https://doi.org/10.1023/A:1022699900025>.
- P. E. Utgoff, N. C. Berkman, and J. A. Clouse. Decision tree induction based on efficient tree restructuring. *Machine Learning*, 29(1):5–44, Oct 1997. ISSN 1573-0565. doi: 10.1023/A:1007413323501. URL <https://doi.org/10.1023/A:1007413323501>.
- J. Wang, P. Neskovic, and L. N. Cooper. Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. *Pattern Recognition*, 39(3):417 – 423, 2006. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2005.08.009>. URL <http://www.sciencedirect.com/science/article/pii/S0031320305003365>.
- D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241 – 259, 1992. ISSN 0893-6080. doi: [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1). URL <http://www.sciencedirect.com/science/article/pii/S0893608005800231>.
- S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang. Efficient knn classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1774–1785, May 2018. ISSN 2162-237X. doi: 10.1109/TNNLS.2017.2673241.
- Y. Zhu and D. Shasha. Chapter 32 - statstream: Statistical monitoring of thousands of data streams in real time\*work supported in part by u.s. nsf grants iis-9988345 and n2010:0115586. In P. A. Bernstein, Y. E. Ioannidis, R. Ramakrishnan, and D. Papadias, editors, *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases*, pages 358 – 369. Morgan Kaufmann,



San Francisco, 2002. ISBN 978-1-55860-869-6. doi: <https://doi.org/10.1016/B978-155860869-6/50039-1>. URL <http://www.sciencedirect.com/science/article/pii/B9781558608696500391>.

- I. Zliobaite. Learning under concept drift: an overview. *CoRR*, abs/1010.4784, 01 2010.