2021-2022

Aprendizaje Automático

2. Conceptos básicos de aprendizaje automático

Enrique Vidal Ruiz
(evidal@dsic.upv.es)

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

Departament de Sistemas Informàtics i Computació (DSIC)

Universitat Politècnica de València (UPV)

Septiembre, 2021

Index

Index

Planteamiento formal de aprendizaje inductivo

Recordatorio de los elementos que intervienen en el aprendizaje automático:

- *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: $g \subset \mathcal{X} \times \mathcal{Y}$.

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".

Mínimo riesgo global o mínima esperanza de error de decisión:

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

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".

Mínimo riesgo global o mínima esperanza de error de decisión:

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

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

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

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:

a) 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

b) La decisión óptima de clasificación de esta flor y la probabilidad de que dicha decisión sea errónea

c) La mejor decisión y la correspondiente probab. de error para tamaños de pétalos $1, 2, \ldots, 10$ cm^2

d) La mínima probabilidad de error de decisión esperada para cualquier flor Iris; es decir, $P_\star(\text{error})$

e) 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%)

Marco estadístico de AA

- *Aprendizaje*: Estimación de $P(y \mid x)$ mediante algún criterio adecuado.
- *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.

Marco estadístico de AA: 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:

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

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

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

MV es consistente con la minimización de la esperanza del error de decisión.

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

Marco estadístico de AA: regla de Bayes

Frecuentemente puede simplificarse el aprendizaje por MV descomponiendo las probabilidades conjunta y/o a posteriori, mediante la *regla de Bayes*:

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

De esta forma, la función de decisión resulta:

$$f^\star(x) = \arg\max_y P(y)\ P(x \mid y)$$

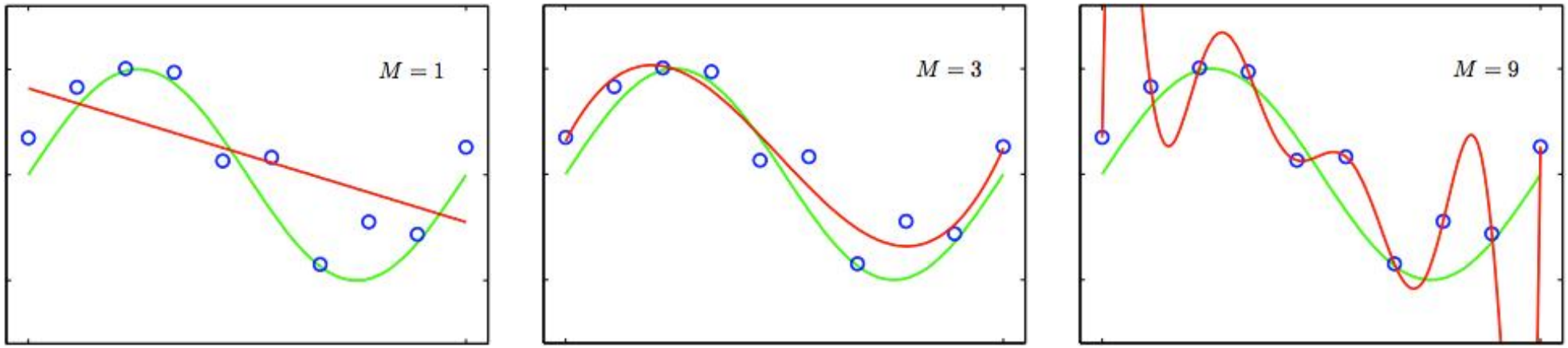
donde $P(y)$ es la *probabilidad a priori* de y y $P(x \mid y)$ es la *probabilidad condicional* de x dada y.

Ejercicio: Aplicar el criterio de máxima verosimilitud en este caso, asumiendo los correspondientes modelos que $P(y) \approx P(y; \boldsymbol{\theta})$ y $P(x \mid y) \approx P(x \mid y; \boldsymbol{\theta})$

Index

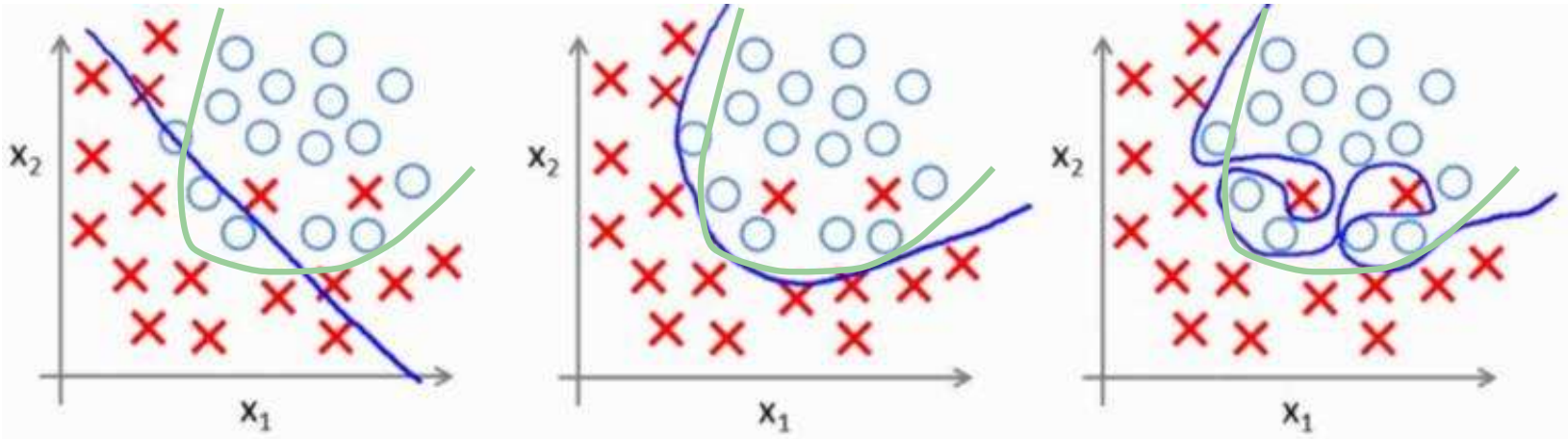
Sobregeneralización y sobreajuste: ejemplos

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 $\mathcal{X} \equiv \mathbb{R}^{d_x}$ y $\mathcal{Y} \equiv \mathbb{R}^{d_y}$.
- Se dispone de un conjunto de entrenamiento $S \subset \mathcal{X} \times \mathcal{Y}$ de tamaño $|S| = N$.
- La función "verdadera" a aprender $g : \mathcal{X} \to \mathcal{Y}$ es observada mediante sensores imprecisos como $\boldsymbol{y} = g(\boldsymbol{x}) + \epsilon$, con $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, 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 de generalización 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:

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

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

Sesgo y varianza

- Sesgo: $B_S(g, f, \boldsymbol{x}) \;=\; \mathbb{E}_S\left[f_S(\boldsymbol{x})\right] - g(\boldsymbol{x})$
- Varianza: $V_S(f, \boldsymbol{x}) \;=\; \mathbb{E}_S\left[\left(\mathbb{E}_S\left[f_S(\boldsymbol{x})\right] - f_S(\boldsymbol{x})\right)^2\right]$
- Propiedad: Si la función de error E es el error cuadrático, se verifica:

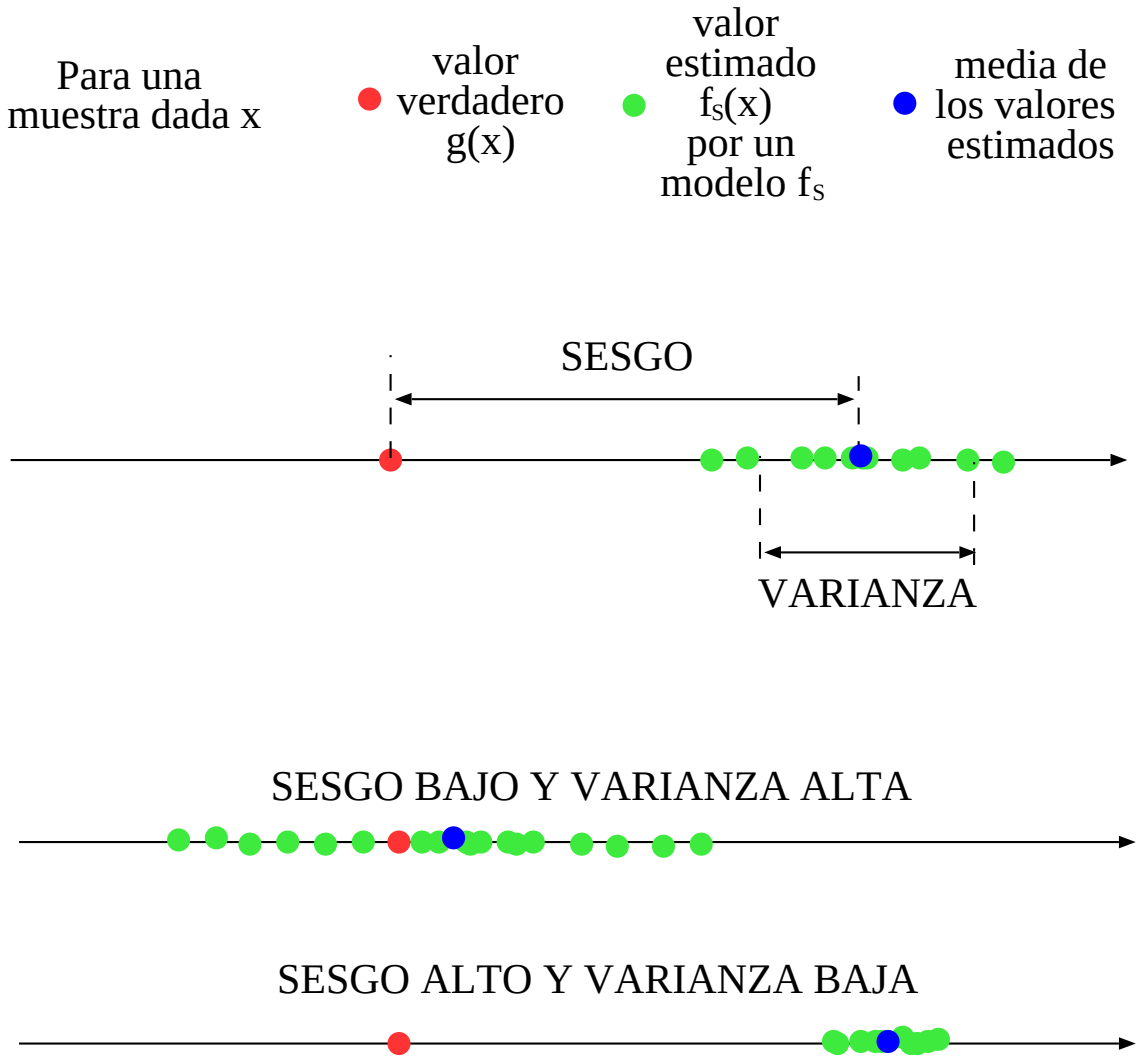
  $$R_g(f, \boldsymbol{x}, \boldsymbol{y}) \;\equiv\; \mathbb{E}_{S,\epsilon}\left[E(f_S(\boldsymbol{x}), \boldsymbol{y})\right] \;=\; N(g, \boldsymbol{x}, \boldsymbol{y}) + \left(B_S(g, f, \boldsymbol{x})\right)^2 + V_S(f, \boldsymbol{x})$$

  donde $N(g, \boldsymbol{x}, \boldsymbol{y})$ es el error irreducible causado por el ruido ϵ:

  $$N(g, \boldsymbol{x}, \boldsymbol{y}) \;=\; \mathbb{E}_\epsilon\left[(\boldsymbol{y} - g(\boldsymbol{x}))^2\right] \;=\; \mathbb{E}_\epsilon\left[\epsilon^2\right] \;\stackrel{\text{def}}{=}\; \tau^2$$

- Estimación del sesgo y la varianza: por valización cruzada, asumiendo que el ruido es nulo ($\tau = 0$).

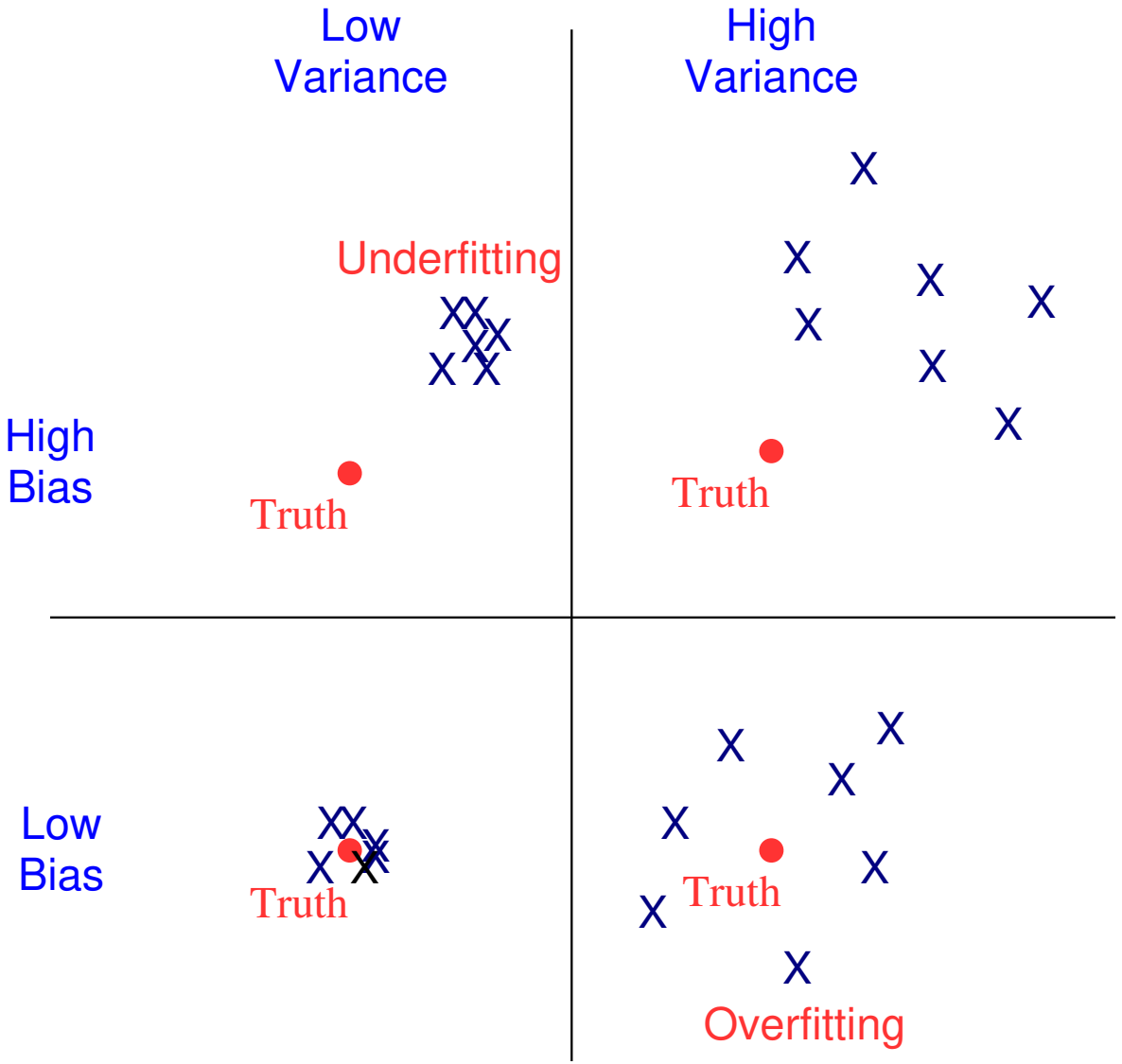
Sesgo y varianza



Los puntos verdes son las predicciones de f_S al variar los datos de entrenamiento, S.

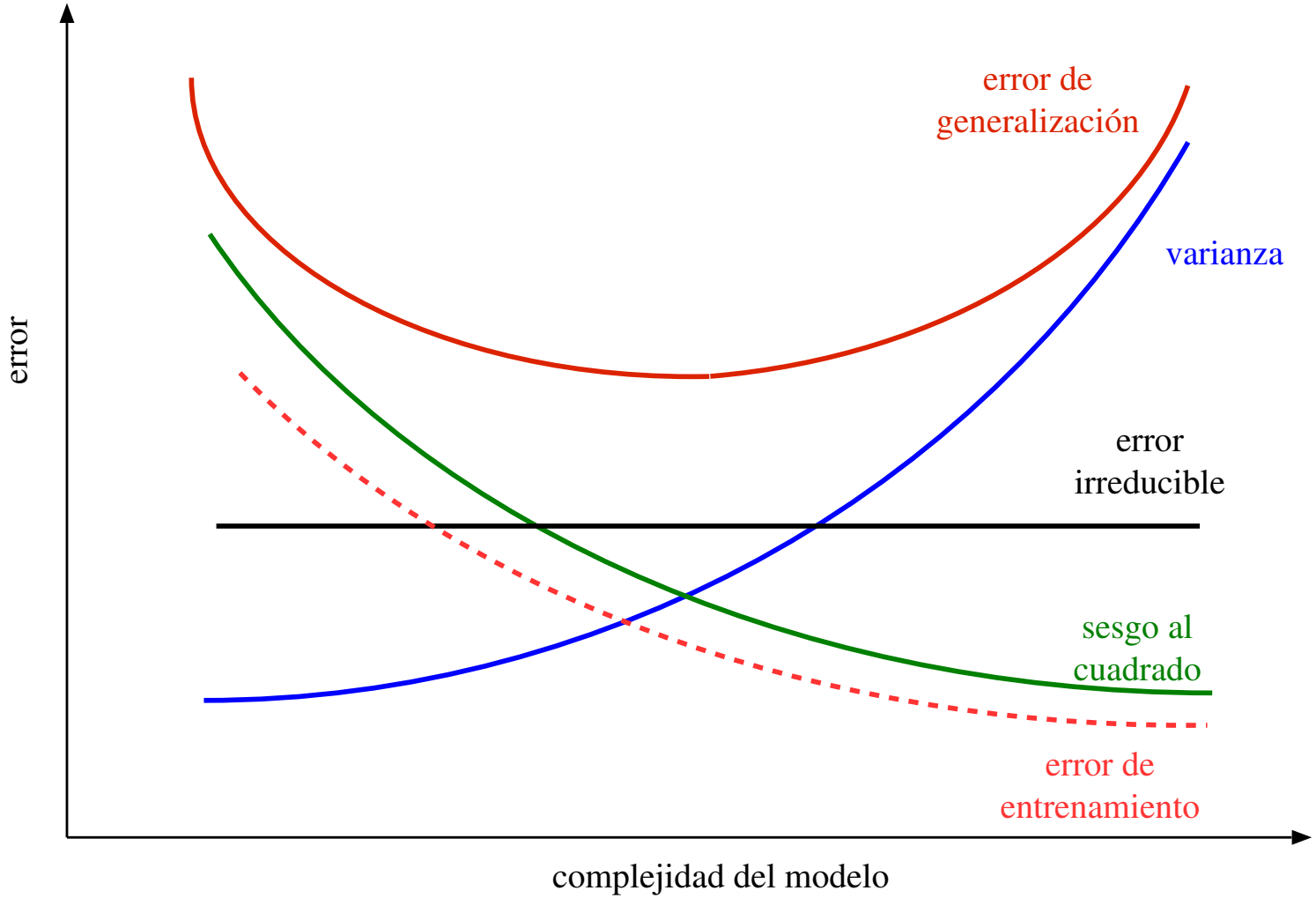
Sobregeneralización y sobreajuste

[Singh. Understanding the Bias-Variance Tradeoff. 2018]



Sobregeneralización y sobreajuste

[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 sobreentrenamiento.
  - 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.

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

Evaluación: Esperanza de error

Sea $f : \mathcal{X} \to \mathcal{Y}$ la función obtenida mediante un sistema de AA. La *probabilidad de error* de f es la esperanza estadística de que la salida de f sea incorrecta.

Supongamos que $\mathcal{X}$ es un dominio discreto y sea $P(x)$ la (verdadera) distribución de probabilidad incondicional de las entradas $x \in \mathcal{X}$.

$$E_x[\text{error}(f(x))] = \sum_{x \in \mathcal{X}} \text{error}(f(x))P(x)$$

Si $\mathcal{X}$ es continuo (por ejemplo, $\mathcal{X} = \mathbb{R}^d$):

$$E_{\boldsymbol{x}}[\text{error}(f(\boldsymbol{x})] = \int_{\boldsymbol{x} \in \mathcal{X}} \text{error}(f(\boldsymbol{x}))\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

donde $p(\boldsymbol{x})$ es ahora la *densidad de probabilidad* incondicional.

Esta es la *"verdadera"* probabilidad de error, pero normalmente no se conoce $P(x)$, o solo se conocen aproximaciones que no permiten calcular $E[\text{error}(f)]$.

En la práctica, $E[\text{error}(f)]$ se suele *estimar* mediante datos de *"test etiquetados"*; es decir, datos similares a los de entrenamiento, que contienen información de *entrada* y la correspondiente "etiqueta" (información de *salida correcta*).

Estimación de la probabilidad de error

Sea $p = E[\text{error}(f)]$ la *verdadera esperanza* (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%:

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

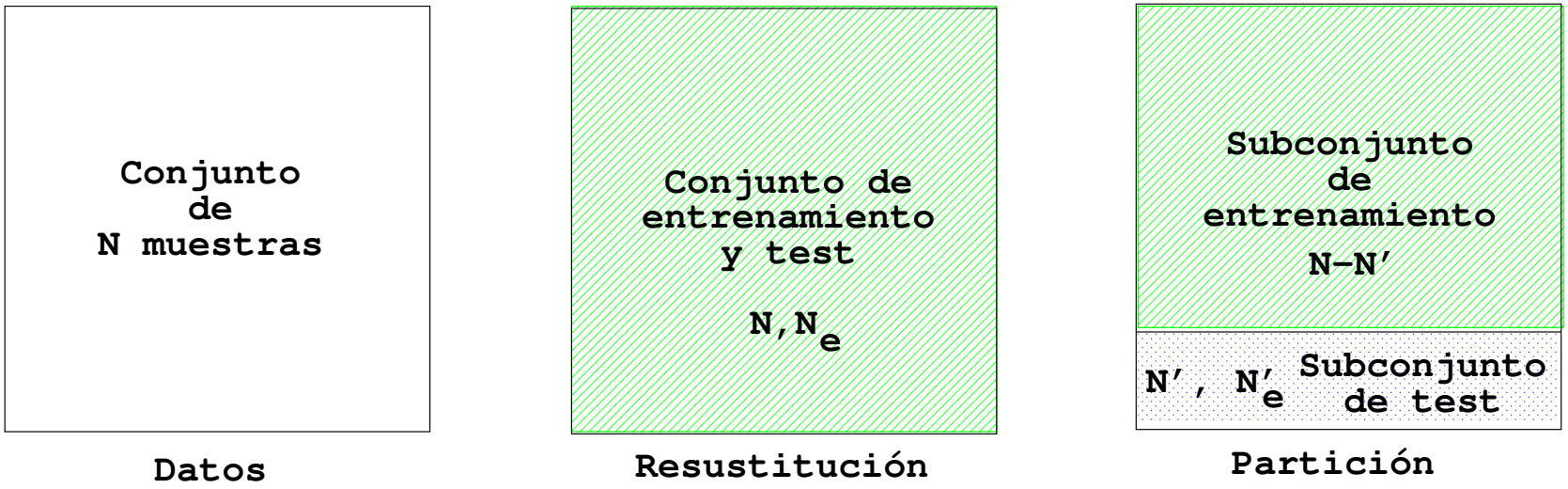
Index

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.

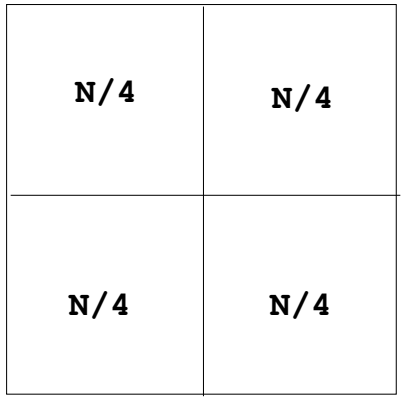
Resustitución y partición

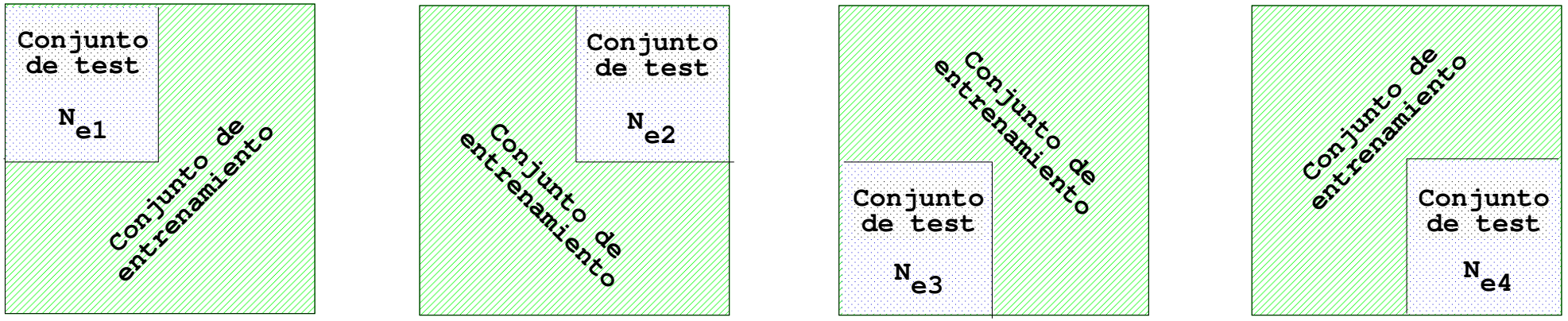


- Resustitución. Error: $\frac{N_e}{N}$. Talla de entrenamiento: N.
- Partición: Error: $\frac{N'_e}{N'}$. Talla de entrenamiento: $N - N'$.

Validación cruzada

B=4





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

Index

Notación

- $\mathcal{A}$: algoritmo de aprendizaje
- $\mathcal{G}$: clase de funciones a aproximar o "aprender"
- $\mathcal{F}$: Clase funciones con las que se representan los "modelos"
- S: muestra finita de aprendizaje ($S \subset \mathcal{X} \times g(\mathcal{X})$)
- $|g|$: talla de la descripción de g
- $\ln$ y $\log$: logaritmo neperiano y decimal, respectivamente
- $P(x;\theta)$: probabilidad de $x \in \mathcal{X}$ segun un modelo de parámetros dados θ
- $\max_x(.)$ ($\min_x(.)$) y $\arg\max_x(.)$ ($\arg\min_x(.)$): operadores que devuelven el máximo (mínimo) variando x y el argumento x que maximiza (minimiza) una función, respectivamente
- $\prod$ y $\sum$: operadores productorio y sumatorio, respectivamente
- $L_S(.)$: logaritmo de la verosimilitud de un conjunto S
- $\phi : \mathcal{X} \to \mathbb{R}$: una función discriminante
- $\mathcal{Q}$ conjunto de estados.
- r y R funciones de recompensa y recompensa acumulada, respectivamente
- τ: distribución de probabilidad de las transiciones