Ciencia de Datos: un primer acercamiento

Aprendizaje supervisado

Álvaro Cabana (FPsico)

UNIVERSIDAD DE LA REPUBLICA URUGUAY

Aprender

¿Qué es? ¿Hay más de una forma?

Aprender

Cambiar estados internos en base a la experiencia, mejorando un desempeño o comportamiento.

Precisamos entonces:

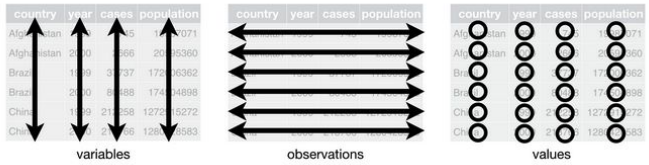
- Experiencia: exposición a información del ambiente (datos)
- Modificar estados internos: algo cambia en base a la experiencia
- Métrica de desempeño: el cambio no es caprichoso, sino que mejora el comportamiento del sistema

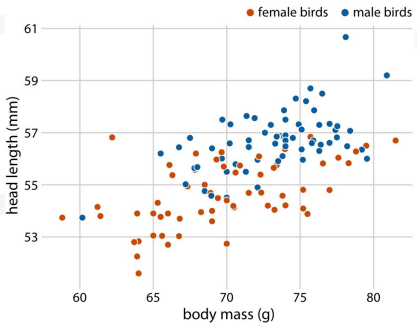
El ciclo de los datos

predicción /
testeo

modelo

inferencia /
entrenamiento



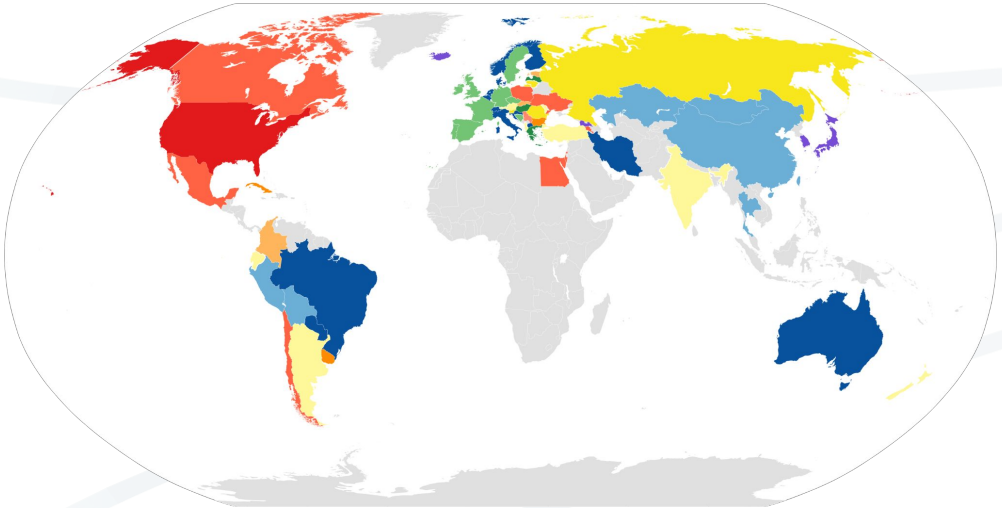


descripción

Media, mediana, desvío, IQR, etc...

Modelos

Representaciones útiles de la realidad



"All models are wrong, some are useful" George E.P. Box

Del Rigor en la Ciencia

Minicuento de J.L. Borges

Tipos de aprendizaje

Según la relación entre la métrica de desempeño y el proceso de cambio de estados internos

- **Supervisado:** El desempeño ideal está disponible, guía el aprendizaje.
- **No supervisado:** El desempeño ideal no está presente, o no existe.
- **Por refuerzo:** No está disponible el desempeño ideal, sino sólo un escalar qué representa qué tan bueno es el desempeño actual.

Tipos de aprendizaje

En el Sistema Nervioso Central de los mamíferos...

- **Supervisado:** Cerebelo y otras estructuras de control motor. La señal de error (esperado-observado) está presente y retroalimenta (closed loop).
- **No supervisado:** Cortezas cerebrales de asociación.
- **Por refuerzo:** Circuito de recompensa (VTA / NAcc / Corteza prefrontal).

Tipos de aprendizaje

En aprendizaje automático:

- **Supervisado:** Regresión, SVM, RF, muchos modelos de redes neurales.
- **No supervisado:** PCA, ICA, clustering, algunos modelos de RN (Self Organizing Maps).
- **Por refuerzo:** Modelos de Reinforcement Learning (y DeepRL).

Hoy veremos:

Clasificación
Métricas de desempeño
Cómo aprender - función de pérdida
Sobreajuste y regularización

Asociacionismo

Un tipo importante de aprendizaje: el aprendizaje asociativo

Estímulo → respuesta

Acción → recompensa

Entrada → salida

Imagen → etiqueta

Aprender es aproximar una función deseada:

$y = f(x)$

f() = perro

f() = gato

En el aprendizaje supervisado, **conocemos** un conjunto de pares x,y

Aprendizaje supervisado

Tenemos un conjunto de pares x,y

Imágenes y sus etiquetas

Entradas y sus correspondientes salidas

En base a ese conjunto de datos estimo f(x) [entrenamiento]
$\hat{y} = \hat{f}(x)$

perro

perro

gato

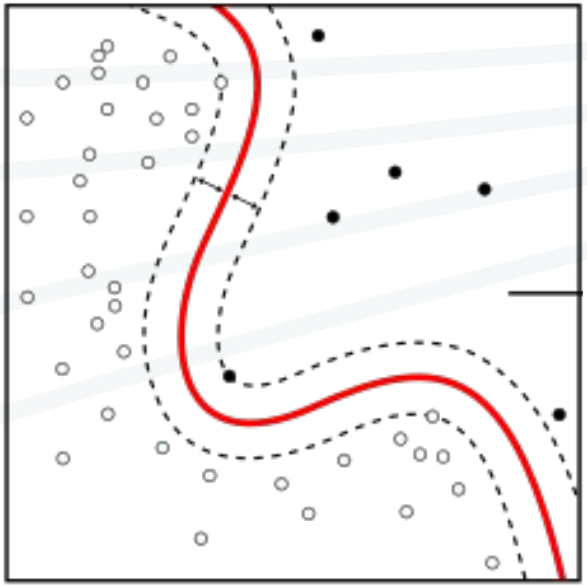
perro

gato

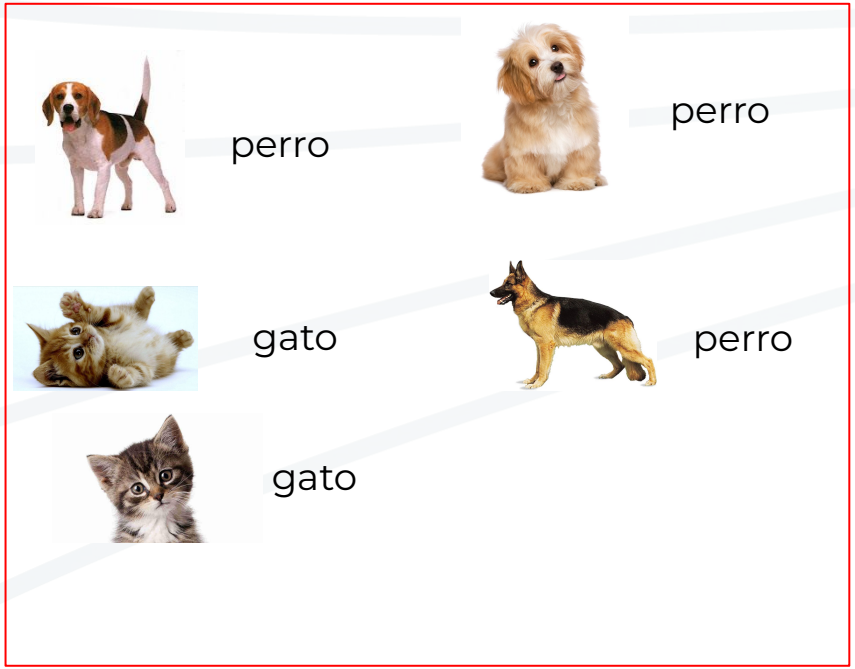
En este caso los *y* son **categorías discretas.** Ésta es una tarea de **categorización o clasificación.**

Si y es continua es una tarea de regresión.

Aprendizaje Supervisado: Clasificación



Clasificación



Conjunto de entrenamiento (Training)

Ejemplo

X e y suelen ser vectores: $x \in \mathbb{R}^n \; y \in \mathbb{R}^m$

y = {0,1} 1: es mamífero, 0: no es mamífero

x = [x_1 x_2] x_1: tiene pelo x_2: pone huevos

Conjunto de entrenamiento (Training)

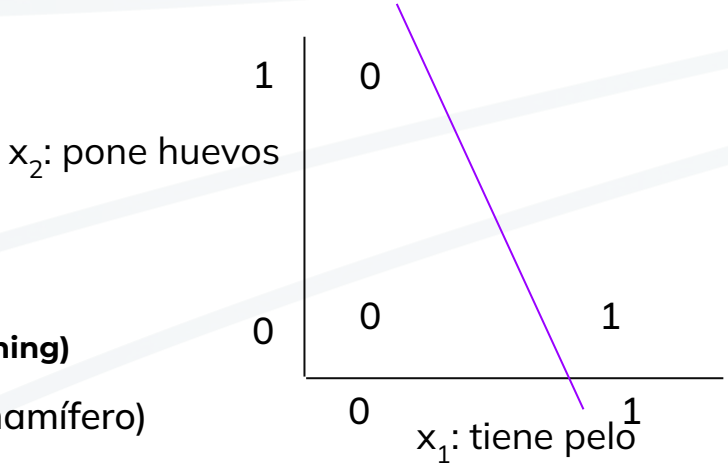
[0 0] → 0 (salamandra, no mamífero)
[1 0] → 1 (perro, mamífero)
[0 1] → 0 (sapo, no mamífero)

Las dimensiones de entrada suelen ser llamadas *rasgos (features)*

Ejemplo

y = {0,1} 1: es mamífero, 0: no es mamífero

x = [x_1 x_2] x_1: tiene pelo x_2: pone huevos



$\hat{y} = H(a_1x_1 + a_2x_2 + b)$

$\hat{y} = H(2x_1 - 1x_2 + -.5)$

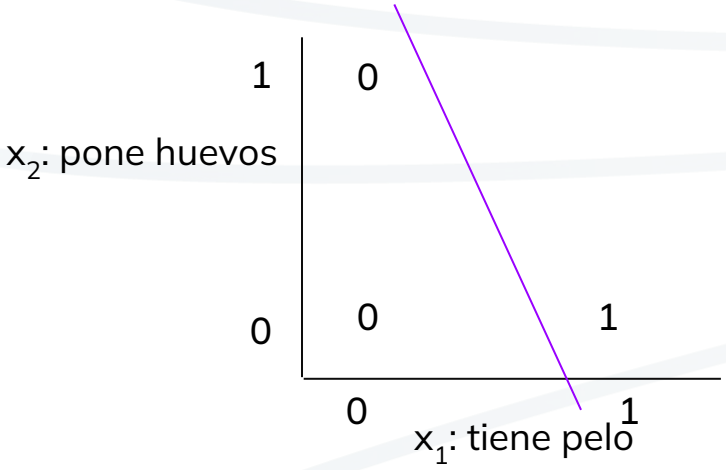
$\hat{y}(0,0) = H(-5) = 0$
$\hat{y}(0,1) = H(-1.5) = 0$
$\hat{y}(1,0) = H(1.5) = 1$

Conjunto de entrenamiento (Training)

[0 0] → 0 (salamandra, no mamífero)
[1 0] → 1 (perro, mamífero)
[0 1] → 0 (sapo, no mamífero)

Ejemplo

$\hat{y} = H(a_1x_1 + a_2x_2 + b)$



$\hat{y} = H(2x_1 - 1x_2 + -.5)$

$\hat{y}(0,0) = H(-5) = 0$
$\hat{y}(0,1) = H(-1.5) = 0$
$\hat{y}(1,0) = H(1.5) = 1$

Es un buen **clasificador**?

Conjunto de entrenamiento (Training)

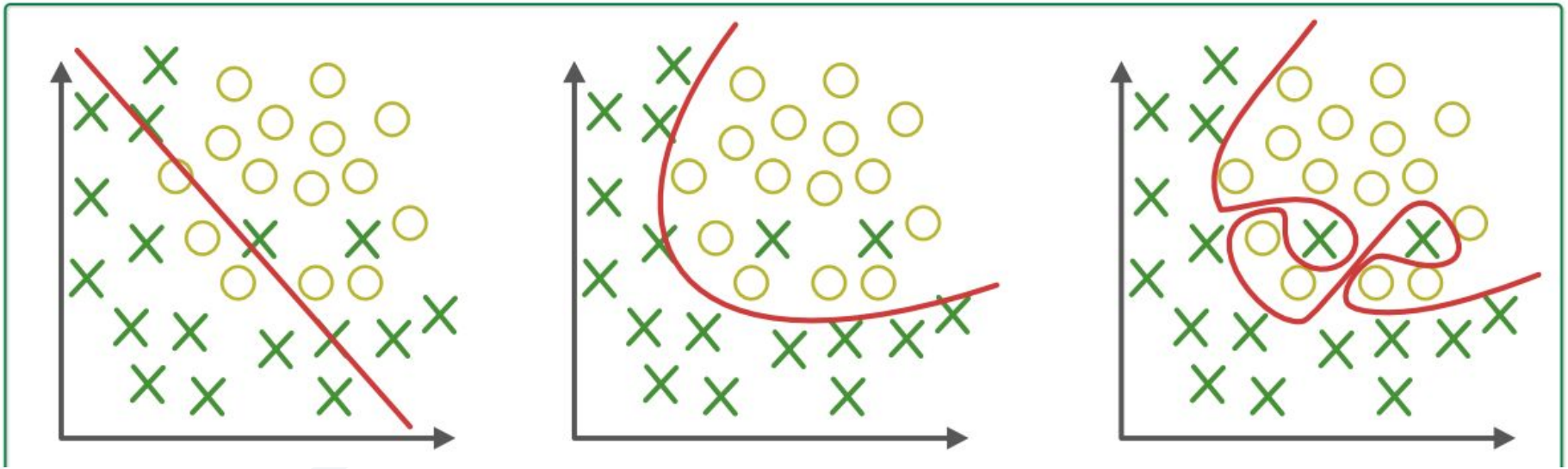
[0 0] → 0 (salamandra, no mamífero)
[1 0] → 1 (perro, mamífero)
[0 1] → 0 (sapo, no mamífero)

$$\text{Accuracy} = \frac{\#\text{ elementos correctamente clasificados}}{\#\text{ elementos en total}}$$

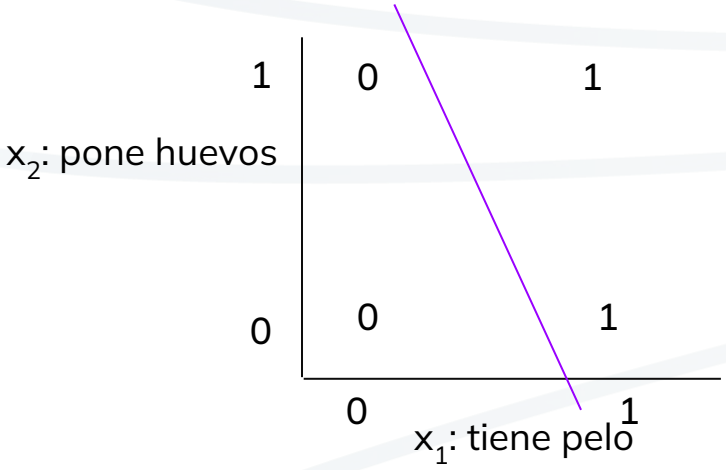
$$\text{Accuracy} = \frac{3}{3} = 1 \text{ !!}$$

Clasificación

La clasificación (en especial la binaria) puede visualizarse como una partición del espacio de entrada



Ejemplo



Conjunto de entrenamiento (Training)

[0 0] → 0 (salamandra, no mamífero)
[1 0] → 1 (perro, mamífero)
[0 1] → 0 (sapo, no mamífero)

Es un buen **clasificador**?

$$\text{Accuracy} = \frac{3}{3} = 1 \text{ !!}$$

Qué tan bien **generaliza** a datos no observados hasta el momento?

Conjunto de testeo, o validación (Testing, validation)

[1 1] → 1 (ornitorrinco, mamífero)

$$\hat{y} = H(2x_1 - 1x_2 + -.5)$$

$$\hat{y}(1,1) = H(.5) = 1 \qquad \text{Accuracy} = \frac{1}{1} = 1 \text{ !!}$$

Clasificación


perro

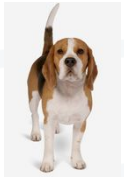

perro


gato


perro


gato

Conjunto de entrenamiento (Training)


perro


perro


gato


perro

gato

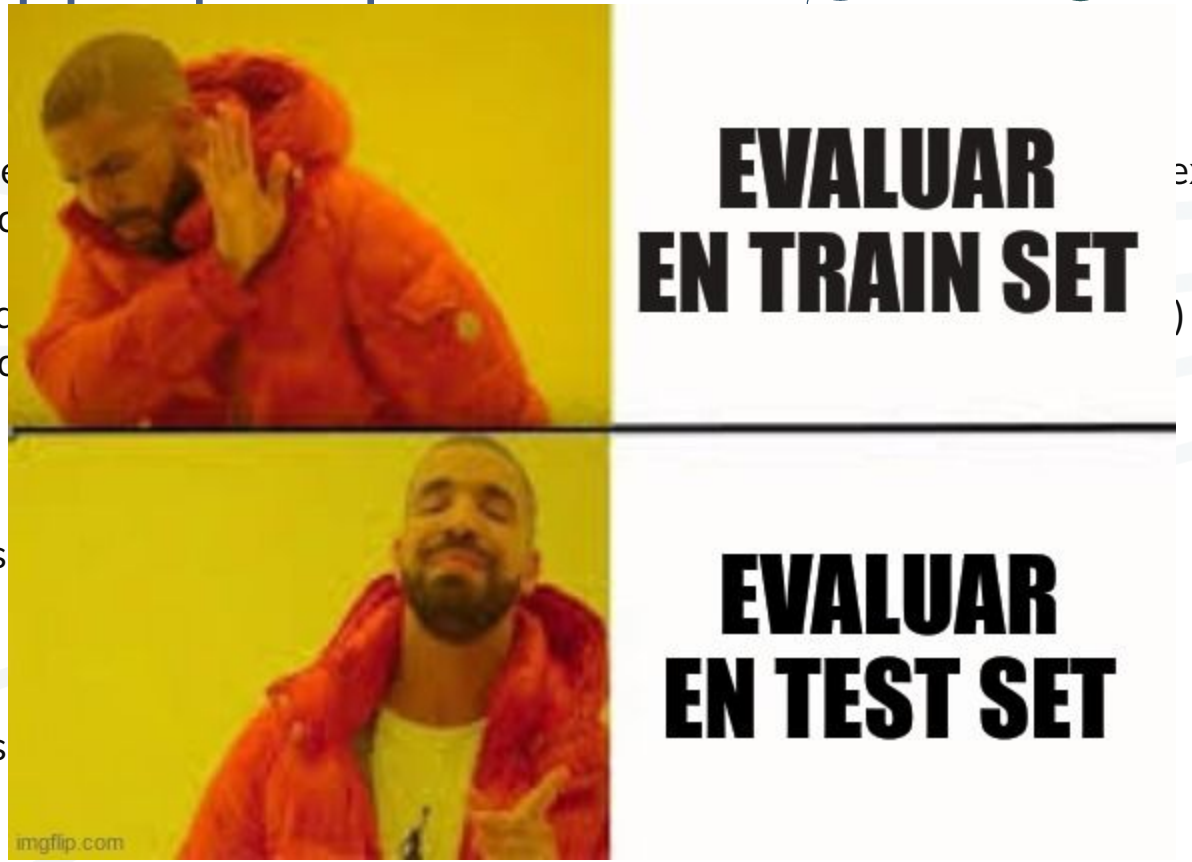
Conjunto de testeo (Test)

Aprendizaje explícita:
Asoci

Dada una c)
Evalu

Evaluar tas

Evaluar tas



Ahora...

Tarea:

Detectar si une niñe está en riesgo de desarrollar dislexia en 1° de escuela {0=no, 1=si}

Rasgos:

Desempeño en 3 tareas psicolingüísticas (en nivel 5).

Clasificador:

Acierto del 95%.

Clasificador f(x) = 0.

A *todas* las entradas les da salida **0**. (no en riesgo)

En el conjunto de validación el 95% de les niñes NO está en riesgo.

Matriz de confusión

GOLD STANDARD o GROUND TRUTH	LO QUE DICE EL MODELO: f(x)=1 (positivo)	LO QUE DICE EL MODELO: f(x)=0 (negativo)
y=1 (positivo)	TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
y=0 (negativo)	FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)

Para tareas dicotómicas

Medidas de desempeño

PRECISIÓN

$$\text{PRECISIÓN} = \frac{TP}{TP + FP} =$$

Cuántos de los que piensa con positivos, son de verdad positivos (VPP: Valor predictivo positivo)

$$\text{VPN} = \frac{TN}{TN + FN} =$$

Cuántos de los que piensa con negativos, son de verdad negativos (Valor predictivo negativo)

Medidas de desempeño

RECALL o COBERTURA

$$\text{ESPECIFICIDAD} = \frac{TN}{TN + FP} =$$

Cuántos de los negativos de verdad, los encuentra negativos

$$\text{RECALL (sensibilidad)} = \frac{TP}{TP + FN} =$$

Cuántos de los positivos de verdad, los encuentra positivos

Ahora...

Tarea:

Detectar si une niñe está en riesgo de desarrollar dislexia en 1° de escuela {0=no, 1=si}

Rasgos:

Desempeño en 3 tareas psicolingüísticas (en nivel 5).

Clasificador:

Acierto del 95%.

Clasificador f(x) = 0.

A *todas* las entradas les da salida **0**. (no en riesgo)

En el conjunto de validación el 95% de les niñes NO está en riesgo.

PRECISIÓN: 0%
VPN: 95%
SENSIBILIDAD: 0%
ESPECIFICIDAD: 100%

TP	FN
FP	TN

0	5
0	95

Ahora...

Tarea:

Detectar si une niñe está en riesgo de desarrollar dislexia en 1° de escuela {0=no, 1=si}

Rasgos:

Desempeño en 3 tareas psicolingüísticas (en nivel 5).

Clasificador:

Acierto del 95%.

Clasificador f(x) = 1.

A *todas* las entradas les da salida **1**. (en riesgo)

En el conjunto de validación el 95% de les niñes NO está en riesgo.

PRECISIÓN: 5%
VPN: 0%
SENSIBILIDAD: 100%
ESPECIFICIDAD: 0%

TP	FN
FP	TN

5	0
95	0

Curva ROC

Hay un compromiso entre especificidad y sensibilidad

Clasificador f(x) = p

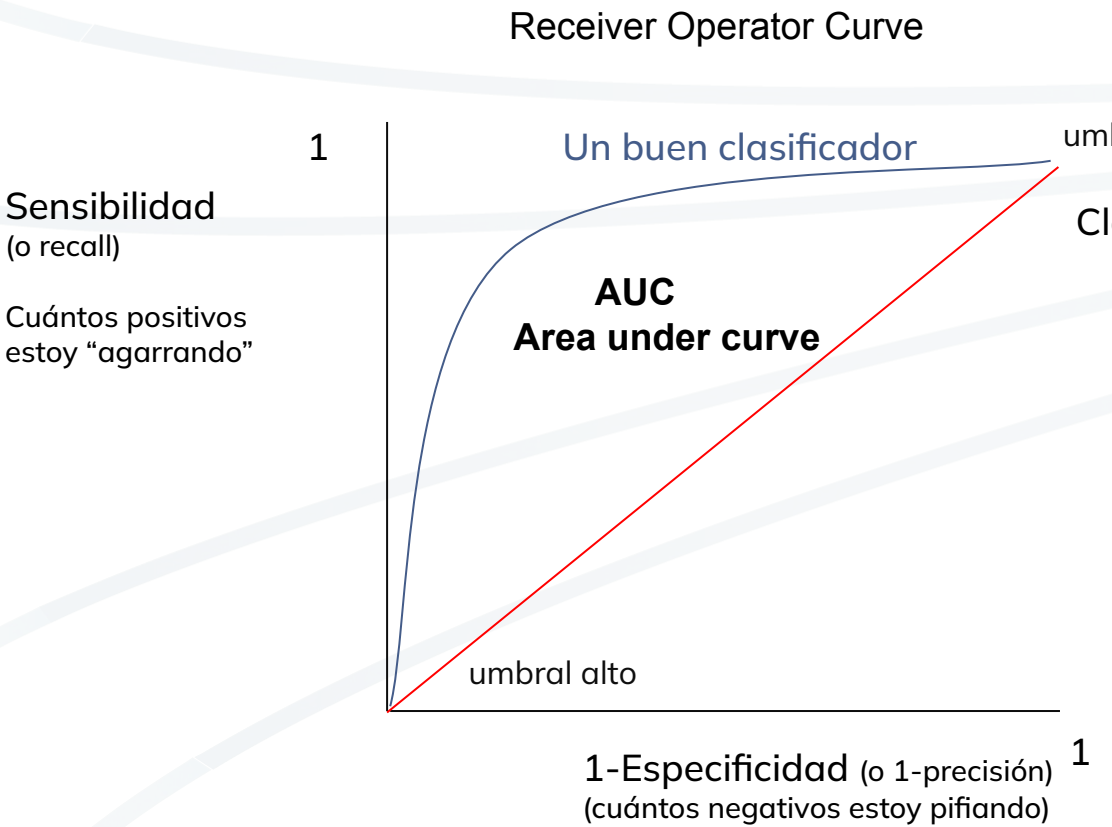
Podemos intepretar p como la “probabilidad”, o el grado de “sospecha” (algo así como el p-valor...)

Para convertir p en {0,1} debo elegir un umbral... (alfa?)

Un umbral **alto**... no encuentro positivos, f(x) casi siempre me termina dando 0 (digo que nadie está en riesgo: especificidad alta, sensibilidad baja)

Un umbral **bajo**... todos los encuentro positivos, f(x) casi siempre me termina dando 1 (todos en riesgo, especificidad baja, sensibilidad alta)

Curva ROC



http://www.navan.name/roc/

Curvas ROC

Hasta ahora

Existen otras formas de medir desempeño de un clasificador

Muchas son útiles cuando las categorías están desbalanceadas

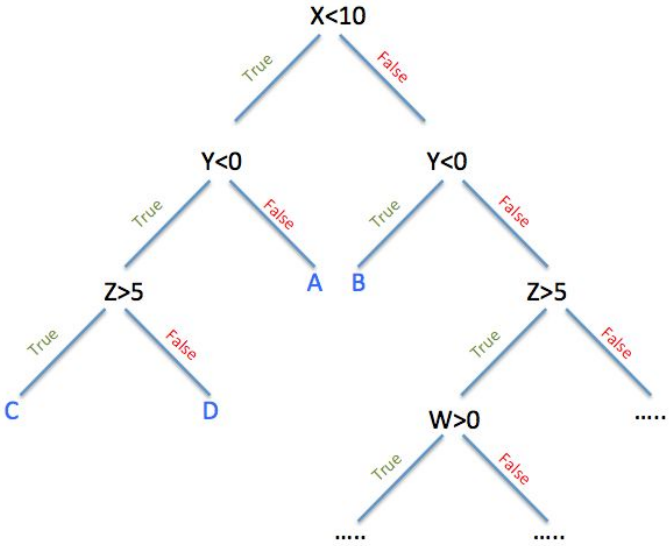
O cuando los valores de predicción están desbalanceados (es peor un falso negativo que un falso positivo).

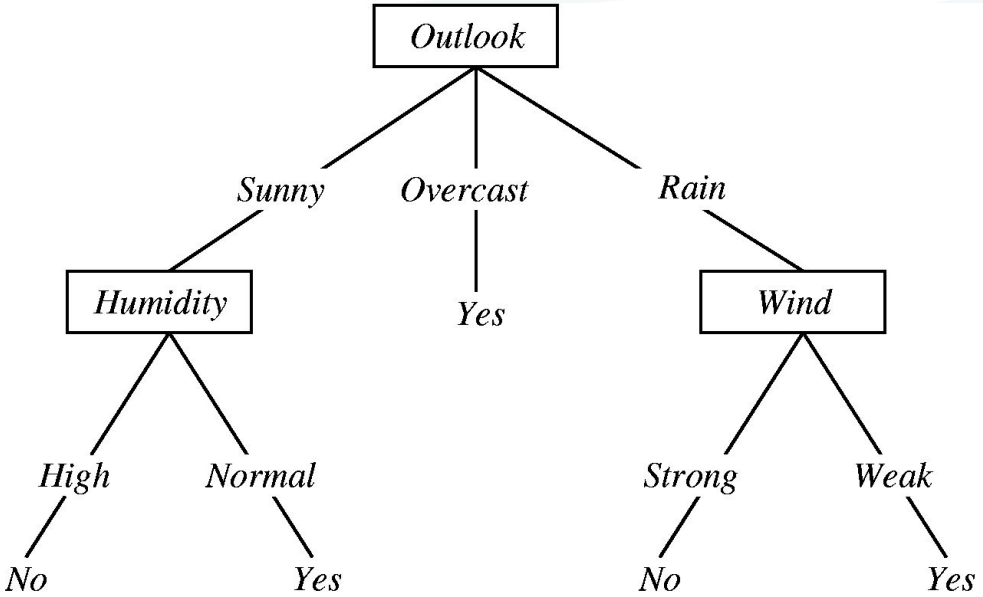
El área bajo una curva ROC (AUC) es un estadístico muy usado para medir la calidad de un clasificador (típicamente binario).

Otras medidas son la media armónica entre Precisión y Cobertura (F1), o el área bajo una curva Precisión-Cobertura (AUCPR).

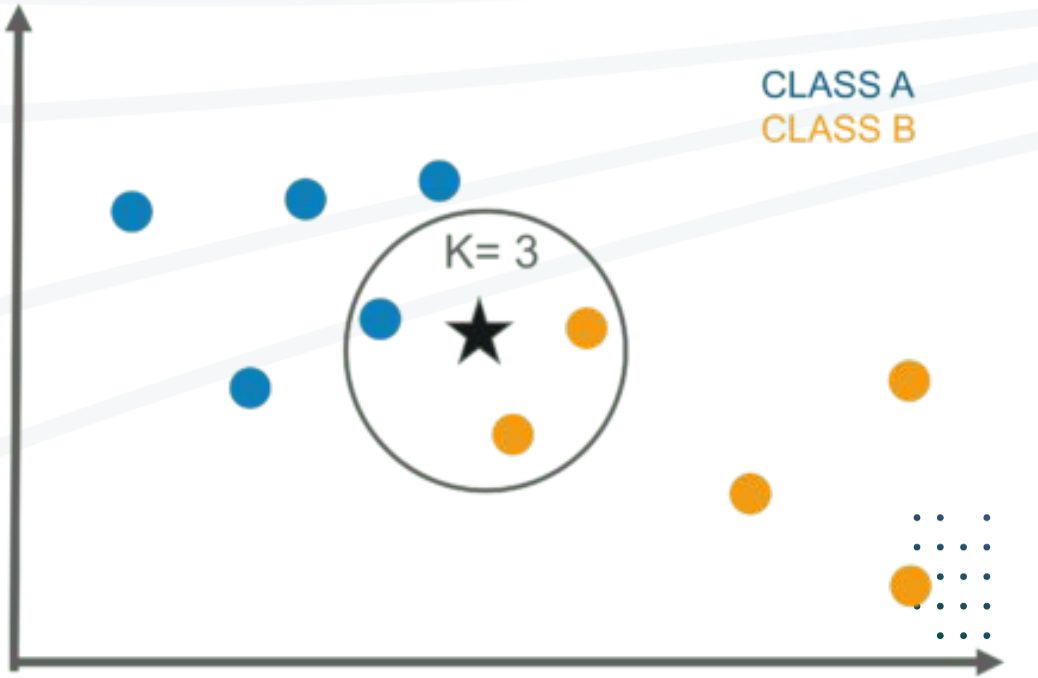
Algunos métodos de clasificación

Árboles de decisión

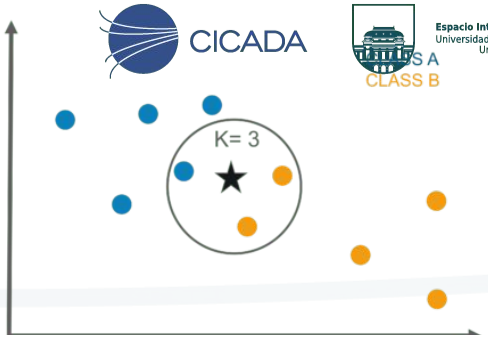


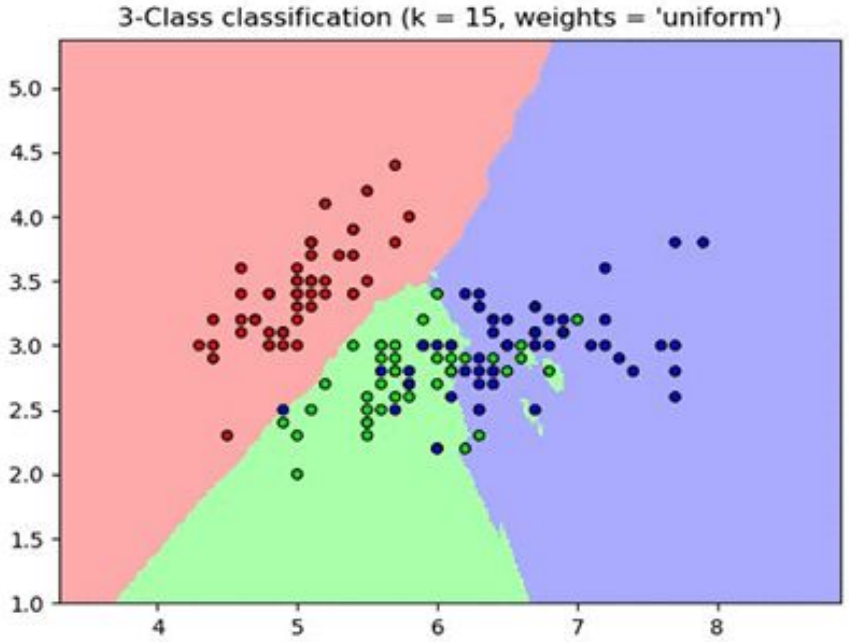


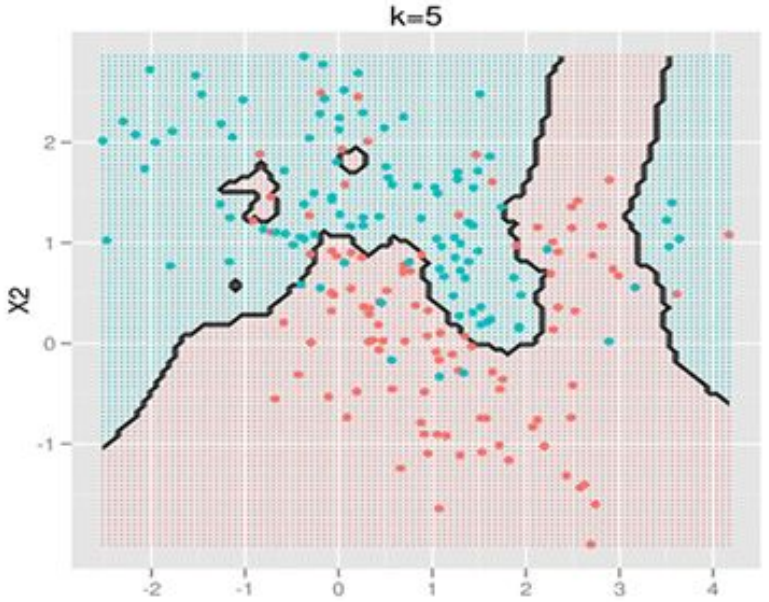
k-nearest neighbors



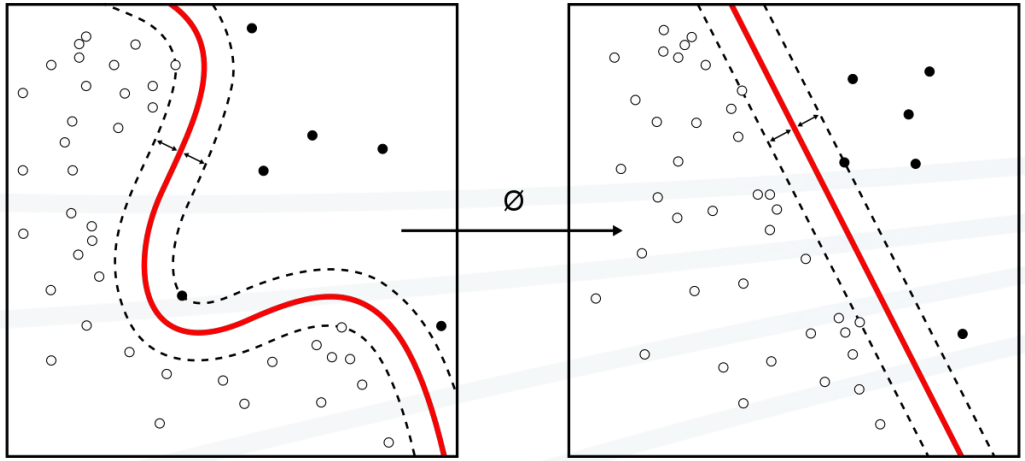
k-nearest neighbors

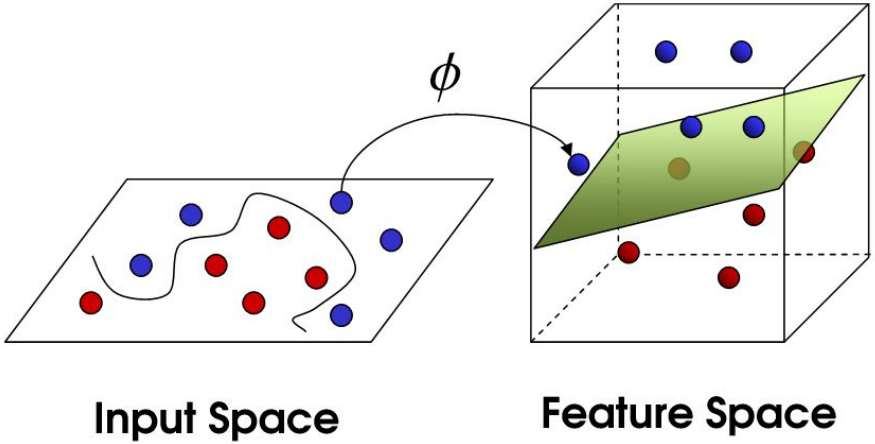




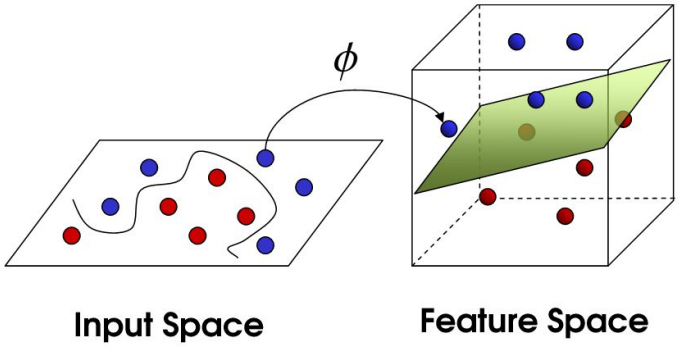


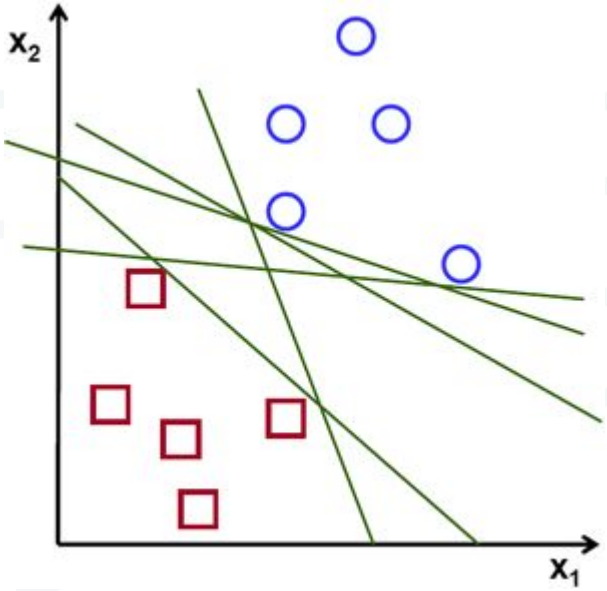
SVM
Máquinas de soporte vectorial

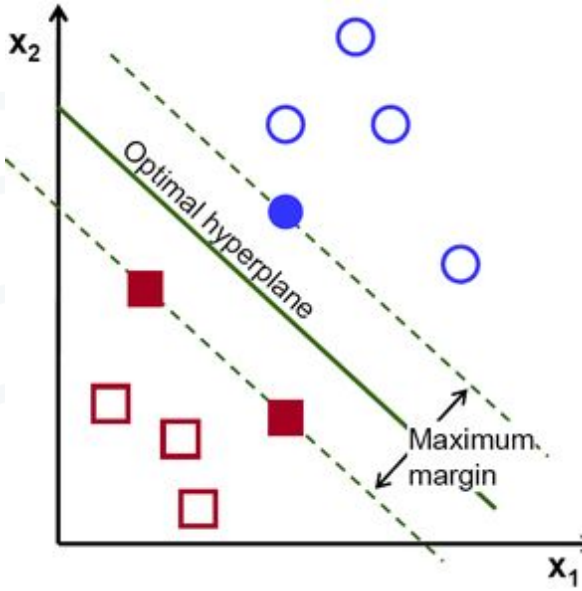




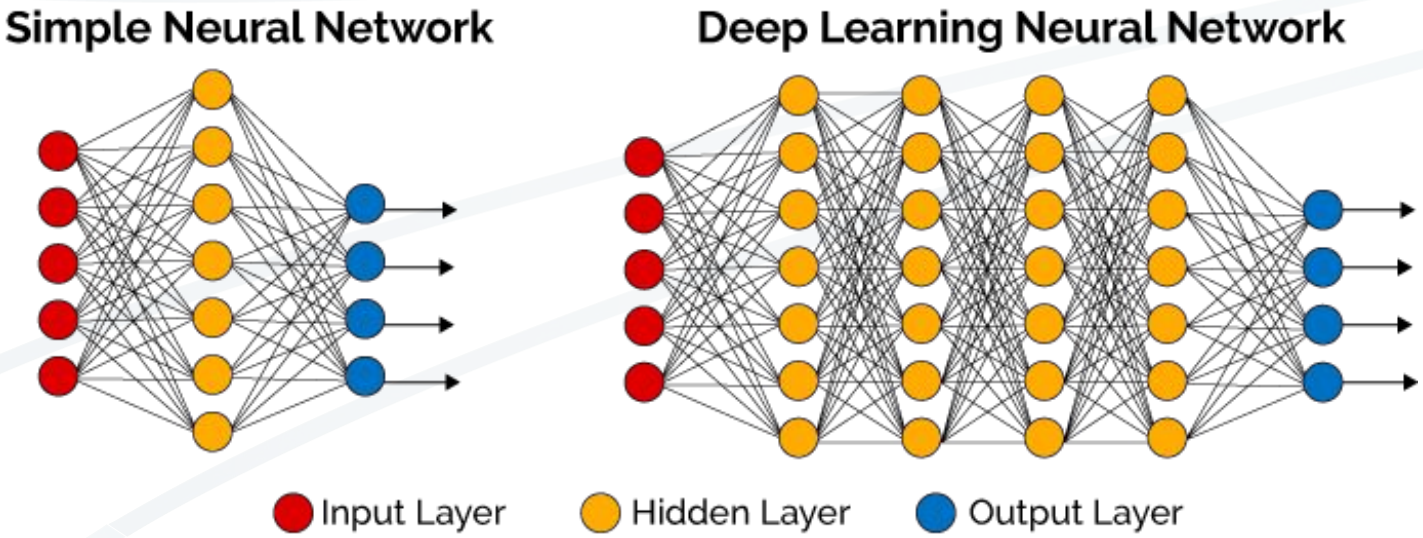
SVM
Máquinas de soporte vectorial







redes neuronales



Parámetros?

La mayoría de los métodos de clasificación (así como los de regresión) tienen parámetros que deben estimarse.

El valor de los parámetros afecta el desempeño del clasificador.

¿Cómo elegir el conjunto óptimo de parámetros?

Aprendizaje Supervisado: Regresión

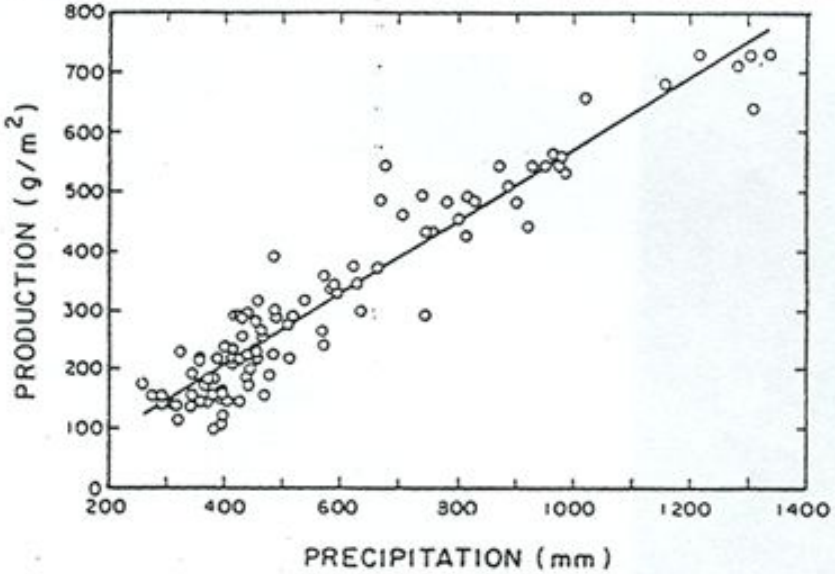
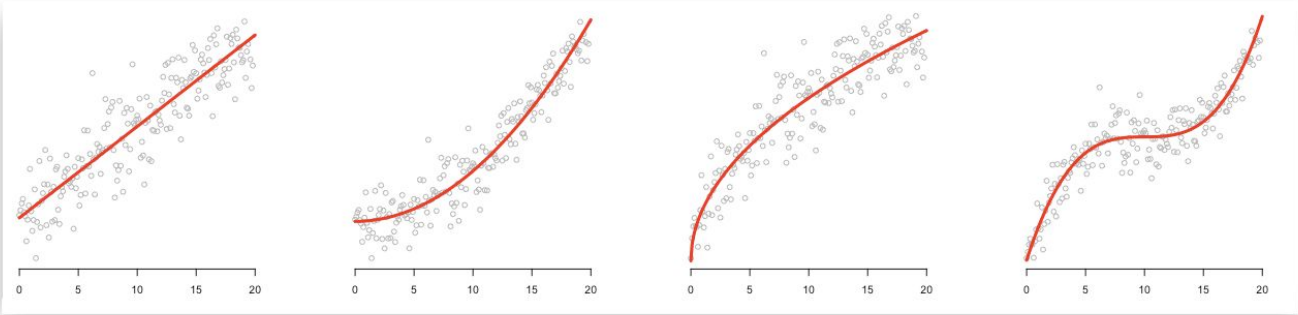


FIG. 2. Relationship between mean annual precipitation and mean aboveground net primary production (ANPP) for 100 major land resource areas across the Central Grassland region. ANPP = −34 + 0.6·APPT; $r^2 = 0.90$.

El agua como recurso limitante para el crecimiento de la vegetación

Aprendizaje Supervisado:
Regresión



How you can use linear regression models to predict quadratic, root, and polynomial functions

Función de pérdida

“Loss function”, función de “error” (o costo), típicamente depende de y e ŷ:

Discrepancia entre actual y predicho (y-ŷ)

En regresión suele usarse mínimos cuadrados ordinarios (OLS, MCO):

$L = \sum (y-\hat{y})^2$

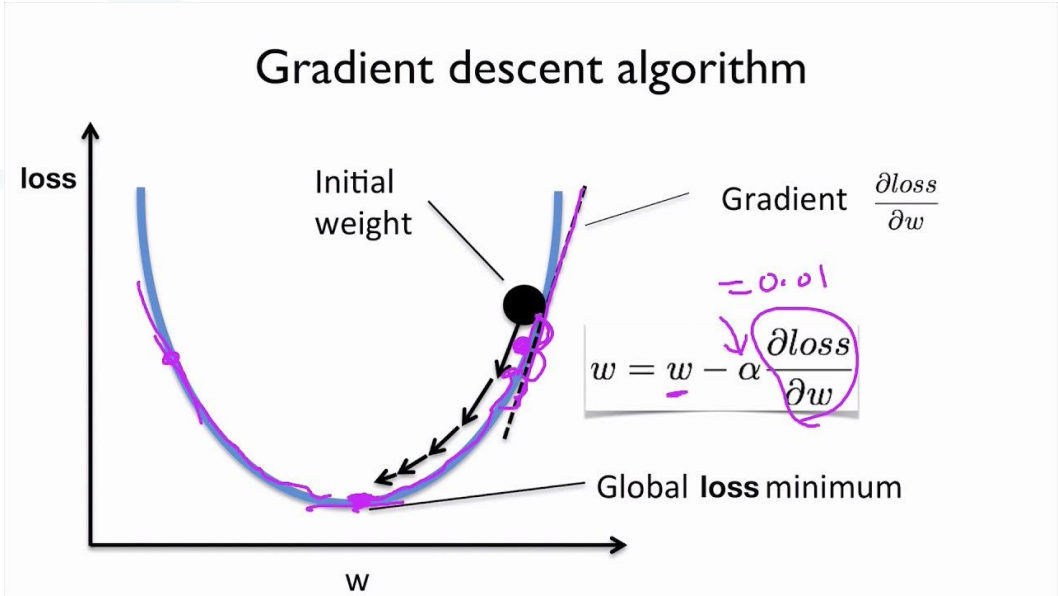
En regresión lineal hay fórmula cerrada (analítica) para los parámetros que *minimizan* el error cuadrático

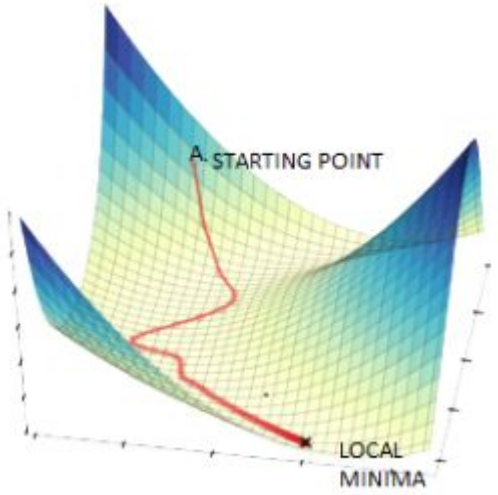
En clasificación suele usarse la entropía cruzada (cross-entropy)

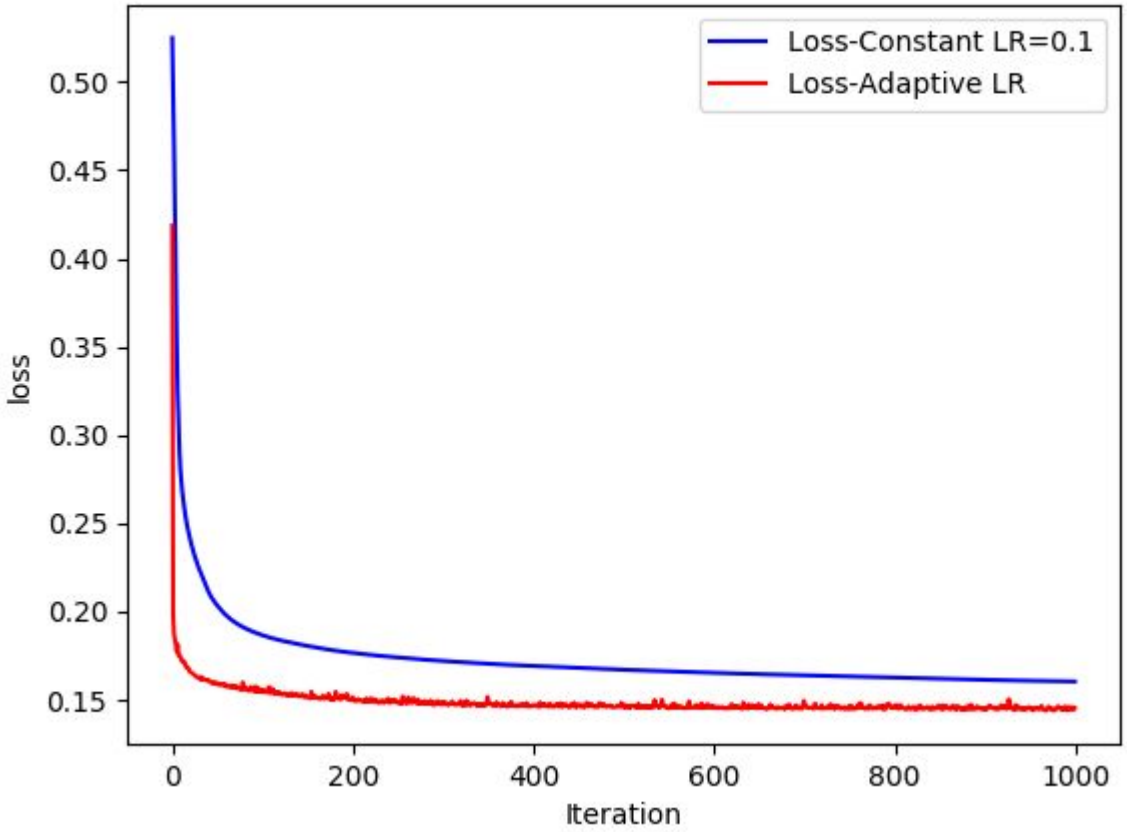
$L = \sum \log(p)$

Si no hay fórmula cerrada para minimizar la pérdida, suelen usarse algoritmos iterativos

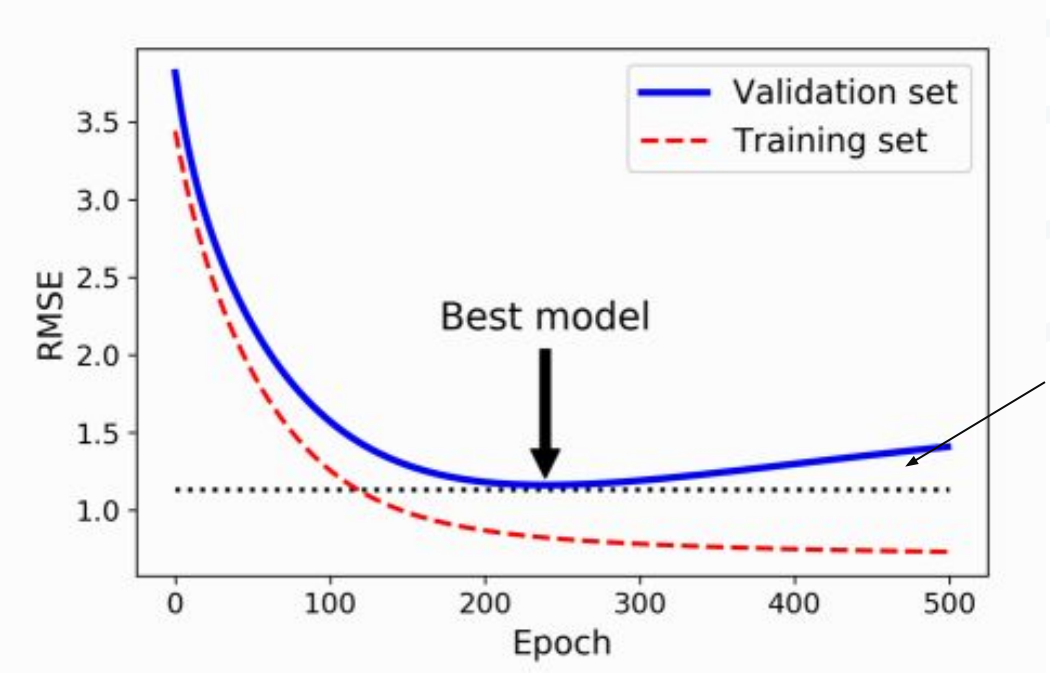
Métodos de gradiente descendiente y similares


Gradient descent algorithm
loss
Initial
weight
Gradient $\frac{\partial loss}{\partial w}$
$= 0.01$
$w = w - \alpha \frac{\partial loss}{\partial w}$
Global loss minimum
w


A STARTING POINT
LOCAL
MINIMA


Loss-Constant LR=0.1
Loss-Adaptive LR
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
loss
0
200
400
600
800
1000
Iteration

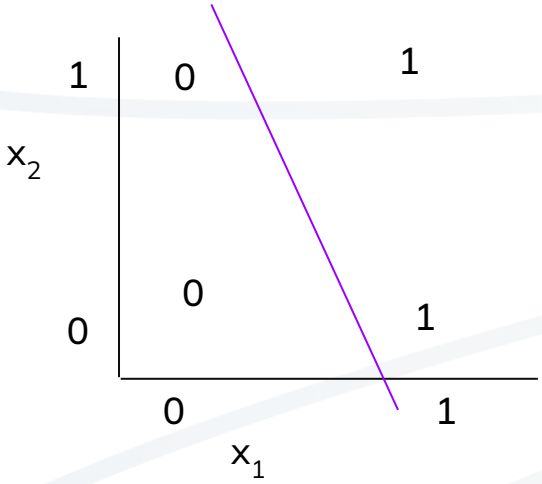
Curva de aprendizaje



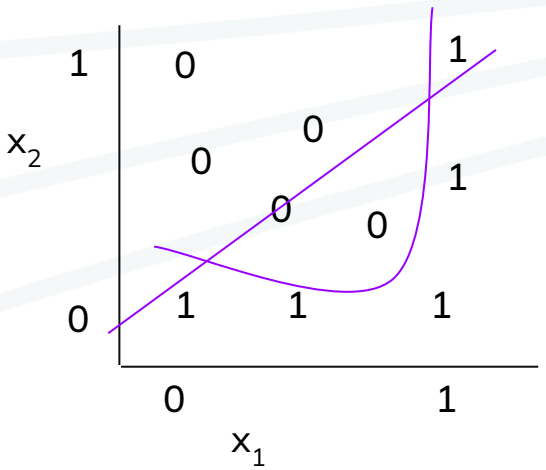
Sobreajuste a los datos de entrenamiento!

Más parámetros, mejor?

Ejemplo



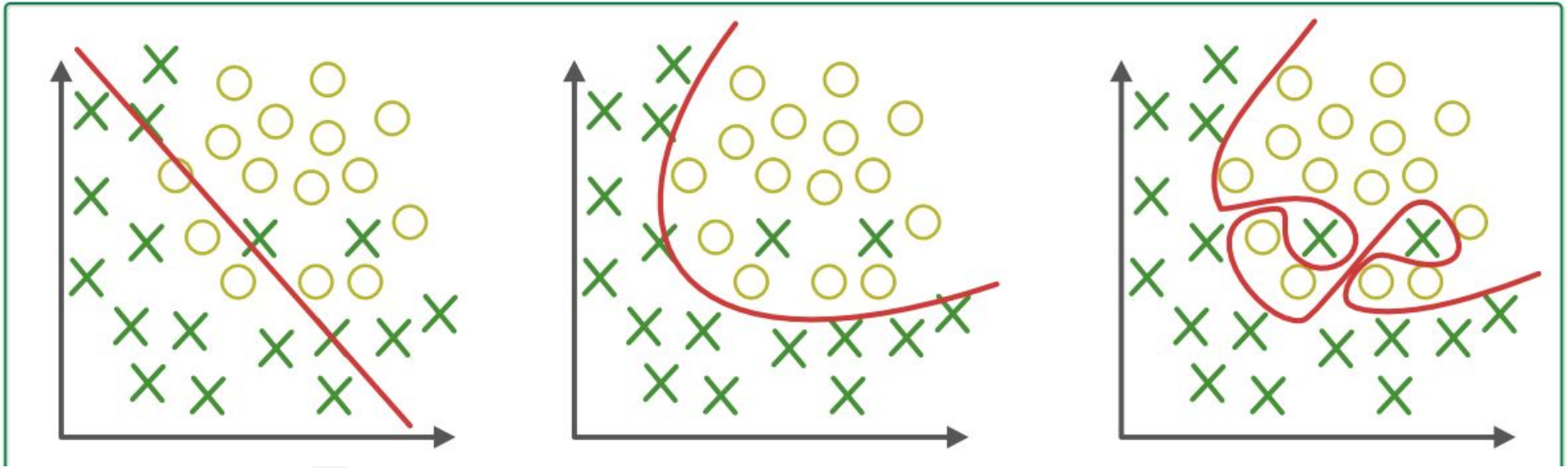
$\hat{y} = H(a_1x_2 + a_2x_2 + b)$



$\hat{y} = H(a_1x_2 + a_2x_2 \ \ a_{12}x_1x_2 + b)$

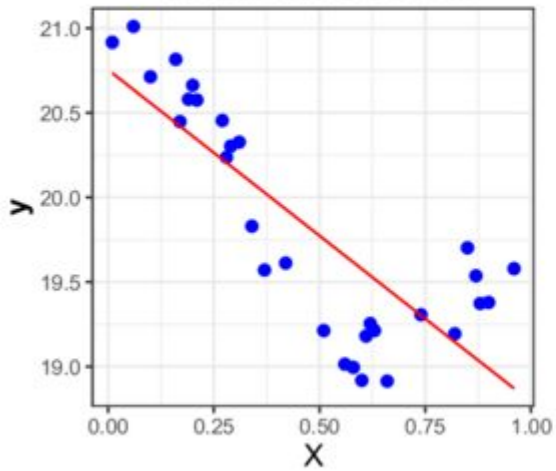
Complejidad

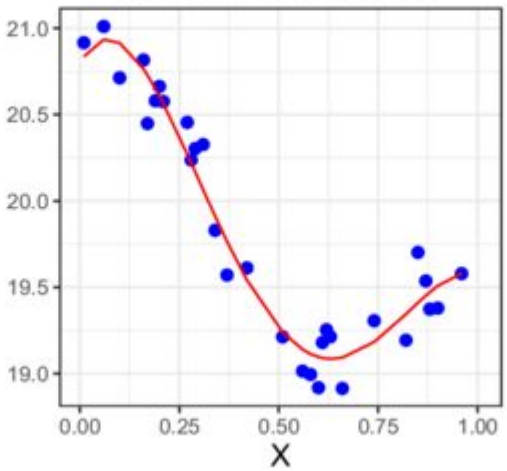
Para una misma clase de modelos, mayor cantidad de parámetros nos da un mejor ajuste (menor pérdida)

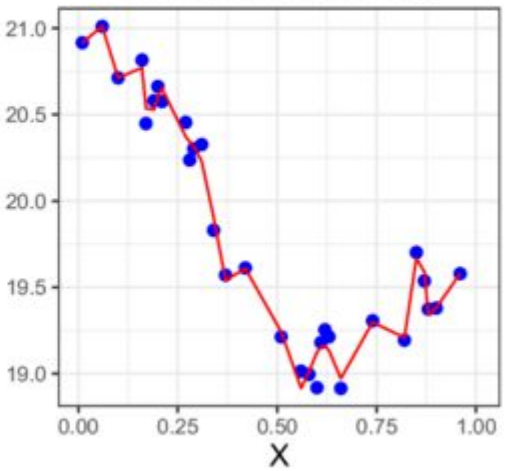


Complejidad

Para una misma clase de modelos, mayor cantidad de parámetros nos da un mejor ajuste (menor pérdida)

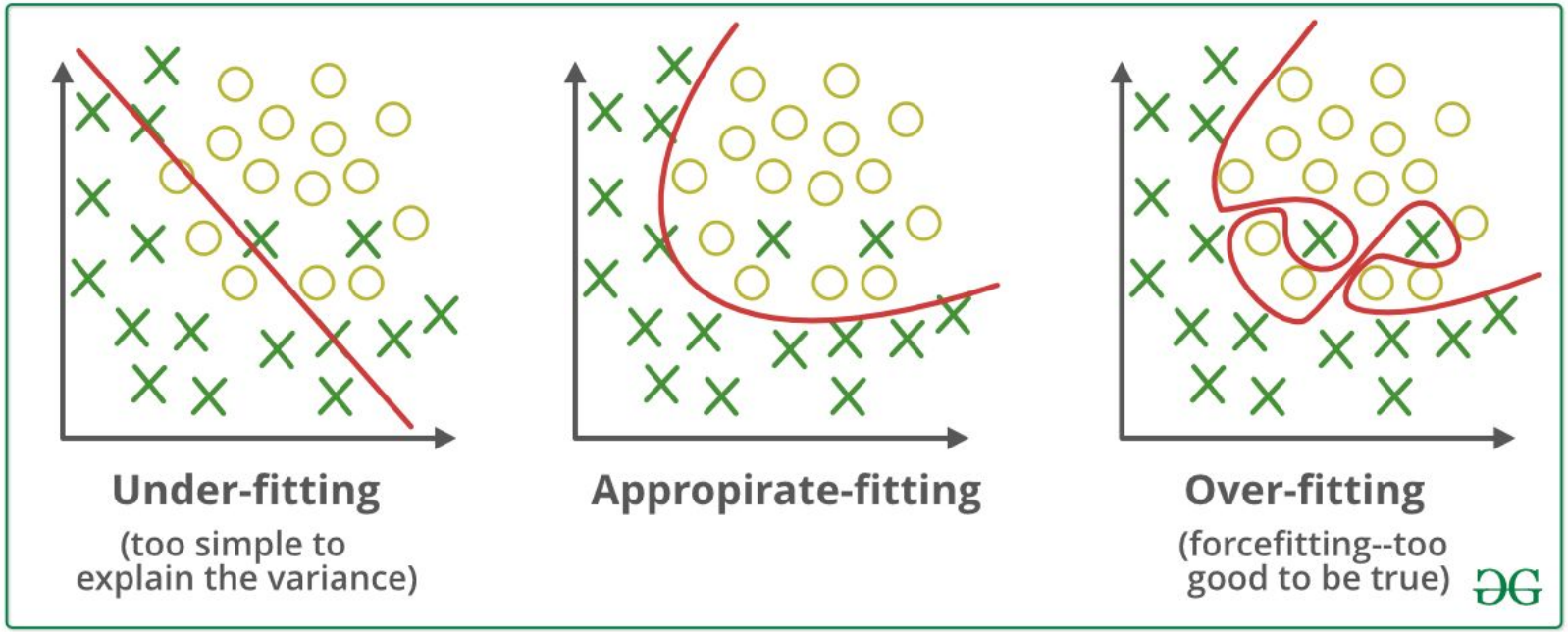






Complejidad

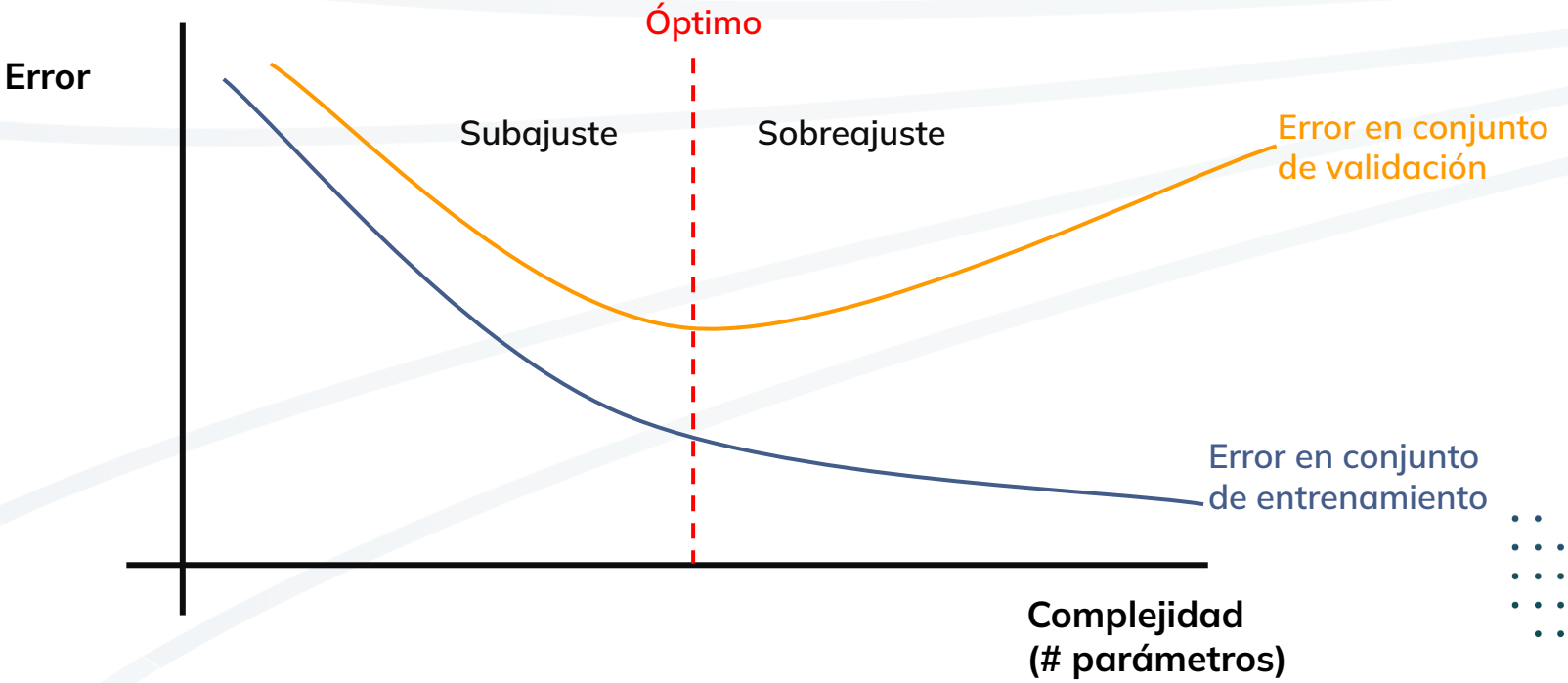
Para una misma clase de modelos, mayor cantidad de parámetros nos da un mejor ajuste (menor pérdida)



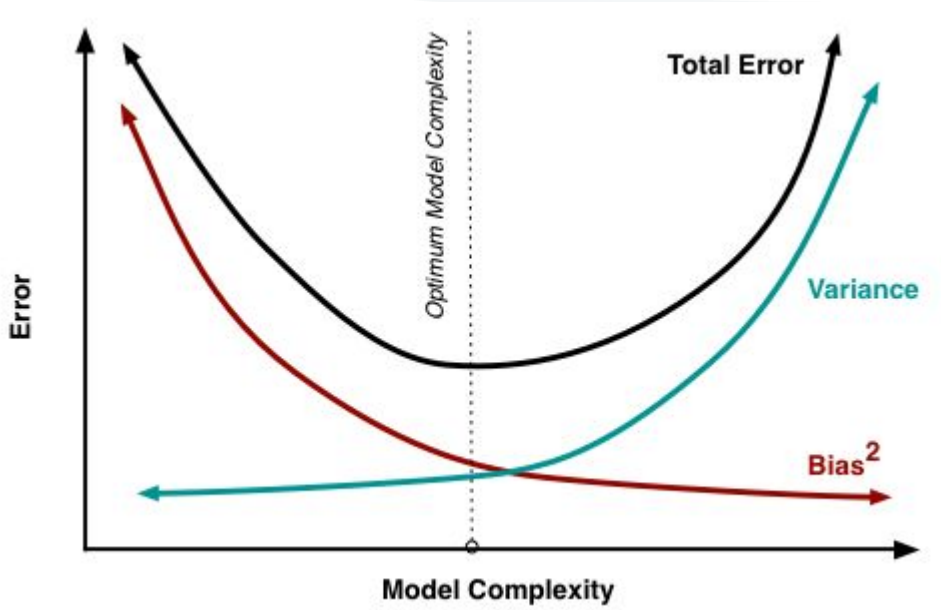
Mucho error de sesgo

Mucho error de varianza

La curva de complejidad



Trade-off: sesgo y varianza



Componentes del error:

Sesgo: error debido al "sesgo" que impone un modelo simple.

Varianza: error debido a la variabilidad en el conjunto de datos.

Regularización

Se puede reducir el sobreajuste inducido por un modelo de mayor complejidad?

Regularización

Introducir penalización en la función de pérdida

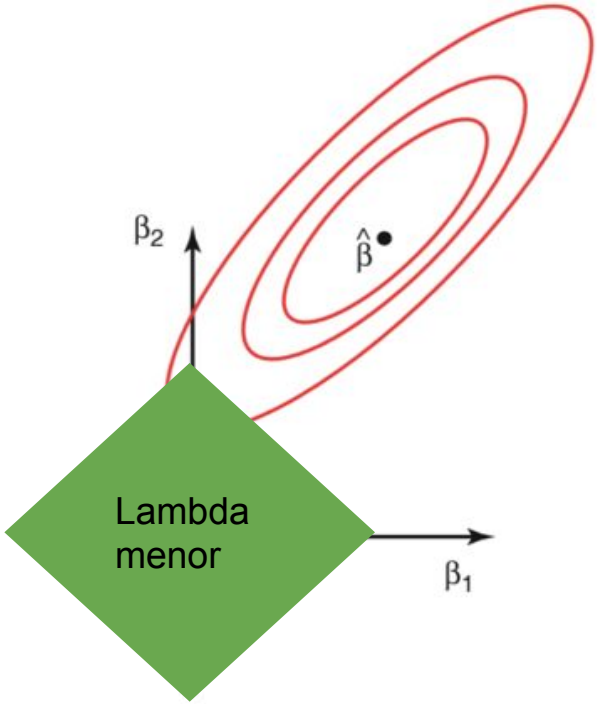
Sesgar a favor de soluciones de parámetros que sean más sencillas

Regresión de Tikhonov → Ridge, o LASSO

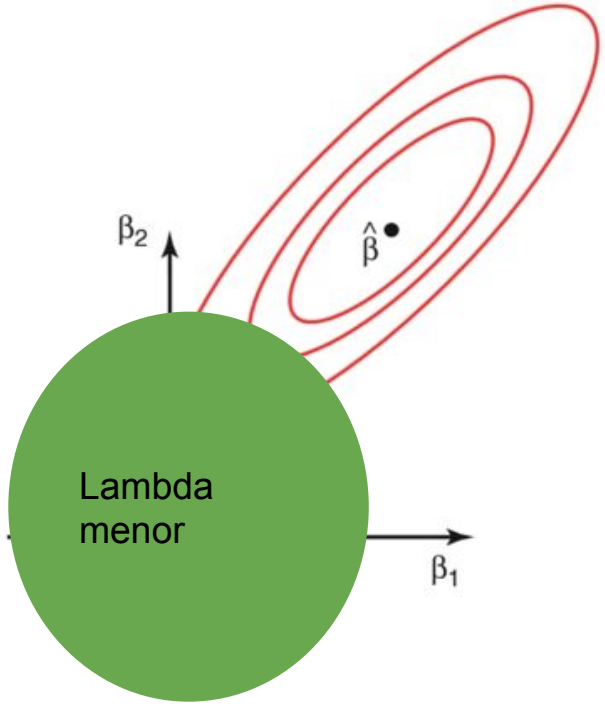
Función de pérdida depende de la norma del vector de parámetros.

$L = \sum (y-\hat{y})^2 + \lambda ||w||^2_p$ p=1, LASSO, p=2, Ridge, λ: hiperparámetro de regularización

Regularización

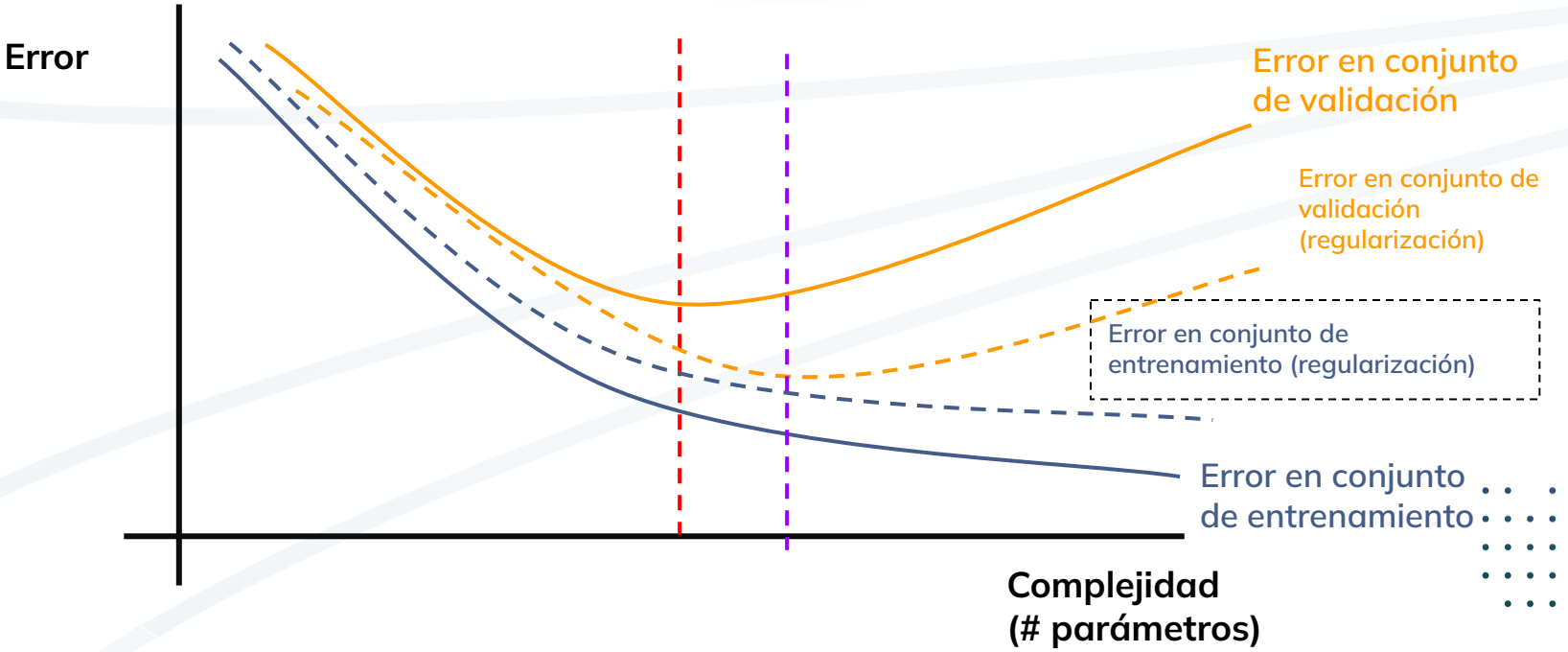


LASSO



Ridge

La curva de complejidad



Hasta ahora

Muchas veces encontrar el modelo óptimo no tiene forma analítica cerrada

- Proceso iterativo (aprendizaje)
- Ajuste de los parámetros del modelo
- Función de pérdida
- Gradiente-descendiente o similar
- Puede encontrar mínimos locales, o sobreajustar al conjunto de entrenamiento

Diferentes modelos tienen distinta complejidad

- Mayor complejidad, mejor ajuste al conjunto de entrenamiento
- Peligro de sobreajuste (error de generalización)
- Balance sesgo-varianza (simplicidad modelo vs error por sobreajuste)
- Regularización permite encontrar mejores óptimos, restringiendo grados de libertad en los parámetros

Preguntas

CICADA

Espacio Interdisciplinario
Universidad de la República
Uruguay

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY