2023-2024

Aprendizaje Automático

1. Introducción al Aprendizaje Automático

Francisco Casacuberta Nollaz
(fcn@dsic.upv.es)

Alfons Juan Císcar
(ajuan@dsic.upv.es)

(Con material de Enrique Vidal Ruiz)

Departament de Sistemas Informàtics i Computació (DSIC)

Universitat Politècnica de València (UPV)

Septiembre, 2023

Index

Index

Introducción

Aprendizaje automático (AA), *aprendizaje computacional* o *"machine learning"* (ML):

- Tecnologías desarrollados en el marco de varias disciplinas relacionadas con los *sistemas inteligentes*: *reconocimiento de formas*, *cibernética*, *inteligencia artificial*, *estadística aplicada*, entre otras.

- Modernamente se suele considerar como un planteamiento *integrador* de todas estas disciplinas.

- *No* solo se interesa en el "aprendizaje de modelos" propiamente dicho, sino en todo el proceso de resolución de problemas, basado más o menos explícitamente en una aplicación rigurosa de la *teoría de la decisión estadística*.

Aprendizaje automático: predicción y generalización

Aprendizaje:

- Se asume la existencia de *datos de aprendizaje o entrenamiento*; típicamente datos de *entrada* $x \in \mathcal{X}$ y *salida* $y \in \mathcal{Y}$ de un sistema o proceso.

- El objetivo es obtener un modelo (o función $f : \mathcal{X} \to \mathcal{Y}$) que *generalice* estos datos adecuadamente.

- "Generalizar" frecuentemente consiste en *predecir* la salida a partir de nuevos datos de entrada diferentes de los de entrenamiento.

Index

Orígenes y evolución histórica del AA

Desde los años 40 del pasado siglo, se han venido desarrollando de forma más o menos paralela dos enfoques principales para la disciplina que modernamente se conoce como *sistemas inteligentes* (SI):

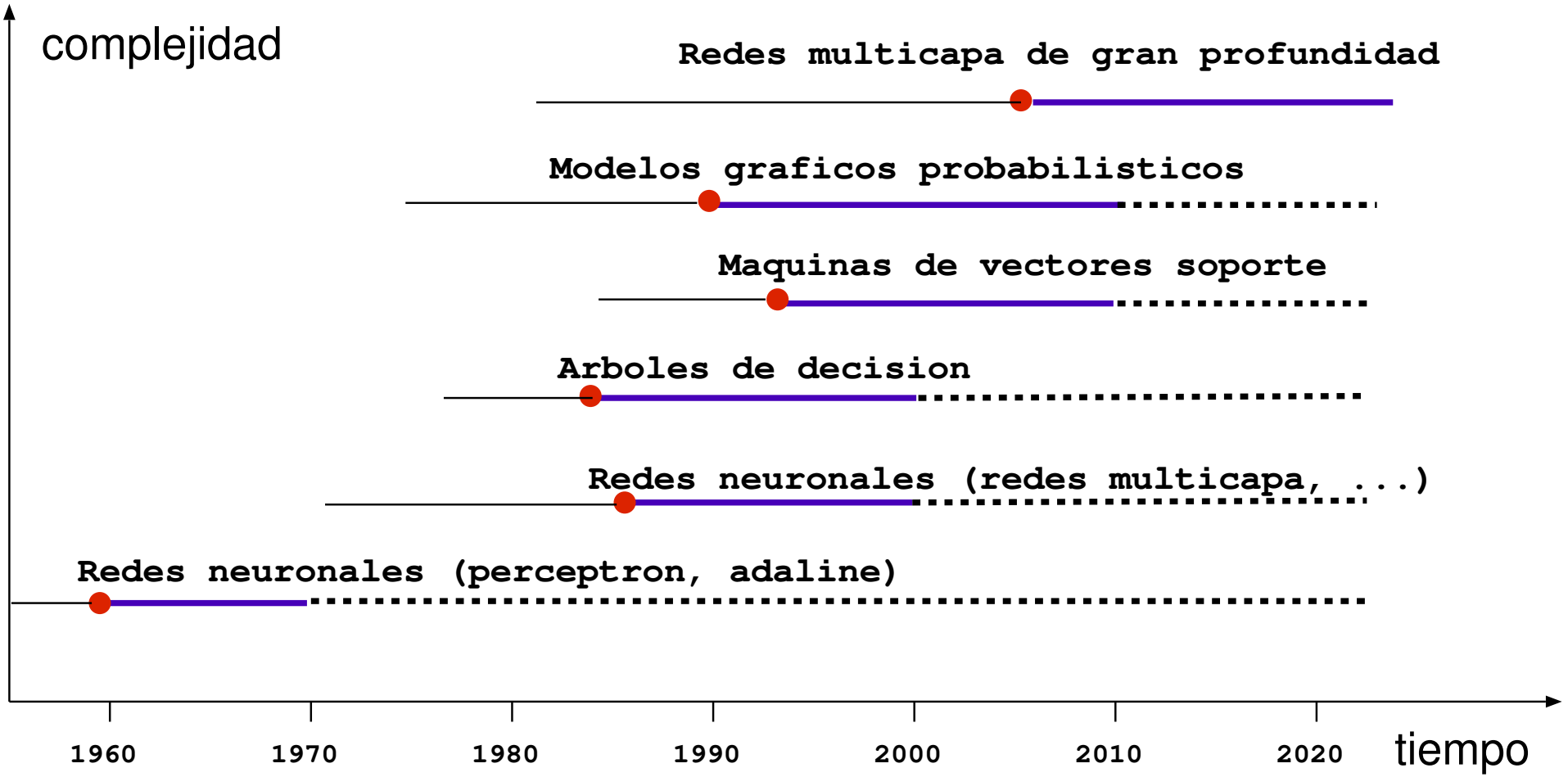
- *Inteligencia artificial* (propiamente dicha, o "clásica" – IA), que se ocupa principalmente de los aspectos mas cognitivos, con claras relaciones con la lógica, el conocimiento y su procesamiento.

- *Reconocimiento de formas* (RF – también "reconocimiento de patrones" o, en inglés, "pattern recognition"), que se ocupa de aspectos más "perceptivos", relacionados con la visión, el habla, etc.

El *aprendizaje automático* surge en los años 80-90 como planteamiento integrador de los enfoques IA y RF, entre otros.

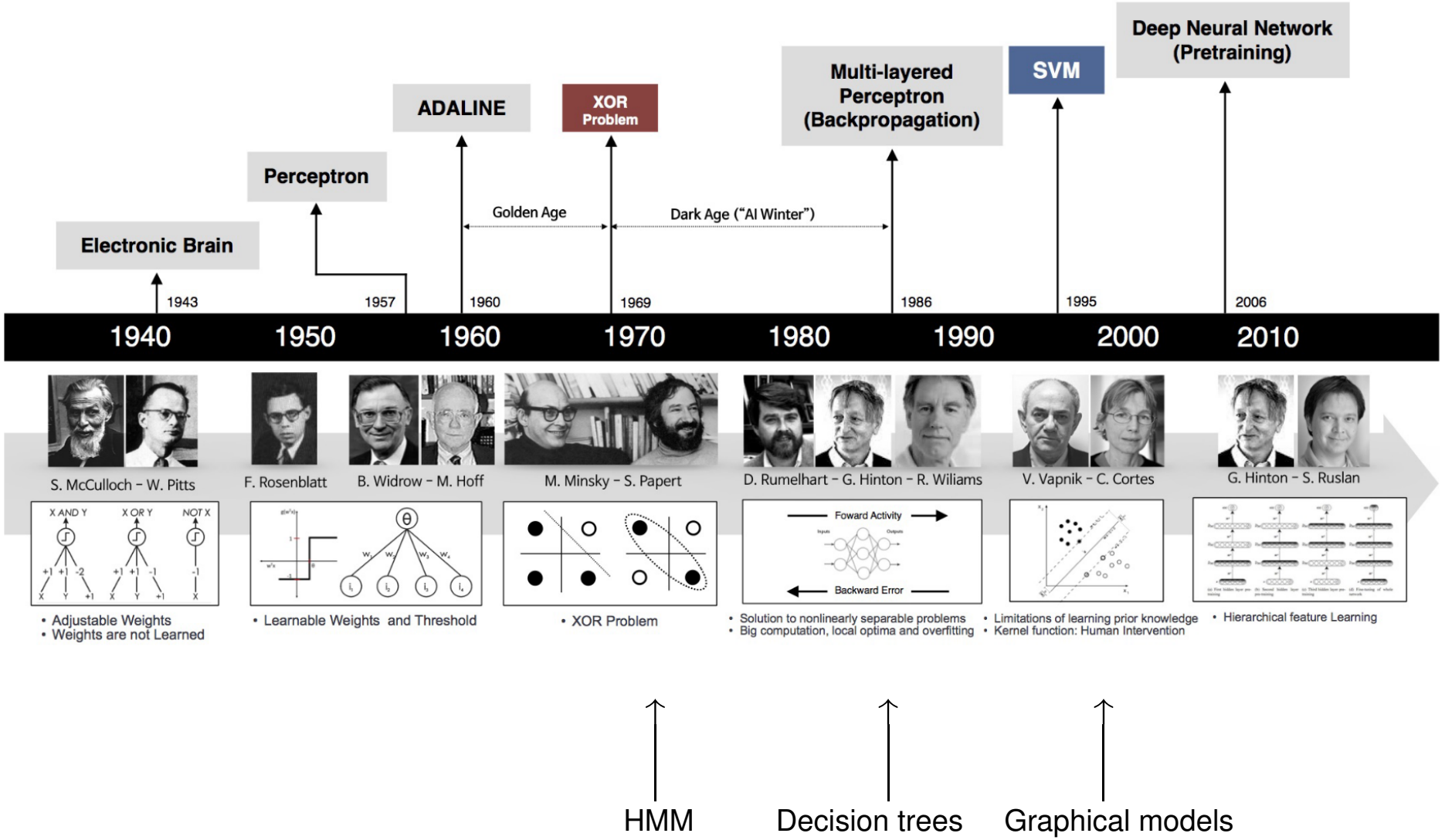
Grandes avances y espectaculares resultados prácticos en los últimos 20 años.

Evolución de algunas tecnologías importantes de AA

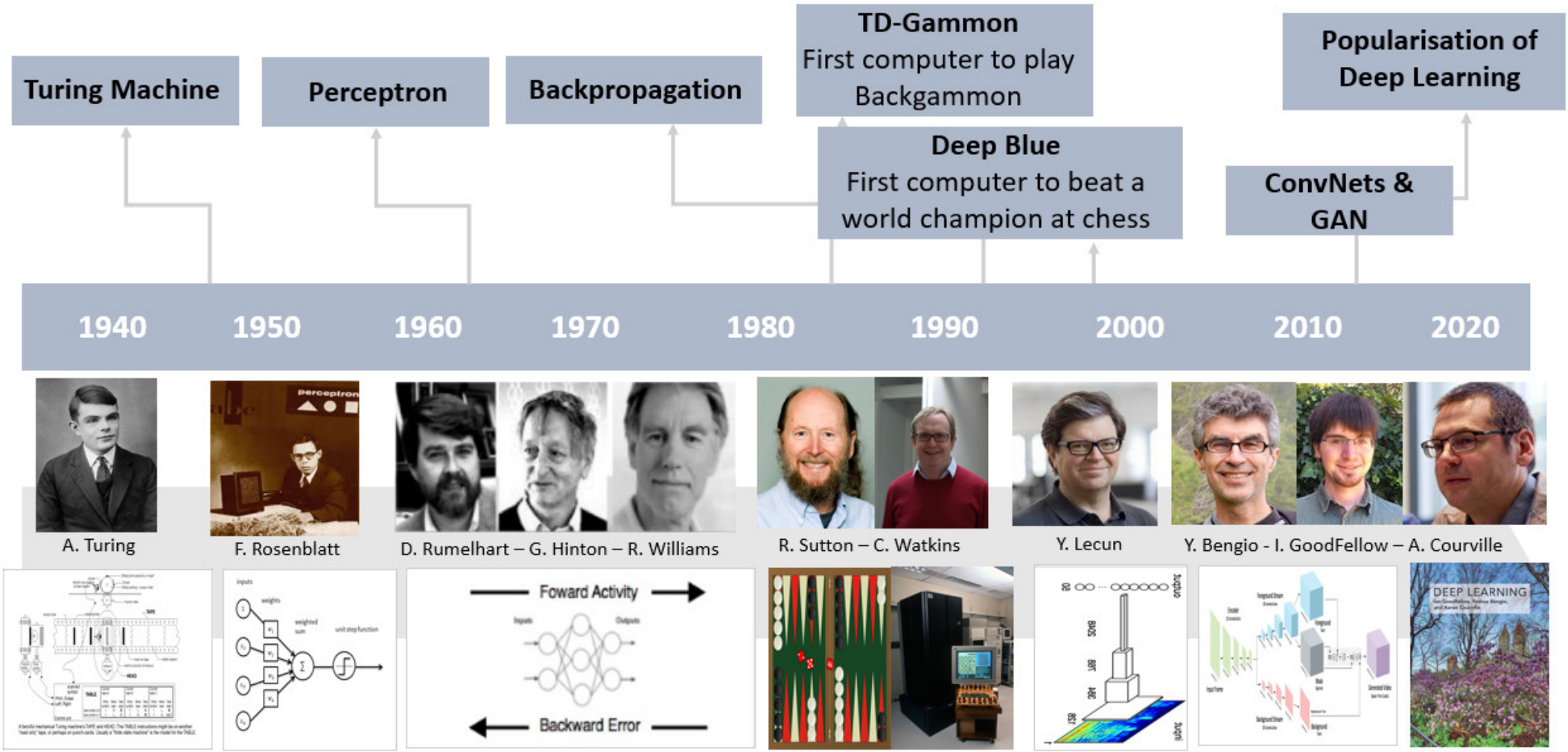


Para cada tecnología, la línea continua indica el periodo de desarrollo teórico-experimental y la de puntos el periodo de vigencia como tecnología consolidada.

Aprendizaje profundo [Serengil 2017]



Cronología del aprendizaje automático escalable [Hanini 2017]



Index

Aproximaciones al AA

- Aprendizaje de modelos probabilísticos, tales como los modelos de Markov, modelos de n-gramas o *modelos gráficos* en general.
- Aprendizaje de funciones discriminantes en máquinas de vectores soporte, redes neuronales, etc.
- Aprendizaje por *proximidad* y *memorización* (almacenamiento de prototipos), basado en medidas de *disimilitud o distancia*.

Aprendizaje supervisado y no supervisado

Aprendizaje supervisado: Información (completa) de *entrada* y *salida* en los datos de entrenamiento.

Aprendizaje no supervisado:

- Los datos de entrenamiento solo contienen información de la *entrada* $x \in \mathcal{X}$.
- El objetivo es obtener información sobre la estructura del dominio de *salida*, $\mathcal{Y}$.
- En problemas de *clasificación*, esta información se refiere a la (posible) estructura en clases de los datos $x \in \mathcal{X}$. En este caso, el problema se conoce como *agrupamiento* o *"clustering"*.

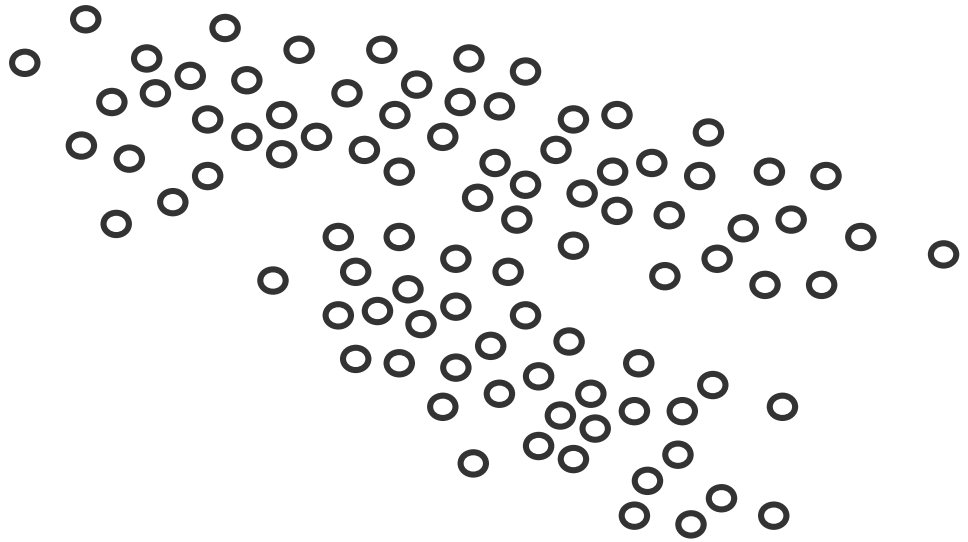
Aprendizaje supervisado y no supervisado

Aprendizaje supervisado: Información (completa) de *entrada* y *salida* en los datos de entrenamiento.

Aprendizaje no supervisado:

- Los datos de entrenamiento solo contienen información de la *entrada* $x \in \mathcal{X}$.
- El objetivo es obtener información sobre la estructura del dominio de *salida*, $\mathcal{Y}$.
- En problemas de *clasificación*, esta información se refiere a la (posible) estructura en clases de los datos $x \in \mathcal{X}$. En este caso, el problema se conoce como *agrupamiento* o *"clustering"*.

Ejemplo: Datos de entrenamiento en $\mathcal{X} = \mathbb{R}^2$:



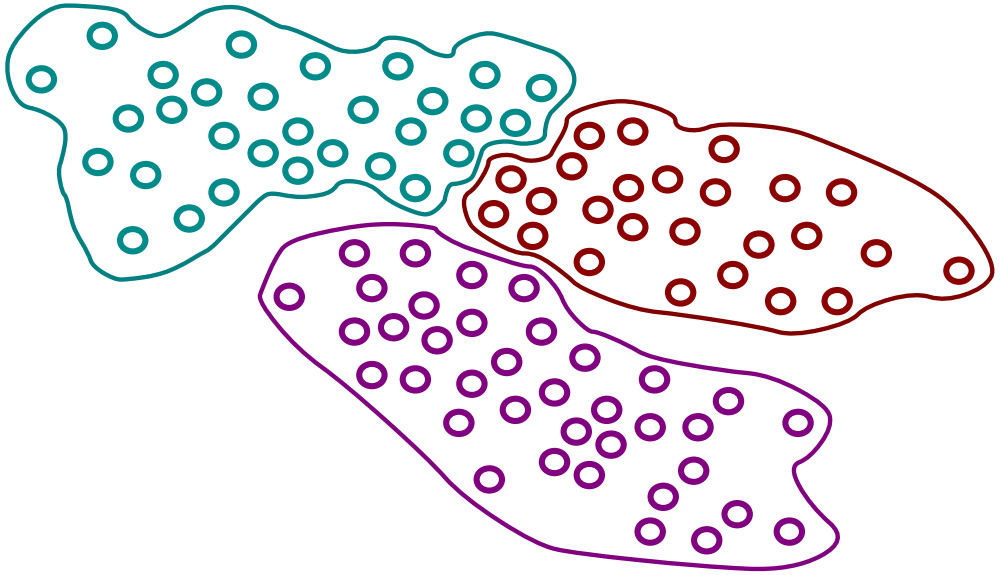
Aprendizaje supervisado y no supervisado

Aprendizaje supervisado: Información (completa) de *entrada* y *salida* en los datos de entrenamiento.

Aprendizaje no supervisado:

- Los datos de entrenamiento solo contienen información de la *entrada* $x \in \mathcal{X}$.
- El objetivo es obtener información sobre la estructura del dominio de *salida*, $\mathcal{Y}$.
- En problemas de *clasificación*, esta información se refiere a la (posible) estructura en clases de los datos $x \in \mathcal{X}$. En este caso, el problema se conoce como *agrupamiento* o *"clustering"*.

Ejemplo: Datos de entrenamiento en $\mathcal{X} = \mathbb{R}^2$, agrupados en tres clases:



Otros modos de aprendizaje automático

- *Aprendizaje "semi-supervisado" (ASS):* se refiere a planteamientos de AA situados entre el aprendizaje totalmente supervisado y totalmente no-supervisado.

- *Aprendizaje adaptativo (AAD):* se parte de un modelo previo, cuyos parámetros se modifican ("adaptan") usando los (nuevos) datos de entrenamiento.

- *Aprendizaje "on-line" (AOL):* no hay distinción explícita entre las fases de "entrenamiento" y "test"; el sistema aprende (posiblemente partiendo de cero) mediante el propio proceso de predicción, con supervisión humana.

- *Aprendizaje activo (AAC):* no se dispone de la salida, y, de cada dato (x) de entrenamiento y el sistema escoge las muestras x más adecuadas para que un agente externo (humano) las etiquete con su y correcta.

- *Aprendizaje por refuerzo (AR):* Puede considerarse como un caso de AOL y ASS en el que la supervisión es "incompleta"; típicamente una información (booleana) de *premio* o *castigo* con respecto a la salida predicha por el sistema.

Index

Planteamiento formal de aprendizaje inductivo

En *aprendizaje inductivo* el sistema posee escaso conocimiento a-priori sobre la tarea a resolver y obtiene su(s) modelo(s) principalmente mediante *ejemplos* o muestras de *entrada/salida* de dicha tarea.

En un planteamiento formal intervienen los siguientes elementos:

- *Método o algoritmo de aprendizaje,* $\mathcal{A}$.
- *Clase* $\mathcal{G}$ *de funciones a aproximar o "aprender"*. Toda $g \in \mathcal{G}$ es de la forma $g\colon \mathcal{X} \to \mathcal{Y}$. Para cada tarea, se asume que existe alguna $g \in \mathcal{G}$ que la representa exactamente[1].
- *Clase* $\mathcal{F}$ *funciones con las que se representan los "modelos" resultado del aprendizaje*. Toda $f \in \mathcal{F}$ es de la forma: $f\colon \mathcal{X} \to \mathcal{Y}$.
- *Muestra finita de aprendizaje* $S \subset \mathcal{X} \times g(\mathcal{X})$.
- *Modo de presentación de la muestra*. Indica cómo se extraen las muestras S de $\mathcal{X} \times g(\mathcal{X})$.
- *Criterio de éxito*. Indica qué se espera de $\mathcal{A}$ al final del aprendizaje.

1. Esta definición obvia la existencia de una *función de representación* que asigna a cada objeto real un elemento del *espacio de representación*, $\mathcal{X}$. Cuando esto se tiene en cuenta, en general $\mathcal{G}$ no puede ser un espacio de *funciones*, sino de *relaciones* de la forma: $g \subset \mathcal{X} \times \mathcal{Y}$.

Regresión y clasificación

- *Regresión*: Tanto los datos de entrada como los de salida pertenecen a dominios $(\mathcal{X}, \mathcal{Y})$ arbitrarios.

  Ejemplos:

  - $\mathcal{X} \subset \mathbb{R}^d$ (vectores de d componentes reales), $\mathcal{Y} \subset \Sigma^\star$ (cadenas de símbolos).
  - Un caso simple: $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ y el modelo predictor es una función $f : \mathbb{R} \to \mathbb{R}$.

- *Clasificación*: $\mathcal{X}$ es arbitrario, pero $\mathcal{Y}$ es un conjunto finito (y generalmente pequeño) de C elementos llamados *clases*. Sin pérdida de generalidad, se puede asumir que $\mathcal{Y} = \{1, 2, \ldots, C\} \subset \mathbb{N}$.

  Ejemplos:

  - Reconocimiento de imágenes de dígitos: $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{1, 2, \ldots, 10\}$.
  - Detección de *"spam"*: $\mathcal{X} \subset \Sigma^\star$, $\mathcal{Y} = \{1, 2\}$, donde Σ es el alfabeto ASCII (o UTF) y las etiquetas $\{1, 2\}$ corresponden a *spam* y *no-spam*.

Index

Teoría de la decisión estadística

- La "función"[1] a aprender, g, es arbitraria (es decir, no se hace ninguna asunción sobre la clase $\mathcal{G}$).
- El modelo que se aprende, $f: \mathcal{X} \to \mathcal{Y}$, se considera una *"función de decisión"*.
- Para definir un *criterio de éxito* (simplificado, pero útil en la práctica) se asume que, para cada $x \in \mathcal{X}$, la *"decisión"* $f(x)$ solo puede ser *"acertada"* o *"errónea"* (por ej., según $f(x)$ sea idéntica o bastante similar a $g(x)$, o no).
- Toda decisión tiene un *coste*. El planteamiento más simple asume que si la decisión $f(x)$ es *acertada* su coste es 0 y si es *errónea* su coste es 1.
- La función de decisión, f, se basa en la *probabilidad a posteriori*, $P(y \mid x)$, estimada a partir de una muestra de entrenamiento $S \subset \mathcal{X} \times g(\mathcal{X})$.
- El *criterio de éxito* es minimizar la esperanza estadística del coste de las decisiones sobre todos los posibles datos de entrada $x \in \mathcal{X}$. Con la simplificación de coste $0/1$, esto equivale a minimizar la probabilidad de error.
- *Este planteamiento es la base del marco estadístico en AA y RF*.

1. En este caso, g *no* es una *función* propiamente hablando, sino una *relación* de la forma $g :\subset \mathcal{X} \times \mathcal{Y}$.

Teoría de la decisión estadística: minimizar el riesgo de error

Sea $x \in \mathcal{X}$ un dato de *entrada* y sea $y \in \mathcal{Y}$ una *decisión* que se toma para x. $P(y \mid x)$ representa la probabilidad de que la decisión y sea correcta.

Probabilidad de error si se toma la decisión y:

$$P_y(\text{error} \mid x) = 1 - P(y \mid x)$$

Teoría de la decisión estadística: minimizar el riesgo de error

Sea $x \in \mathcal{X}$ un dato de *entrada* y sea $y \in \mathcal{Y}$ una *decisión* que se toma para x. $P(y \mid x)$ representa la probabilidad de que la decisión y sea correcta.

Probabilidad de error si se toma la decisión y:

$$P_y(\text{error} \mid x) = 1 - P(y \mid x)$$

Mínima probabilidad de error:

$$\forall x \in \mathcal{X}: \quad P_\star(\text{error} \mid x) = \min_{y \in \mathcal{Y}} P_y(\text{error} \mid x) \;=\; 1 - \max_{y \in \mathcal{Y}} P(y \mid x)$$

Para cada x la probabilidad de error se minimiza si se toma la decisión con mayor $P(y \mid x)$; o sea, la decisión que es "probablemente más acertada".

Teoría de la decisión estadística: minimizar el riesgo de error

Sea $x \in \mathcal{X}$ un dato de *entrada* y sea $y \in \mathcal{Y}$ una *decisión* que se toma para x. $P(y \mid x)$ representa la probabilidad de que la decisión y sea correcta.

Probabilidad de error si se toma la decisión y:

$$P_y(\text{error} \mid x) = 1 - P(y \mid x)$$

Mínima probabilidad de error:

$$\forall x \in \mathcal{X}: \quad P_\star(\text{error} \mid x) = \min_{y \in \mathcal{Y}} P_y(\text{error} \mid x) \;=\; 1 - \max_{y \in \mathcal{Y}} P(y \mid x)$$

Para cada x la probabilidad de error se minimiza si se toma la decisión con mayor $P(y \mid x)$; o sea, la decisión que es "probablemente más acertada".

Función de decisión de mínimo riesgo de error o de Bayes:

$$\forall x \in \mathcal{X}: \quad f^\star(x) = \arg\max_{y \in \mathcal{Y}} P(y \mid x)$$

Teoría de la decisión estadística: minimizar el riesgo de error

Sea $x \in \mathcal{X}$ un dato de *entrada* y sea $y \in \mathcal{Y}$ una *decisión* que se toma para x. $P(y \mid x)$ representa la probabilidad de que la decisión y sea correcta.

Probabilidad de error si se toma la decisión y:

$$P_y(\text{error} \mid x) = 1 - P(y \mid x)$$

Mínima probabilidad de error:

$$\forall x \in \mathcal{X}: \quad P_\star(\text{error} \mid x) = \min_{y \in \mathcal{Y}} P_y(\text{error} \mid x) \;=\; 1 - \max_{y \in \mathcal{Y}} P(y \mid x)$$

Para cada x la probabilidad de error se minimiza si se toma la decisión con mayor $P(y \mid x)$; o sea, la decisión que es "probablemente más acertada".

Función de decisión de mínimo riesgo de error o de Bayes:

$$\forall x \in \mathcal{X}: \quad f^\star(x) = \arg\max_{y \in \mathcal{Y}} P(y \mid x)$$

Esta función garantiza la *minimización del riesgo global o de la esperanza de error de decisión:*

$$P_\star(\text{error}) \;=\; \int_{x \in \mathcal{X}} p_\star(\text{error}, x)\, dx \;=\; \int_{x \in \mathcal{X}} P_\star(\text{error} \mid x)\, p(x)\, dx$$

Ejercicio (recordatorio de la asignatura SIN)

Un problema clásico de decisión consiste en clasificar flores de la familia *Iris* en tres clases; *setosa, versicolor* y *virgínica*, en base a los tamaños de sus pétalos y sépalos (x).

Para ello se han calculado sendos histogramas de las superficies de los pétalos de una muestra de 50 flores de cada clase. Normalizando estos histogramas, se ha estimado la siguiente distribución de tamaños de pétalos para cada clase (y):

	tamaño de los pétalos en cm^2											
$P(x \mid y)$	<1	1	2	3	4	5	6	7	8	9	10	>10
SETO	0.90	0.10	0	0	0	0	0	0	0	0	0	0
VERS	0	0	0	0.20	0.30	0.32	0.12	0.06	0	0	0	0
VIRG	0	0	0	0	0	0	0.08	0.12	0.24	0.14	0.20	0.22

Asumiendo que las clases son equiprobables, calcular:

1. Las probabilidades a posteriory $P(y \mid x),\ y \in \{\text{SETO}, \text{VERS}, \text{VIRG}\}$, para una flor cuyo tamaño de pétalos es $x = 7$ cm^2
2. La decisión óptima de clasificación de esta flor y la probabilidad de que dicha decisión sea errónea
3. La mejor decisión y la correspondiente probab. de error para tamaños de pétalos $1, 2, \ldots, 10$ cm^2
4. La mínima probabilidad de error de decisión esperada para cualquier flor Iris; es decir, $P_\star(\text{error})$
5. Repetir los calculos anteriores, asumiendo que las probabilidades a priori son:
   $P(\text{SETO}) = 0.3,\ P(\text{VERS}) = 0.5,\ P(\text{VIRG}) = 0.2$

Algunas soluciones: a) 0.0, 0.33, 0.67; b) VIRG, 0.33; d) 0.05 (5%) e.a) 0.0, 0.55, 0.44; e.b) VERS, 0.44; e.d) 0.04 (4%)

Index

Estimación de la probabilidad de error

Sea $P_*(\text{error})$ la verdadera probabilidad de error asociada a la función de decisión de Bayes,

Sea $P_f(\text{error}) \stackrel{\text{def}}{=} p$ la verdadera probabilidad de error de un sistema basado en f.

Una estimación empírica ($\hat{p}$) de p puede obtenerse contabilizando el número de errores de decisión, N_e, que se producen en una *muestra de test* con N datos:

$$\hat{p} = \frac{N_e}{N}$$

Si $N >>$, podemos asumir que $\hat{p}$ se distribuye normalmente como:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{N}\right)$$

Intervalo de confianza al 95%:

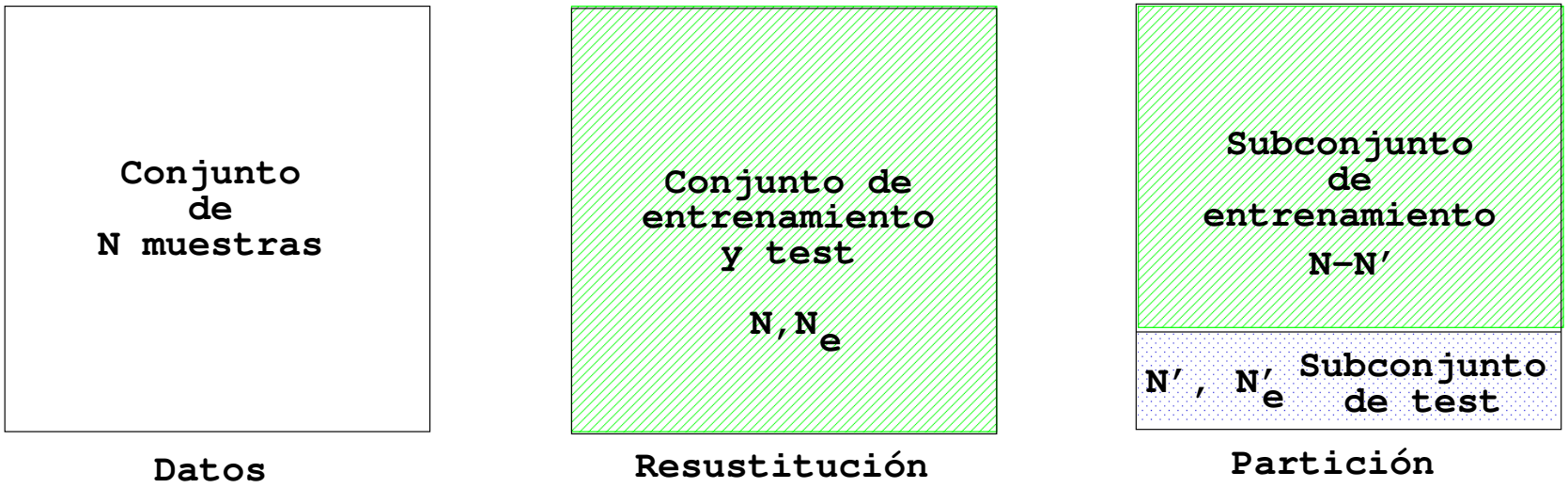
$$P(\hat{p} - \epsilon \leq p \leq \hat{p} + \epsilon) = 0.95; \qquad \epsilon = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Métodos de partición de datos

Para evaluar un sistema de *Aprendizaje Automático*, se necesitan datos etiquetados, no solo para estimar el error, sino para aprender los modelos de decisión. Dado un conjunto de datos, este se puede dividir de diversas formas en subconjuntos de *entrenamiento* y de *test*:

- *Resustitución (Resubstitution):* Todos los datos disponibles se utilizan tanto para para entrenamiento como para test. Inconveniente: es *(muy) optimista*.
- *Partición (Hold Out):* Los datos se dividen en un subconjunto para entrenamiento y otro para test. Inconveniente: desaprovechamiento de datos.
- *Validación Cruzada en B bloques (B-fold Cross Validation):* Los datos se dividen aleatoriamente en B bloques. Cada bloque se utiliza como test para un sistema entrenado con el resto de bloques. Inconvenientes: Reduce el número de datos de entrenamiento (sobre todo cuando B es pequeño) y el coste computacional se incrementa con B.
- *Exclusión individual (Leaving One Out):* Cada dato individual se utiliza como dato único de test de un sistema entrenado con los $N-1$ datos restantes. Equivale a Validación Cruzada en N bloques. Inconveniente: máximo coste computacional.

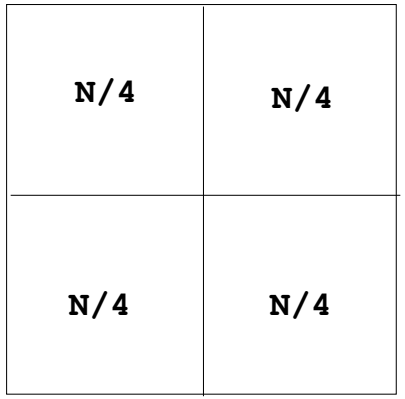
Resustitución y partición

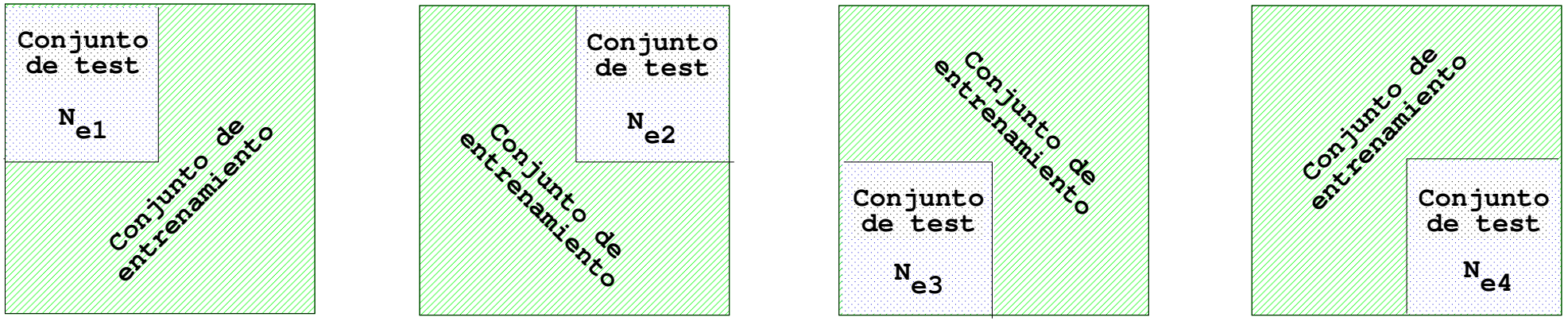


- Resustitución. Error: $\frac{N_e}{N}$. Talla de entrenamiento: N.
- Partición: Error: $\frac{N'_e}{N'}$. Talla de entrenamiento: $N - N'$.

Validación cruzada

B=4





- Error: $\frac{N_{e1}+N_{e2}+N_{e3}+N_{e4}}{N}$.
- Talla de entrenamiento efectiva: $\frac{3N}{4}$.

Parámetros e hiperparámetros

Muchos modelos constan de parámetros e hiperparámetros. En el caso de redes neuronales, los parámetros son los pesos de las conexiones y los hiperparámetros el número de capas de una red neuronal o el número de neuronas por capa entre otros.

- Conjunto de aprendizaje:
  - Conjunto de entrenamiento propiamente dicho para obtener los parámetros del modelo (mediante el algoritmo de aprendizaje).
  - Conjunto de validación para obtener los hiperparámetros del modelo (normalmente por exploración).

- Conjunto de test.

Index

Marco estadístico de Aprendizaje Automático

- *Aprendizaje*:

  Estimación de $P(y \mid x)$ mediante algún criterio adecuado.

  Generalmente el aprendizaje se basa en estimar $P(x, y)$, ya que $P(y \mid x) = P(x, y)/P(x)$.

- *Decisión* o *Inferencia* (clasificación o regresión):

  Para cada dato de *test*, $x \in \mathcal{X}$, calcular $f^\star(x)$; es decir, obtener una y tal que $P(y \mid x)$ sea máxima:

$$f^\star(x) = \arg\max_{y \in \mathcal{Y}} P(y \mid x)$$

  La función $f^\star(\cdot)$ es siempre la misma para cualquier problema; lo que cambia es la distribución $P(y \mid x)$.

Aprendizaje estadístico: Máxima Verosimilitud

Criterio de máxima verosimilitud (MV) para estimación de $P(y \mid x)$:

Se asume que la función de probabilidad conjunta $P(x, y)$ depende de un *vector de parámetros*[1], $\boldsymbol{\theta}$; es decir, $P(x, y) \equiv P(x, y; \boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^D$.

Dado un conjunto de entrenamiento $S \subset \mathcal{X} \times \mathcal{Y}$, la probabilidad (o *"verosimilitud"*) de que $P(x, y; \boldsymbol{\theta})$ genere S, y su logaritmo, son:

$$P(S\,; \boldsymbol{\theta}) \;=\; \prod_{(x,y)\in S} P(x, y; \boldsymbol{\theta})\,, \qquad L_S(\boldsymbol{\theta}) \;=\; \sum_{(x,y)\in S} \log P(x, y; \boldsymbol{\theta})$$

Estimación de máxima verosimilitud de $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L_S(\boldsymbol{\theta})$$

- Si $P(x, y; \boldsymbol{\theta}^\star)$ es la verdadera distribución de la que se ha extraido S, $P(x, y; \hat{\boldsymbol{\theta}}) \to P(x, y; \boldsymbol{\theta}^\star)$ cuando $|S| \to \infty \Rightarrow$
- *MV es consistente con la minimización de la esperanza del error de decisión*

1. Ej: (μ, σ) de una Gausiana, o probabilidades de transición y emisión en modelos ocultos de Markov discretos.

Aprendizaje por MV: regla de Bayes

Frecuentemente puede simplificarse el aprendizaje por MV descomponiendo la probabilidad conjunta como:

$$P(y \mid x) = \frac{P(y)\, P(x \mid y)}{P(x)}$$

donde $P(y)$ es la *probabilidad a priori* de y y $P(x \mid y)$ es la *probabilidad condicional, o verosimilitud* de x dada y.

Así, la log-verosimilitud se descompone como:

$$L_S(\boldsymbol{\theta}) = \sum_{(x,y)\in S} \log P(x, y; \boldsymbol{\theta}) = \sum_{(x,y)\in S} \log P(y; \boldsymbol{\theta}_1) + \sum_{(x,y)\in S} \log P(x \mid y; \boldsymbol{\theta}_2)$$

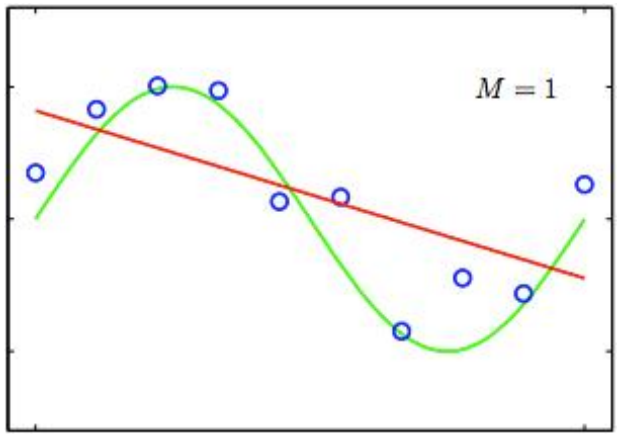
Como ambos sumandos son negativos, para que $L_S(\boldsymbol{\theta})$ sea máximo basta maximizar cada sumando por separado.

$\Longrightarrow$ $P(y; \boldsymbol{\theta}_1)$ y $P(x \mid y; \boldsymbol{\theta}_2)$ pueden estimarse de forma independiente.

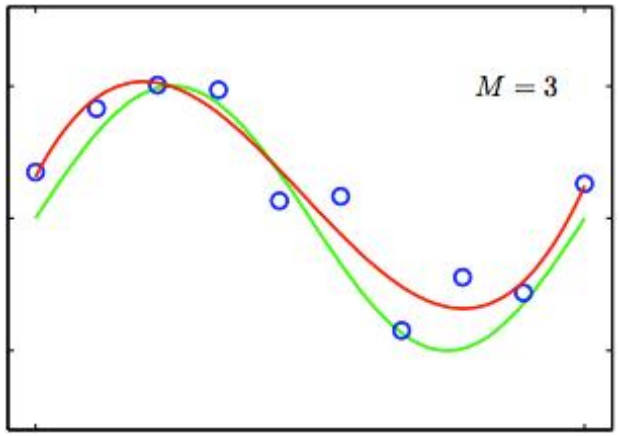
Index

Sobregeneralización y sobreajuste: ejemplos

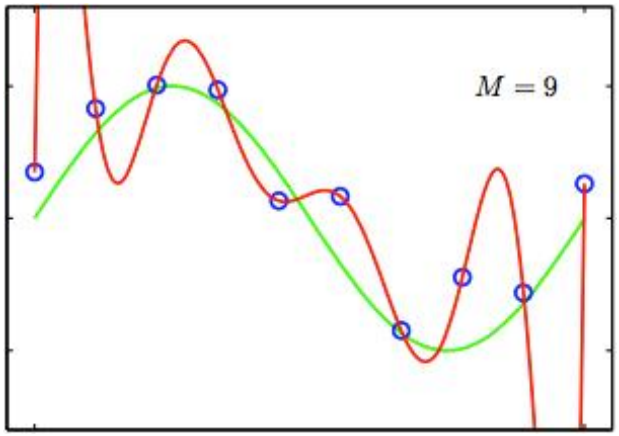
Modelos de regresión f (en rojo) que aproximan a $g : \mathbb{R} \to \mathbb{R}$ (en verde)



sobregeneralización

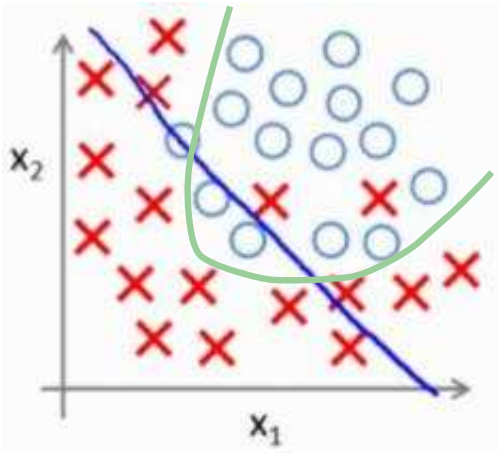


O.K.

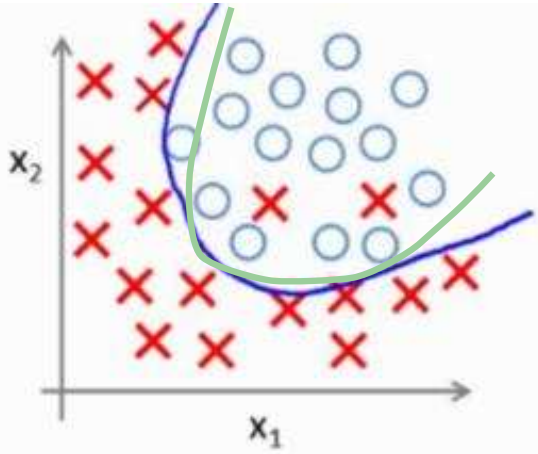


sobreajuste

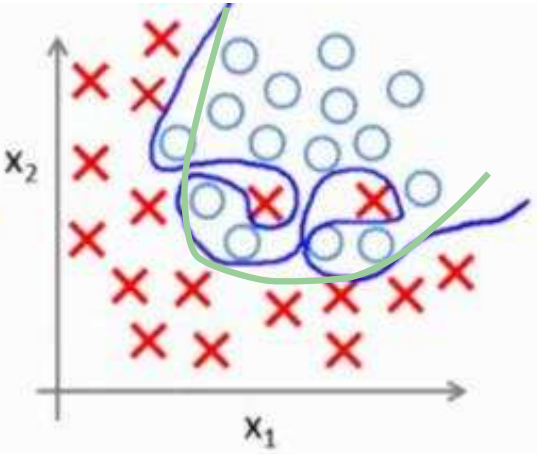
Front. de decisión (azul) que aproximan a las de un clasificador $g : \mathbb{R}^2 \to \{\times, \circ\}$ (verde)



sobregeneralización



O.K.



sobreajuste

Sesgo y varianza

- El compromiso entre sobreajuste y sobregeneralización está estrechamente relacionado con el compromiso entre el sesgo y la varianza.
- Se asume que $\mathcal{X} \equiv \mathbb{R}^{d_x}$, $\mathcal{Y} \equiv \mathbb{R}$ (sin pérdida de generalidad).
- Se dispone de conjuntos de entrenamiento $S \subset \mathcal{X} \times \mathcal{Y}$ de tamaño fijo $|S| = N$.
- La función "verdadera" a aprender $g : \mathcal{X} \to \mathcal{Y}$ es observada mediante sensores imprecisos como $y = g(\boldsymbol{x}) + \epsilon$, con $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, y una imprecisión o "ruido" ϵ que verifica: $\mathbb{E}[\epsilon] = 0$, $\mathbb{V}[\epsilon] = \tau^2$.
- El problema consiste en encontrar un "buen" modelo f, aprendido a partir del conjunto de entrenamiento S. El criterio de éxito es conseguir que la esperanza de error para cualquier $\boldsymbol{x} \in \mathcal{X}$ sea pequeña, teniendo en cuenta que g es desconocida.

  La esperaza de error de generalización se puede expresar como:

$$R_g(f, \boldsymbol{x}, y) \stackrel{\text{def}}{=} \mathbb{E}_{S,\epsilon}\left[\mathcal{E}(f_S(\boldsymbol{x}), y)\right]$$

  donde $\mathcal{E}(\cdot)$ es una función de error que mide la disimilitud entre la predicción de nuestro modelo aprendido $f_S(\boldsymbol{x})$ y lo que observamos y.

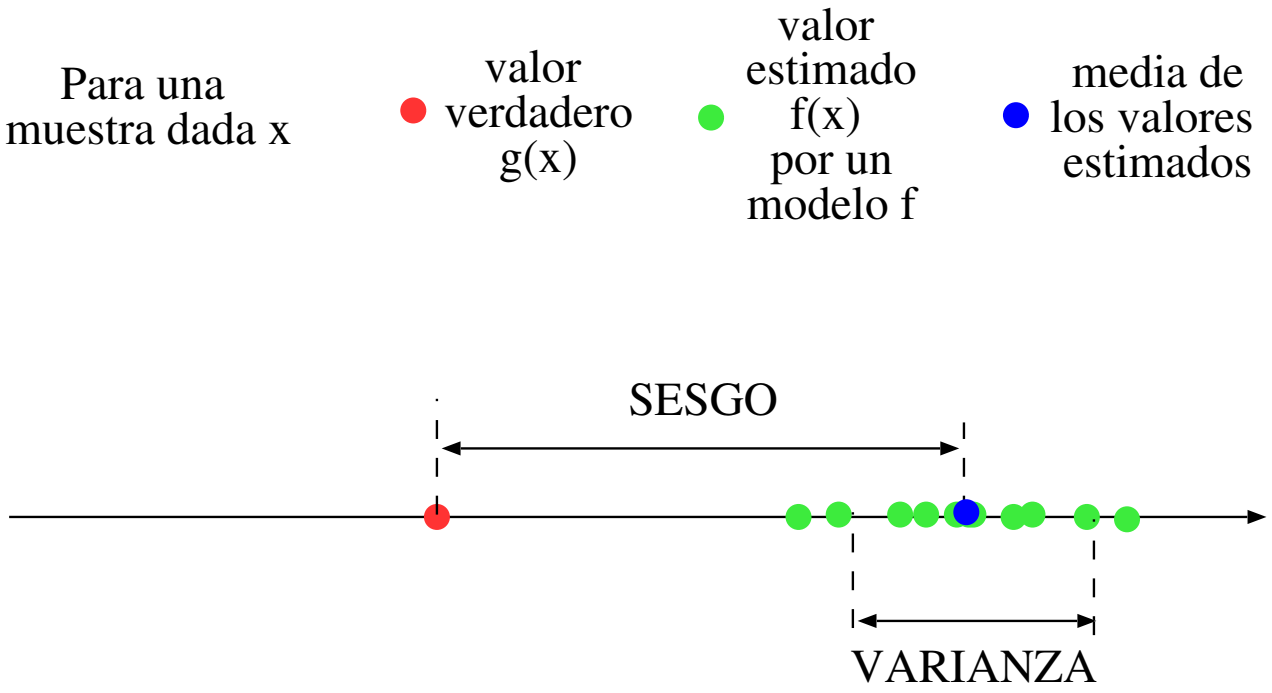
Sesgo y varianza (cont.)

- Sesgo (o *Bias*): $B_S(g, f, \boldsymbol{x}) = \mathbb{E}_S\left[f_S(\boldsymbol{x})\right] - g(\boldsymbol{x})$
- Varianza: $V_S(f, \boldsymbol{x}) = \mathbb{E}_S\left[\left(\mathbb{E}_S\left[f_S(\boldsymbol{x})\right] - f_S(\boldsymbol{x})\right)^2\right]$
- Error irreducible causado por el ruido ϵ: $\mathbb{V}[\epsilon] = \mathbb{E}_\epsilon\left[(y - g(\boldsymbol{x}))^2\right] = \tau^2$
- Propiedad: Si la función de error $\mathcal{E}(\cdot)$ es el *error cuadrático*, se verifica:

$$R_g(f, \boldsymbol{x}, y) \equiv \mathbb{V}[\epsilon] + \left(B_S(g, f, \boldsymbol{x})\right)^2 + V_S(f, \boldsymbol{x})$$

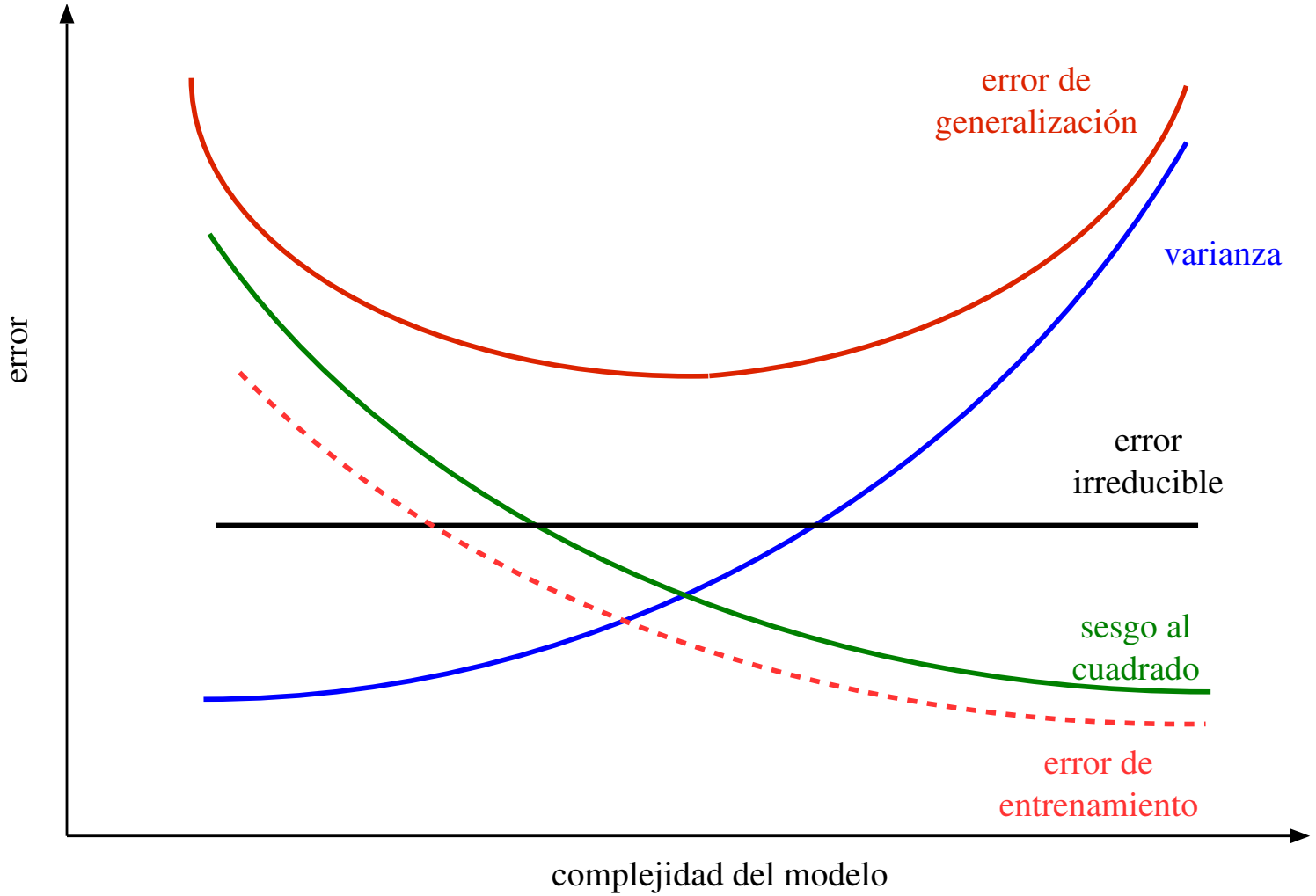
- Estimación del sesgo y la varianza: por validación cruzada, asumiendo $\tau = 0$.

Sesgo y varianza



Sesgo y varianza

[Papachristoudis. The Bias-Variance Tradeoff. 2019]



Sesgo y varianza

[Avati. Bias-Variance Analysis:Theory and Practice, 2020]

- *Varianza alta.* Síntoma: si el error de entrenamiento es bajo pero el error de validación cruzada es alto, seguramente el modelo tendrá una gran varianza:
  - Probable *sobreajuste*.
  - Es inútil cambiar a modelo más grande.
  - Soluciones para reducir la varianza: aumentar la regularización, obtener un conjunto de datos más grande, disminuir el número de características, usar un modelo más pequeño, etc.

- *Sesgo alto.* Síntoma: si el modelo no se ajusta bien a los datos de entrenamiento, seguramente el sesgo será alto.
  - Probable *sobregeneralización*.
  - Es inútil gastar tiempo y recursos en obtener más datos.
  - Soluciones para reducir el sesgo: disminuir la regularización, usar más características, usar un modelo más grande, etc.

Index

La amenaza de la dimensionalidad

- Si $\mathcal{X} \equiv \mathbb{R}^d$, cuando d es muy grande, aparecen diversos fenómenos adversos que se conocen comúnmente como la *"amenaza de la dimensionalidad"*.
- La causa común de estos problemas es que, cuando aumenta d, el volumen del espacio (por ej., de un hipercubo) aumenta exponencialmente y los datos aparecen muy dispersos.
- Ej.: bastan $10^2 = 100$ puntos para muestrear un intervalo unidad (un hipercubo en $\mathbb{R}^1$) para que los puntos no disten más de $10^{-2} = 0.01$ entre sí. Pero en $\mathbb{R}^{10}$ harían falta 10^{20} puntos.
- *Curiosidad* relacionada con lo anterior:
  si $d{>>>}$, *¡los puntos de un hipercubo tienden a concentrarse "cerca de sus vértices"!*.

  Si $d{>>>}$, el volumen de un hipercubo de lado $2r$ es $(2r)^d$, mientras que el de una hiperesfera de radio r (contenida en él) es *mucho* menor: $2r^d\pi^{d/2}/d\,\Gamma(d/2)$.

  Al aumentar d, el volumen de la hiperesfera resulta insignificante con respecto al del hipercubo:

$$d \to \infty \quad \Rightarrow \quad \frac{2r^d\pi^{d/2}}{d\,\Gamma(d/2)}\,\frac{1}{(2r)^d} \to 0$$

- Con frecuencia, estos fenómenos causan problemas de *sobregeneralización* y *sobreajuste*.
- Soluciones: determinación de la *"dimensionalidad intrínseca"*, técnicas de *reducción de la dimensionalidad*, etc.

Index

Notación

- $\mathbb{R}$, $\mathbb{N}$ y $\mathbb{B}$; espacios de los reales, de los naturales y de los booleanos, respectivamente.
- $\mathbb{R}^d$: espacio vectorial de d dimensiones.
- Σ^*: espacio de cadenas de longitud finita de símbolos.
- $\mathcal{X}, \mathcal{Y}$: espacios de datos de entrada y de salida, respectivamente.
- x, y: un dato de entrada y un dato de salida, respectivamente.
- $f, g : \mathcal{X} \to \mathcal{Y}$: funciones entre el espacio de entrada y el de salida.
- C: número de clases en un problema de clasificación.

Conceptos básicos de Estadística y Probabilidad

Variable aleatoria, probabilidad $P(X = x) \equiv P(x) : \quad \sum_x P(x) = 1$

Probabilidad, densidad $P(x),\ p(x) : \quad \sum_x P(x) = 1\ , \quad \int_x p(x)\ d\,x = 1$

Probabilidad conjunta $P(x, y) : \quad \sum_x \sum_y P(x, y) = 1$

Probabilidad condicional $P(x \mid y) : \quad \sum_x P(x \mid y) = 1 \quad \forall y$

Marginales $P(x) \ = \ \sum_y P(x, y), \qquad P(y) \ = \ \sum_x P(x, y)$

Regla de la probabilidad conjunta $P(x, y) \ = \ P(x)\, P(y \mid x)$

Regla de la cadena $P(x_1, x_2, \ldots, x_N) \ = \ P(x_1) \prod_{i=2}^{N} P(x_i \mid x_1, \ldots, x_{i-1})$

Regla de Bayes $P(y \mid x)\ P(x) \ = \ P(y)\, P(x \mid y)$

Esperanza $E_P[f(x)] \ \equiv \ E_X[f(x)] \ \equiv \ \mathbb{E}[f(x)] \ = \ \sum_x f(x)\ P(x)$