

# Fundamentos de Aprendizaje Estadístico

1EST17 - Aprendizaje Estadístico I

Enver G. Tarazona

2022-09-10



# Introducción

## Objetivos de la unidad

- Aprendizaje de máquinas vs. aprendizaje estadístico: ejemplos
- Introducir notación relevante y terminología.
- Estimar  $f$  (regresión, clasificación): precisión en la predicción vs. interpretabilidad del modelo
- Balance del Sesgo-Varianza

# ¿Qué es aprendizaje estadístico?

*Aprendizaje estadístico* es el proceso de aprendizaje a partir de los datos.

Aplicando *métodos estadísticos* a un *conjunto de datos* (llamado el *conjunto de entrenamiento*), nosotros podemos:

- *extraer conclusiones* acerca de las relaciones entre las variables (*inferencia*) o
- *encontrar una función predictiva* para nuevas observaciones. (*predicción*).

Además, nos gustaría encontrar estructuras en los datos que nos ayuden a aprender algo sobre el mundo real.

El aprendizaje estadístico juega un rol muy importante en muchas áreas del conocimiento como en ciencias, finanzas y la industria.

## Ejemplos de problemas de aprendizaje

- Predecir el precio de una acción dentro de 3 meses, en función de las medidas de desempeño de la empresa y los datos económicos. La variable de respuesta es cuantitativa (precio). *Respuesta continua.*
- Detección de spam para correos electrónicos. *Respuesta binaria* (si, no).
- Identificación de factores de riesgo de diabetes. *Respuesta Binaria* (si, no).
- Estimación del riesgo de enfermedad cardíaca o ataque cardíaco, dado el conocimiento sobre la condición, el comportamiento, la edad o las mediciones demográficas, dietéticas y clínicas. *Respuesta binaria* (si, no).
- Reconocimiento de dígitos e imágenes. *Respuesta categórica.*

Datos de Sudáfrica sobre enfermedades cardíacas: 462 observaciones y 10 variables.





## Clasificación de correo electrónico (detección de spam):

El objetivo es construir un filtro de spam. Este filtro puede basarse en las frecuencias de palabras y caracteres en los correos electrónicos. La siguiente tabla muestra el porcentaje promedio de palabras o caracteres en un mensaje de correo electrónico, basado en 4601 correos electrónicos de los cuales 1813 se clasificaron como spam.

	you	free	george	!	\$	edu
not spam	1.27	0.07	1.27	0.11	0.01	0.29
spam	2.26	0.52	0.00	0.51	0.17	0.01



P: ¿Hubieron objetivos y elementos subyacentes comunes en estos ejemplos de aprendizaje?

- Predecir el precio de una acción dentro de 3 meses a partir de las medidas de desempeño de la compañía y los datos económicos.
- Identificar los factores de riesgo para desarrollar diabetes en función de la dieta, la actividad física, los antecedentes familiares y las mediciones corporales.
- El objetivo es construir un filtro de spam.
- Predecir si alguien sufrirá un ataque cardíaco sobre la base de mediciones demográficas, dietéticas y clínicas.
- Identificar los números en un código postal escrito a mano, a partir de una imagen digitalizada.

R:

Sí, el objetivo era la explicación o la predicción, algunas variables de salida eran cualitativas (categóricas) y otras cuantitativas.

# El problema del Aprendizaje Supervisado

## *Punto de partida:*

- Medición del resultado  $Y$ , también llamada variable dependiente, respuesta, objetivo.
- Vector of  $p$  mediciones predictoras  $X = (X_1, \dots, X_p)$ , también llamados entradas, regresores, covariables, características, variables independientes.
- En un **problema de regresión**,  $Y$  es cuantitativa (por ejemplo, precio, presión arterial).
- En un **problema de clasificación**,  $Y$  toma valores en un conjunto finito y desordenado (sobrevivió/ murió, dígito 0-9, clase de cáncer de muestra de tejido).
- Tenemos datos de entrenamiento  $(x_1, y_1), \dots, (x_N, y_N)$ . Estas son observaciones (ejemplos, instancias) de estas medidas.

## El Aprendizaje Supervisado y sus objetivos

Nuestro conjunto de datos (conjunto de entrenamiento) consta de  $n$  medición de la variable de respuesta  $Y$  y de  $p$  covariables  $x$ :

$$(y_1, x_{11}, x_{12}, \dots, x_{1p}), (y_2, x_{21}, \dots, x_{2p}), \dots, (y_n, x_{n1}, x_{n2}, \dots, x_{np}).$$

Sobre la base de los *datos de entrenamiento*, nos gustaría:

- **predecir con precisión** nuevos casos de prueba.
- **comprender** qué entrada afecta los resultados y cómo.
- **evaluar la calidad** de las predicciones e inferencias.

Se llama aprendizaje supervisado porque la variable de respuesta *supervisa nuestro análisis*.

# El Problema del Aprendizaje no Supervisado

- No hay **ninguna variable respuesta**  $y$ , solo un conjunto de predictores (características)  $x_i$  medidos en un conjunto de objetos.
- El objetivo es más difuso: buscar patrones o agrupaciones (ocultas) en los datos, para *obtener información y comprensión*. No hay una respuesta *correcta*.
- Es difícil saber qué tan bien se está realizando la tarea.

Ejemplos en el curso:

- Conglomerados
- Análisis de Componentes Principales
- Reglas de Asociación

# Filosofía general

- Es importante entender primero los métodos más simples para comprender los más sofisticados.
- Es importante evaluar con precisión el desempeño de un método, para saber qué tan bien o qué tan mal está funcionando.

→ **¡Los métodos más simples a menudo funcionan tan bien como los más elegantes!**

- El aprendizaje estadístico es un ingrediente fundamental en la formación de un científico de datos moderno.

# Aprendizaje Estadístico vs. Aprendizaje Automático

- El aprendizaje automático surgió como un subcampo de la inteligencia artificial.
- El aprendizaje estadístico surgió como un subcampo de la estadística.
- Hay mucha superposición: ambos campos se centran en problemas supervisados y no supervisados:
  - El aprendizaje automático tiene un mayor énfasis en las *aplicaciones a gran escala y precisión de la predicción*.
  - El aprendizaje estadístico enfatiza los modelos y su *interpretabilidad*, precisión e incertidumbre.
- La distinción se ha vuelto cada vez más borrosa y hay una gran cantidad de superposición de términos y modelos.
- ¡El aprendizaje automático tiene la ventaja en marketing!

- Existe una controversia y cierto escepticismo en contra de los métodos ML “demasiado sofisticados”.
- Crítica: ML a menudo reinventa métodos existentes y los nombra de manera diferente, pero a menudo sin conocimiento de los métodos existentes en estadística.
- Literatura nueva casi semanal que ofrece comparación. A menudo, los métodos estadísticos “simples” “ganan”.

RESEARCH ARTICLE

## Statistical and Machine Learning forecasting methods: Concerns and ways forward

Spyros Makridakis<sup>1</sup>, Evangelos Spiliotis<sup>2\*</sup>, Vassilios Assimakopoulos<sup>2</sup>

<sup>1</sup> Institute For the Future (IFF), University of Nicosia, Nicosia, Cyprus, <sup>2</sup> Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Zografou, Greece

\* [spiliotis@fsu.gr](mailto:spiliotis@fsu.gr)



Un tweet de uno de los coautores del libro del curso, casualmente...



**Daniela Witten**

@daniela\_witten

...

"When we raise money it's AI, when we hire it's machine learning, and when we do the work it's logistic regression."

(I'm not sure who came up with this but it's a gem 💎)

1:50 PM · Sep 26, 2019 · Twitter Web App

## Términos usados

Aprendizaje estadístico	Aprendizaje automático
modelo	red, gráfico, mapeo
ajustar, estimar	aprender
covariables, entradas, variables independientes, predictoras	atributos, predictoras
variable respuesta, de salida, dependiente	salida, objetivo
conjunto de datos	datos de entrenamiento

Observación: no es una lista exhaustiva, y muchos términos se usan de la misma manera en ambos campos.

## ¿Cuál es el objetivo del aprendizaje estadístico?

Asumimos que<sup>1</sup> :

- Se observa una variable respuesta *cuantitativa*  $Y$  y
- $p$  predictores diferentes  $x_1, x_2, \dots, x_p$ .

Se asume que existe una función  $f$  que relaciona la respuesta con las variables predictoras:

$$Y = f(x) + \varepsilon,$$

donde  $\varepsilon$  es un término aleatorio (error) con media 0 e independiente de  $x$ .

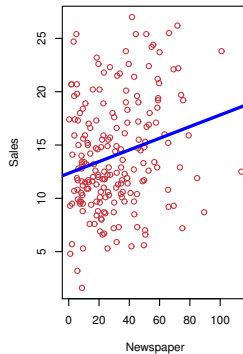
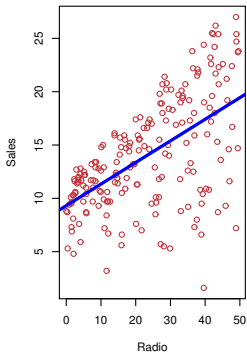
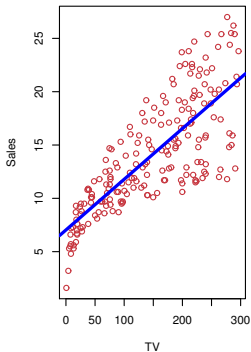
**{El objetivo es estimar  $f$ .}**

---

<sup>1</sup>Para un problema de aprendizaje supervisado.

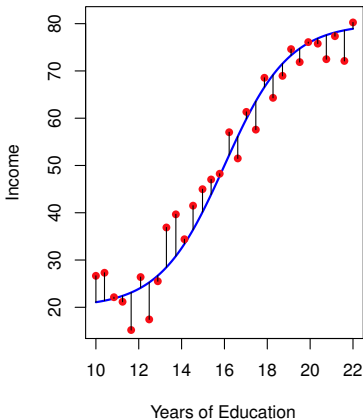
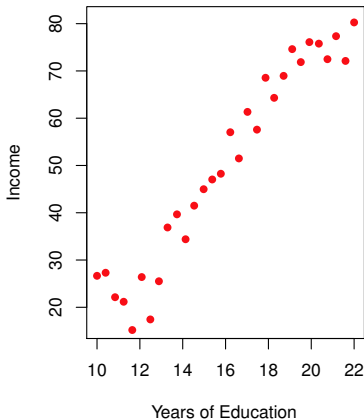
# Ejemplo 1

Ventas de un producto en función de los gastos en publicidad en diferentes medios de comunicación.



## Ejemplo 2

Ingreso para distintos niveles de educación.



Existen dos principales motivos para estimar  $f$ :

- **Predicción**
- **Inferencia**

## Motivo 1: Predicción

**Objetivo:** Predecir la respuesta  $Y$  dadas nuevas observaciones  $x$  de las covariables con la mayor precisión posible.

Notación:

$$\hat{Y} = \hat{f}(x).$$

Donde:

- $\hat{f}$ : la estimación de  $f$
- $\hat{Y}$  predicción de  $Y$  dado  $x$ .
- No es realmente de interés la forma de  $f$  (“caja negra”).

→ ¡no se interpretan los parámetros de regresión cuando el objetivo es exclusivamente predictivo! (la precisión de la predicción es importante.)

Existen dos términos que influyen en la precisión de  $\hat{Y}$  como predicción de  $Y$ :

- El *error reducible* tiene que ver con nuestra estimación  $\hat{f}$  de  $f$ . Este error puede reducirse usando la técnica *apropiada* de aprendizaje supervisado.
- El *error irreducible* proviene del término del error  $\epsilon$  y no puede reducirse mejorando  $f$ . Esto está relacionado con las cantidades no observadas que influyen en la respuesta y posiblemente la aleatoriedad de la situación.

Para una  $\hat{f}$  dada y un conjunto de predictores  $X$  con los que se obtiene  $\hat{Y} = \hat{f}(X)$ , tenemos

$$E[(Y - \hat{Y})^2] = \underbrace{E[(f(X) - \hat{f}(X))^2]}_{\text{reducible}} + \underbrace{Var(\epsilon)}_{\text{irreducible}}$$



P: Si hubiera una relación *determinista* entre la respuesta y un conjunto de predictores, ¿habría un error tanto reducible como irreducible?

R:

Si conocemos todos los predictores y la conexión (determinista) a la respuesta, y no se agrega ningún error aleatorio, entonces no tendremos un error irreducible. Si hay una relación determinista, pero no conocemos todos los valores predictores, los predictores no observados nos darán un error irreducible.

Entonces, muy raramente habrá solo un error reducible presente.

## Motivo 2: Inferencia

**Objetivo:** *Comprender* cómo la variable respuesta se ve afectada por los diversos predictores (covariables).

Nuestro *principal interés*. es la *forma exacta* de  $\hat{f}$ :

- ¿Qué predictores están asociados con la respuesta?
- ¿Cuál es la relación entre la respuesta y cada predictor?
- ¿Puede la relación ser lineal o se necesita un modelo más complejo?

## Regresión y Clasificación

**Regresión** predice un valor numérico.

Ejemplo: Predecir la ganancia dada la cantidad de dinero gastado en publicidad.

**Clasificación** predice la clase de pertenencia.

Ejemplo: dada la presión arterial, el peso y la proporción de cadera predicen si un paciente sufre de diabetes (sí / no).

P:

Dé un ejemplo de una regresión y un problema de clasificación (problema práctico con el conjunto de datos disponible) que le gustaría estudiar en este curso.

R:

# Estimación de $f$

Idea general:

- Usar un conjunto de *datos de entrenamiento* disponible  $(x_1, y_1), \dots, (x_n, y_n)$  para estimar  $\hat{f}$ , tal que  $Y \approx \hat{f}(X)$  para cualquier  $(X, Y)$  (incluso para aquellos que no han sido aún observados).

Dos enfoques principales:

- Métodos paramétricos
- Métodos no paramétricos

# Métodos paramétricos

Los métodos paramétricos se basan en un supuesto sobre la forma y estructura de la función  $f$ .

El modelo de regresión lineal múltiple es un ejemplo de modelo paramétrico. Suponemos aquí que la variable respuesta es una combinación lineal de las covariables con algo de ruido adicional.

$$f(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon.$$

with  $\varepsilon \sim N(0, \sigma^2)$ .

Al hacer esta suposición, la tarea se simplifica en encontrar estimaciones de  $p + 1$  coeficientes  $\beta_0, \beta_1, \dots, \beta_p$ . Para hacer esto, usamos los datos de entrenamiento para ajustar el modelo, de modo que

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p.$$

La estimación de un modelo paramétrico se divide en dos etapas:

1. Seleccionar una forma para la función  $f$ .
2. Estimar los parámetros desconocidos de  $f$  usando el conjunto de entrenamiento.

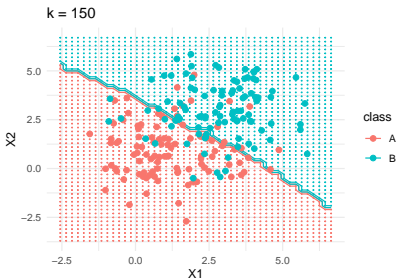
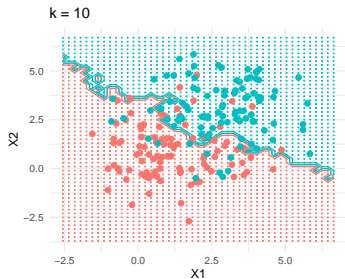
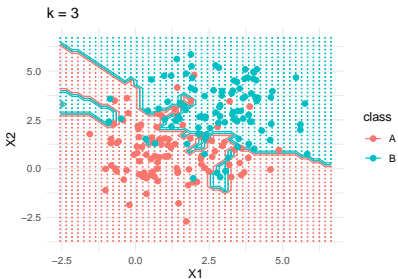
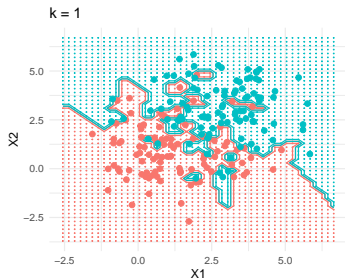
## Métodos no paramétricos

Los métodos no paramétricos buscan una estimación de  $f$  que se acerca a los datos, pero sin hacer suposiciones explícitas sobre la forma de la función  $f$ .

El algoritmo de los  $K$ -vecinos más cercanos es un ejemplo de un modelo no paramétrico. Utilizado mayormente en clasificación, este algoritmo predice una membresía de clase para una nueva observación al hacer un voto mayoritario basado en las  $K$  observaciones más cercanas.



# Ejemplo de KNN



P: ¿Cuáles son las ventajas y desventajas de los métodos paramétricos y no paramétricos?

Sugerencias: interpretabilidad, cantidad de datos necesarios, complejidad, suposiciones hechas, precisión de predicción, complejidad computacional, sobre/sub ajuste.

## R: Métodos paramétricos

Ventajas	Desventajas
Simple de usar y fácil de entender.	La función $f$ está restringida a la forma especificada.
Requiere pocos datos de entrenamiento	La forma de la función $f$ que fue asumida en general no coincidirá con la función verdadera, lo que puede dar una estimación pobre.
Computacionalmente barato	Complejidad limitada

## R: Métodos no paramétricos

Ventajas	Desventajas
Flexible: se puede adaptar una gran cantidad de formas funcionales	Puede sobreajustar los datos
No se hacen suposiciones fuertes sobre la función subyacente	Computacionalmente más caro ya que se deben estimar más parámetros
A menudo puede dar buenas predicciones	Se requieren muchos datos para estimar $f$ (compleja).

# Precisión de la predicción vs. interpretabilidad

(nos estamos preparando para hablar sobre el balance sesgo/varianza)

Los métodos **inflexibles**, o rígidos, son métodos que tienen fuertes restricciones en la forma de  $f$ .

Ejemplos:

- Regresión lineal
- Análisis discriminante lineal
- Selección por subconjuntos y Lasso

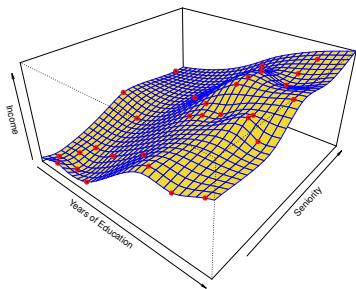
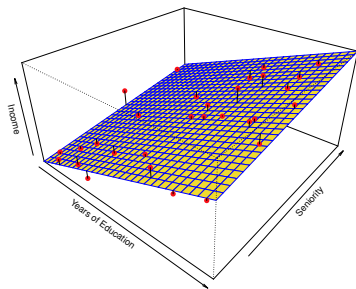
Los métodos **flexibles** tienen menos restricciones sobre la forma de  $f$ .

Ejemplos:

- Clasificación KNN, regresión KNN, splines de suavizamiento
- Bagging y boosting, máquinas de soporte vectorial
- Redes neuronales

¿Por qué preferiría un método inflexible?

Ejemplo: predicción de ingresos de “Años de educación” y “Antigüedad”



Un modelo lineal frente a un ajuste perfecto.

La elección de un método flexible o inflexible depende del objetivo en mente.

Si el objetivo es la inferencia, se preferirá un modelo inflexible, que sea fácil de entender. Por otro lado, si queremos hacer predicciones tan precisas como sea posible, no nos preocupa la forma de  $f$ . Se puede elegir un método flexible, a costa de la interpretación del modelo, y tratamos  $f$  como una caja negra.

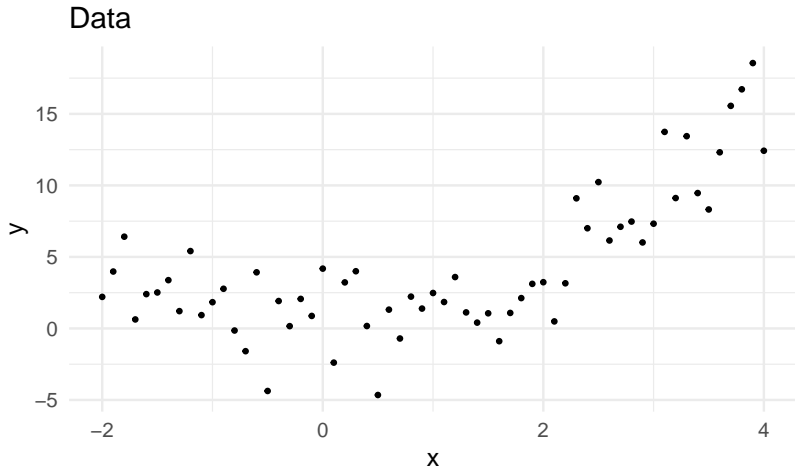
**Sobreajuste** ocurre cuando la función estimada  $f$  se ajusta demasiado a los puntos de datos observados.

**Subajuste** ocurre cuando la función estimada  $f$  es demasiado rígida para capturar la estructura subyacente de los datos.

Ilustramos esto con un ejemplo usando regresión polinomial.

## Ejemplo de Regresión Polinomial

Consideremos la covariable  $x$  observada en la cuadrícula de la recta de los reales de -2 a 4, igualmente espaciada en 0.1, dando  $n = 61$  observaciones.





Asumiremos que la relación teórica entre la respuesta  $Y$  y la covariable  $x$ :

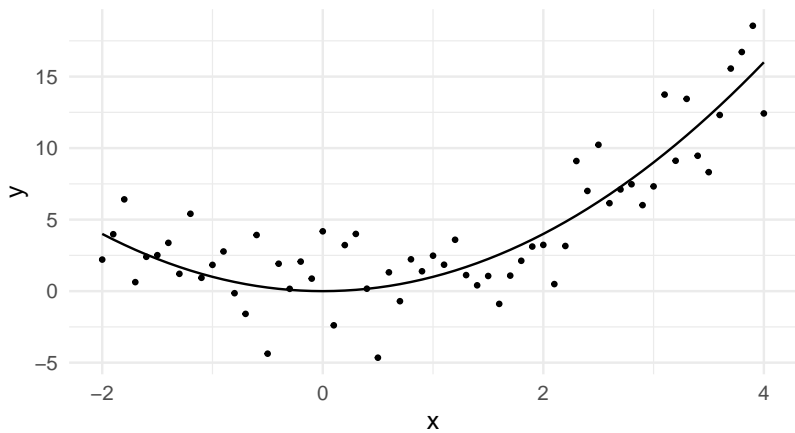
$$Y = x^2 + \varepsilon$$

Donde  $\varepsilon$  es llamado el término de error (o ruido), y es simulado de una distribución normal con media 0 y desviación estándar 2.

Llamaremos a  $Y = x^2$  la *relación verdadera*.

El error agregado se usa como un sustituto de todas las variables no observadas que no están en nuestra ecuación, pero que pueden influir en  $Y$ . Esto significa que no estamos viendo una mera relación *determinística* entre  $x$  y  $Y$ , sino una que permita la aleatoriedad.

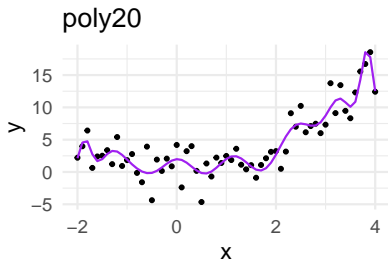
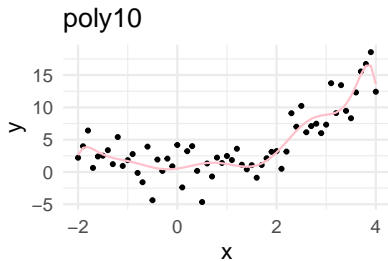
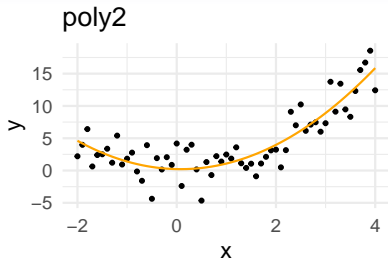
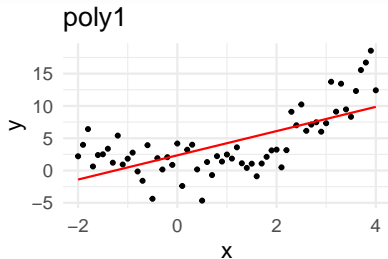
Truth



A continuación, queremos ajustar una función a las observaciones *sin conocer* la verdadera relación, y probamos diferentes funciones polinómicas paramétricas.

- [poly1: izquierda superior ]: La línea roja muestra un modelo lineal simple de la forma  $\beta_0 + \beta_1 x$  ajustada a las observaciones. Esta línea claramente *subajusta* los datos. Vemos que esta función no puede capturar esa naturaleza cuadrática de los datos.
- [poly2: derecha superior]: La línea naranja muestra un ajuste polinómico cuadrático a los datos, de la forma  $\beta_0 + \beta_1 x + \beta_2 x^2$ . Vemos que esta función se ajusta bien y se ve casi idénticamente como la verdadera función.

- [poly10: inferior izquierda]: La línea rosada muestra un polinomio de grado 10 ajustado a los datos, de la forma  $\beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_{10} x^{10}$ . La función captura el ruido en lugar de la estructura subyacente de los datos. La función *sobreajusta* los datos.
- [poly20: inferior derecha]: La línea púrpura muestra un polinomio de grado 20 ajustado a los datos, de la forma  $\beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_{20} x^{20}$ . La función captura el ruido en lugar de la estructura subyacente de los datos. La función *sobreajusta* los datos.



El grado del polinomio es un *parámetro de flexibilidad*.

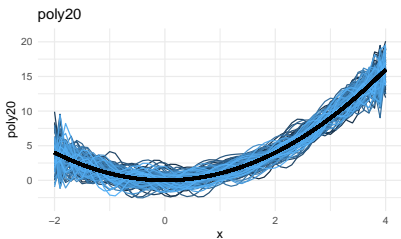
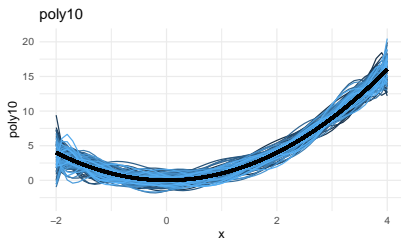
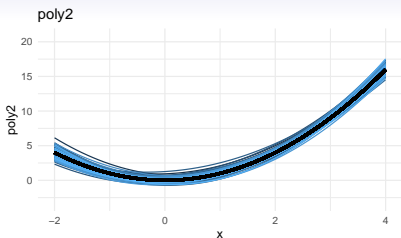
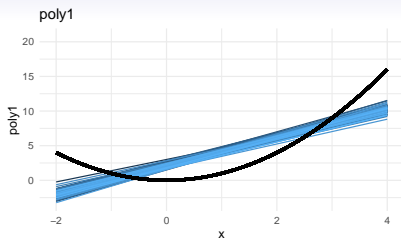
**¿Por qué comparar regresiones con diferentes grados de polinomios?** Porque en aprendizaje supervisado varios métodos incluyen un parámetro que controla la flexibilidad del ajuste del modelo, por lo que podemos hacer algunas generalizaciones con nuestro ejemplo con grados para los polinomios. El  $K$  en el vecino  $K$  más cercano es un parámetro de este tipo. Necesitamos saber cómo elegir este parámetro de flexibilidad.

Ahora podemos preguntar:

- ¿Cuál de estos modelos funciona “mejor”?
- ¿Existe *un* método que domine a todos los demás?

- Supongamos que recopilamos datos nuevos de  $Y$  (pero con nuevos errores agregados distribuidos normalmente) y estimamos nuevas curvas. ¿Cuál de las curvas de color tendría *en promedio* el mejor rendimiento?
- ¿Qué elegiste aquí para definir como “mejor rendimiento”?





Se fijó  $x$  y generaron nuevos errores 100 veces. Se muestran las 100 curvas ajustadas. La línea negra es la verdadera curva  $y = x^2$

# Evaluación de la precisión del modelo

**Ningún método** domina a todos los demás sobre todos los conjuntos de datos posibles.

- Es por eso que necesitamos aprender acerca de muchos métodos diferentes.
- Para un conjunto de datos dado, necesitamos saber cómo decidir qué método produce los *mejores* resultados.
- Necesitamos entender qué significa *mejor*.
- ¿Qué tan cerca está la respuesta predicha del valor de respuesta real?

## Función pérdida

Ahora nos centramos en la predicción.

**P:** ¿Cómo podemos medir la *pérdida* (error) entre una respuesta pronosticada  $\hat{y}_i$  y la respuesta observada  $y_i$ ?

**R:** Posibles funciones de pérdida:

- Pérdida absoluta (norma L1):  $|y_i - \hat{y}_i|$
- Pérdida cuadrática (norma L2):  $(y_i - \hat{y}_i)^2$
- Pérdida 0/1 ( $y$  categórico): Pérdida=0 si  $\hat{y}_i = y_i$  y 1 caso contrario

Problemas: robustez, estabilidad, aspecto matemático.

Usaremos la pérdida cuadrática en regresión.

## MSE de entrenamiento

En regresión, se asume que  $Y = f(x) + \varepsilon$ , y  $\hat{f}(x_i)$  da la respuesta predicha para  $x_i$ , una medida popular es el *MSE de entrenamiento* (error cuadrático medio): media de las diferencias al cuadrado entre predicción y el valor verdadero para los datos de entrenamiento (los mismos valores que se usaron para estimar  $f$ ):

$$\text{MSE}_{\text{train}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

Pero, realmente *no* estamos interesados en cómo funciona el método en los datos de entrenamiento (y, a menudo, ya hemos diseñado el método para que funcione bien en los datos de entrenamiento), queremos saber qué tan bueno es el método cuando lo usamos en *datos de prueba no vistos*, es decir, datos que podamos observar en el futuro.

¿Por qué la formación MSE no es la verdadera medida de interés?

Ejemplo:

- no queremos predecir el precio de las acciones de la semana pasada, queremos predecir el precio de las acciones la próxima semana.
- no queremos predecir si un paciente en los datos de entrenamiento tiene diabetes, queremos predecir si un nuevo paciente tiene diabetes.

```

“{r trainMSE, message=FALSE, warning=FALSE, echo=FALSE,
fig.width=6,fig.height=3,fig.align=“center”,out.width=‘80%’}
library(ggplot2) library(ggpubr) # starting with predarray # M
matrices of size length(x) times nord # first, only look at variability in
the M fits and plot M curves where we had 1
trainMSE=matrix(ncol=nord,nrow=M) for (i in 1:M)
trainMSE[i,]=apply((predarray[i,]-ymat[i,])^2,2,mean)
stackmat=NULL for (i in 1:M)
stackmat=rbind(stackmat,cbind(rep(i,nord),1:nord,trainMSE[i,]))
colnames(stackmat)=c(“rep”,“poly”,“trainMSE”)
sdf=as.data.frame(stackmat) #NB have poly1-20 now - but first only
use 1,2,20 yrange=range(sdf[,3])

p1=ggplot(data=sdf[1:nord,],aes(x=poly,y=trainMSE))+scale__y__continuous
pall= gg-
plot(data=sdf,aes(x=poly,group=rep,y=trainMSE,colour=rep))+scale__y__co
ggarrange(p1,pall) “

```

Izquierda: una repetición, derecha: 100 repeticiones para el MSE de prueba.

**P:** Según el MSE de entrenamiento, ¿qué modelo se ajusta mejor a los datos?

## MSE de prueba

Solución simple: Ajustamos (estimamos  $\hat{f}$ ) de diferentes modelos usando los datos de entrenamiento (tal vez minimizando el MSE), pero elegimos el *mejor* modelo usando un *conjunto de prueba* separado -calculando el *MSE prueba* para un conjunto de  $n_0$  observaciones de prueba  $(x_{0j}, y_{0j})$ :

$$\text{MSE}_{\text{test}} = \frac{1}{n_0} \sum_{j=1}^{n_0} (y_{0j} - \hat{f}(x_{0j}))^2$$

Notación alternativa:

$$\text{Ave}(y_0 - \hat{f}(x_0))^2$$

(tomando el promedio sobre todas las observaciones de prueba disponibles).

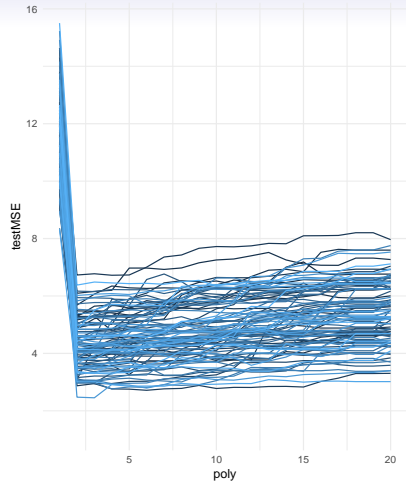
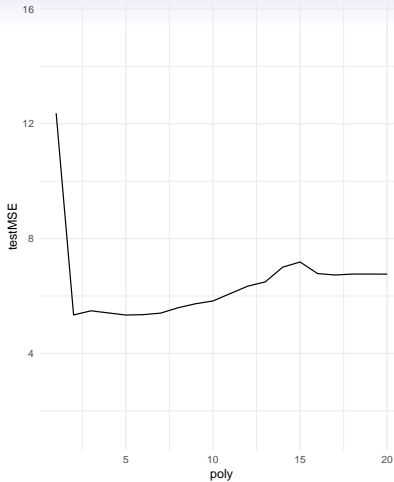
**P:** ¿Qué pasa si no tenemos acceso a los datos de prueba?

**R:** En una siguiente clase analizaremos el uso de *validación cruzada* para imitar el uso de un conjunto de prueba.

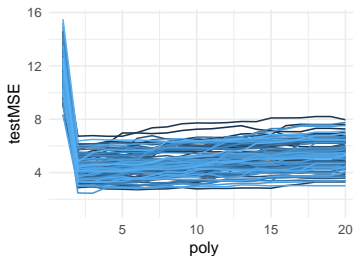
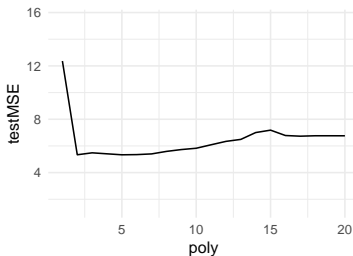
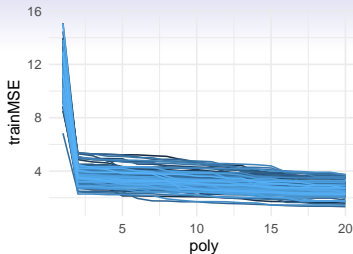
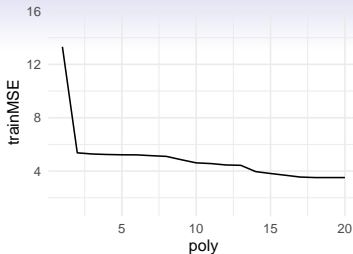
**P:** Pero, ¿podemos usar el MSE de entrenamiento para elegir el modelo? ¿Un error de entrenamiento bajo también debería dar un error de prueba bajo?

**R:** Lamentablemente, no, si utilizamos un modelo flexible, veremos varios casos en los que un error de entrenamiento bajo es un signo de sobreajuste, y daremos un error de prueba alto. Por lo tanto, el error de entrenamiento no es un buen estimador para el error de prueba porque no explica adecuadamente la complejidad del modelo.



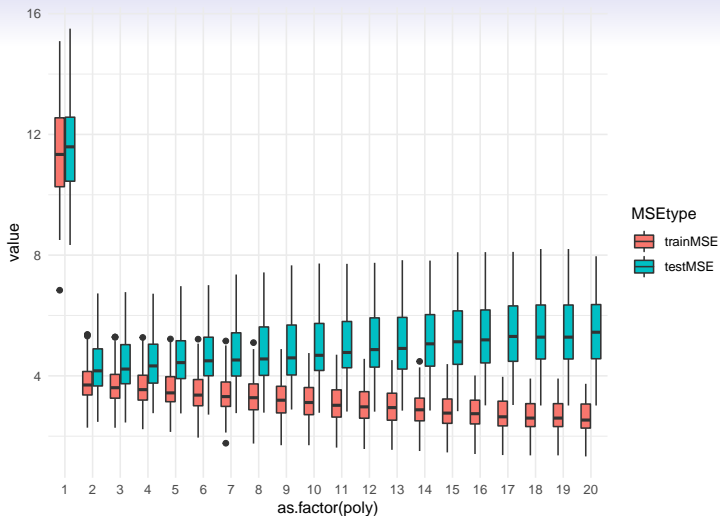


Ejemplo polinomio: Ajuste orden 1-20 cuando el verdadero es 2.  
Izquierda: una repetición, derecha: 100 repeticiones para el MSE de prueba.



**P:** Basado en el MSE prueba - ¿que modelo se ajusta mejor?

**R:** Si se elije flexibilidad basado en  $\text{trMSE}=\text{poly}20$  gana, si se elije  $\text{testMSE}=\text{poly}2$  gana.



Boxplot de las 100 repeticiones (experimento polinomial). Observar la forma de U del error de test.

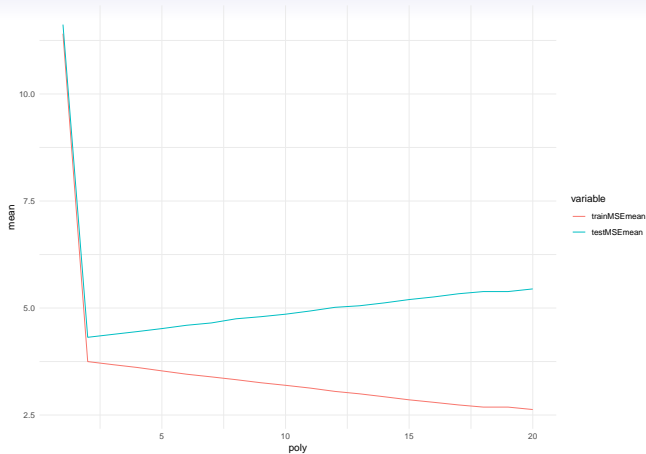
P: ¿Qué puedes leer sobre el diagrama de cajas? ¿Algo nuevo?

**R:**

Los mismos datos que antes, pero ahora se presentan conjuntamente para el MSE de entrenamiento y prueba, para centrarse en la ubicación y la variabilidad:

Boxplot:

- línea negra=mediana,
- caja del 1r al 3r cuartil,
- IQR= Rango Intercuartílico= ancho de la caja
- Bigote de min al max, excepto cuando un dato sea 1.5 veces IQR de la caja, será marcado como outlier.



Media de 100 repeticiones (ejemplo). Observe la forma de U.

**Siguiente: Dejamos el trainMSE e intentamos comprender la curva testMSE: ¡dos propiedades en competencia!**

## El balance entre el sesgo y la varianza

Supongamos que hemos estimado una curva de *regresión*  $Y = f(x) + \varepsilon$  en nuestros datos de entrenamiento, que consisten en pares de observaciones independientes  $\{x_i, y_i\}$  para  $i = 1, \dots, n$ . (Si, solo una covariable  $x$ .)

Asumimos que  $\varepsilon$  es una variable aleatoria no observada que agrega ruido a la relación entre la variable de respuesta y las covariables y se llama error aleatorio, y que los errores aleatorios tienen media cero y varianza constante  $\sigma^2$  para todos los valores de  $x$ .

Este ruido se usa como un sustituto de todas las variables no observadas que no están en nuestra ecuación, pero que influyen  $Y$ .

Supongamos que hemos usado nuestros datos de entrenamiento para producir una curva ajustada, denotada por  $\hat{f}$ .

Queremos usar  $\hat{f}$  para realizar predicciones con una nueva observación  $x_0$ , y estamos interesados en el error asociado con esa predicción. El valor predicho de la variable respuesta será  $\hat{f}(x_0)$ .

El *error cuadrático medio esperado en la prueba* (MSE) para  $x_0$  está definido como:

$$E[Y - \hat{f}(x_0)]^2$$

Observación 1: sí, podríamos haber llamado a la nueva respuesta  $Y_0$  en lugar de  $Y$ .

Observación 2: compare esto con el MSE de prueba para el ejemplo polinómico; observe que aquí tenemos *la versión teórica* donde hemos reemplazado el promedio con la media matemática (valor esperado)

Este MSE de prueba esperado puede ser descompuesto en tres términos:

$$\begin{aligned} \mathbb{E}[Y - \hat{f}(x_0)]^2 &= \mathbb{E}[Y^2 + \hat{f}(x_0)^2 - 2Y\hat{f}(x_0)] \\ &= \mathbb{E}[Y^2] + \mathbb{E}[\hat{f}(x_0)^2] - \mathbb{E}[2Y\hat{f}(x_0)] \\ &= \text{Var}[Y] + \mathbb{E}[Y]^2 + \text{Var}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)]^2 - 2\mathbb{E}[Y]\mathbb{E}[\hat{f}(x_0)] \\ &= \text{Var}[Y] + f(x_0)^2 + \text{Var}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)]^2 - 2f(x_0)\mathbb{E}[\hat{f}(x_0)] \\ &= \text{Var}[Y] + \text{Var}[\hat{f}(x_0)] + (f(x_0) - \mathbb{E}[\hat{f}(x_0)])^2 \\ &= \text{Var}(\varepsilon) + \text{Var}[\hat{f}(x_0)] + [\text{Bias}(\hat{f}(x_0))]^2. \end{aligned}$$



$$E[(Y - \hat{f}(x_0))^2] = \dots = \text{Var}(\varepsilon) + \text{Var}[\hat{f}(x_0)] + [\text{Bias}(\hat{f}(x_0))]^2$$

- Primer término: error irreducible,  $\text{Var}(\varepsilon) = \sigma^2$ , siempre está presente a menos que tengamos mediciones sin error. Este término no se puede reducir independientemente de qué tan bien nuestro modelo estadístico se ajuste a los datos.
- Segundo término: Varianza de la predicción en  $x_0$  o la desviación esperada alrededor de la media en  $x_0$ . Si la varianza es alta, existe una gran incertidumbre asociada con la predicción.
- Tercer término: sesgo al cuadrado. El sesgo da una estimación de cuánto difiere la predicción de la media real. Si el sesgo es bajo, el modelo proporciona una predicción cercana al valor verdadero.

$$E[(Y - \hat{f}(x_0))^2] = \dots = \text{Var}(\varepsilon) + \text{Var}[\hat{f}(x_0)] + [\text{Bias}(\hat{f}(x_0))]^2$$

Este es el **MSE de prueba esperado**. Podemos considerar esto como el MSE promedio de prueba que obtendríamos si estimáramos repetidamente  $f$  usando muchos conjuntos de entrenamiento (como lo hicimos en nuestro ejemplo), y luego probamos esta estimación en  $x_0$ .

Entonces, esto es realmente  $E[(Y - \hat{f}(x_0))^2 \mid X = x_0]$  si asumimos que  $X$  es una variable aleatoria.

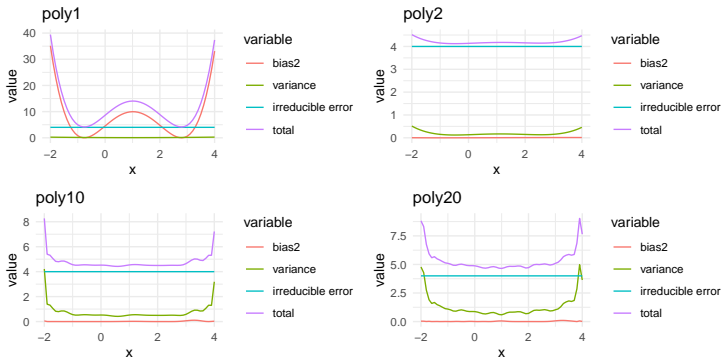
El **MSE de prueba esperado total** \* puede ser calculado promediando el MSE de prueba esperado en  $x_0$  sobre todos los valores posibles de  $x_0$  (promedio con respecto a la frecuencia en el conjunto de prueba), o matemáticamente por la ley de esperanza total  $E\{E[(Y - \hat{f}(X))^2 \mid X]\}$  (a veces también se conoce como la ley de esperanza doble).

## Regla general

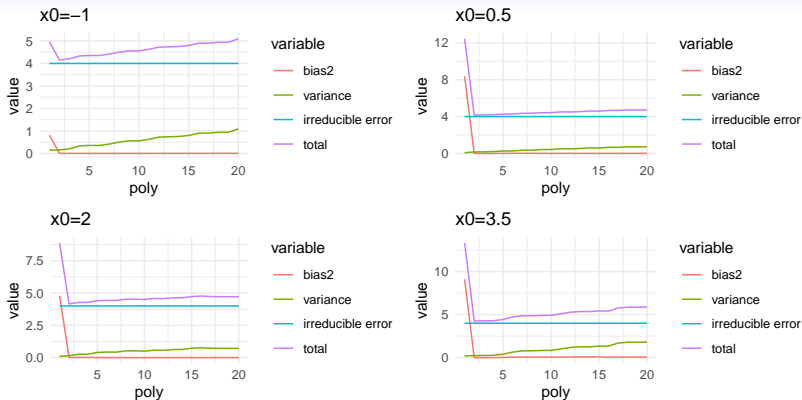
Para modelos más flexibles, la varianza se *incrementa* y el sesgo *disminuye*.

Esto es conocido como *el balance entre el sesgo y la varianza*.

## Ejemplo Modelo Polinomial (cont.)

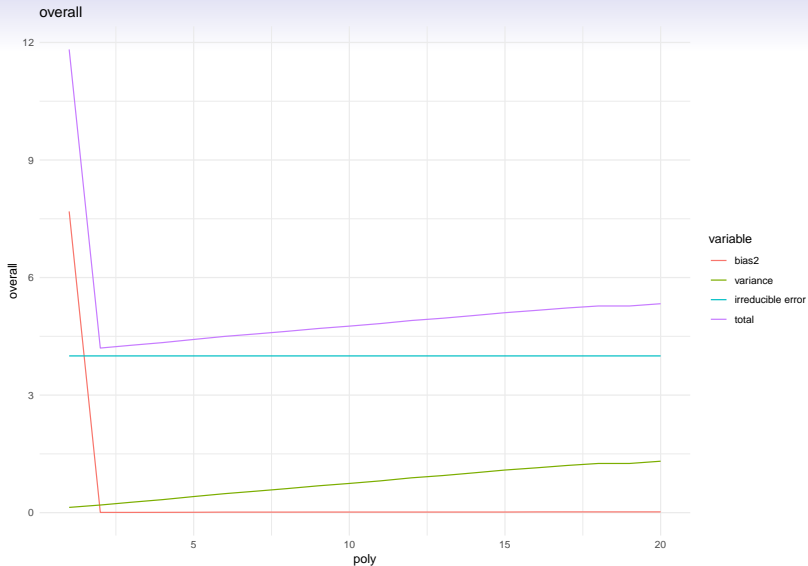


**P:** Resuma las características más importantes de las gráficas. Para 4 modelos polinomiales (poly1,2,10 y 20), el sesgo al cuadrado, la varianza, el error irreducible y la suma total. Gráficos basados en 100 simulaciones para el ejemplo polinómico.



**P:** Resuma las características más importantes de estas gráficas.

Para 4 valores diferentes de  $x_0$ , el sesgo al cuadrado, la varianza, el error irreducible y la suma total. Gráficos basados en 100 simulaciones para el ejemplo polinómico.



Versión general (con un promedio de más de 61 puntos de cuadrícula de  $x$ ).

## Elegir el mejor modelo: observaciones

Cuando se ajusta un modelo estadístico, el objetivo suele ser obtener el modelo más predictivo. A menudo hay muchos modelos candidatos, y la tarea es decidir qué modelo elegir.

- Las observaciones utilizadas para ajustar el modelo estadístico conforman el conjunto de entrenamiento. El error de entrenamiento es la pérdida promedio sobre la muestra de entrenamiento.
- A medida que aumenta la complejidad (y, por lo tanto, la flexibilidad) de un modelo, el modelo se vuelve más adaptable a las estructuras subyacentes y cae el error de entrenamiento.
- El error de prueba es el error de predicción sobre una muestra de prueba.
- La muestra de prueba tendrá nuevas observaciones que no se utilizaron al ajustar el modelo. Uno quiere que el modelo capture las relaciones importantes entre la variable de respuesta y las covariables, de lo contrario, habrá un subajuste. Recuerde la línea roja en la figura correspondiente al ejemplo de simulación anterior.

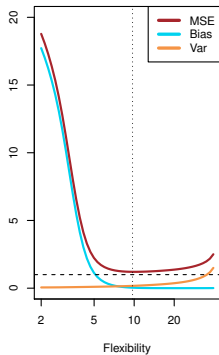
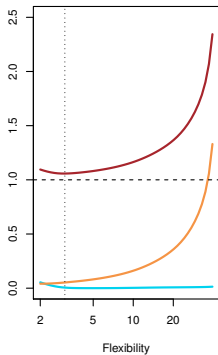
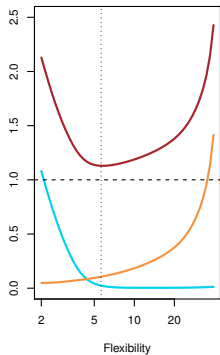
Esta compensación en la selección de un modelo con la cantidad correcta de complejidad/flexibilidad es el **balance entre el sesgo-varianza**.

Para resumir:

- los modelos inflexibles (con pocos parámetros para ajustarse) son fáciles de calcular pero pueden dar lugar a un ajuste deficiente (alto sesgo)
- los modelos flexibles (complejos) pueden proporcionar ajustes más imparciales pero pueden sobreajustar los datos (alta varianza)
- habrá errores irreducibles presentes

Elegir un estimador sesgado puede ser mejor que uno insesgado debido a las diferencias en las variaciones, y los métodos como bagging y boosting pueden reducir la variación mientras prevalece un sesgo bajo.





# Clasificación

**Contexto:** Observaciones de entrenamiento (pares independientes)  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  donde la variable respuesta  $Y$  es cualitativa. E.g  $Y \in \mathcal{C} = \{0, 1, \dots, 9\}$  o  $Y \in \mathcal{C} = \{\text{perro}, \text{gato}, \dots, \text{caballo}\}$ .

**Objetivo:** *Construir* un clasificador  $f(x)$  que asigne una etiqueta de clasificación de  $\mathcal{C}$  a una observación futura sin etiquetar  $x$  y evaluar la *incertidumbre* en esta clasificación. A veces el papel de los diferentes predictores puede ser de gran interés.

**Medidas de Rendimiento:** La más popular es la tasa de error de mala clasificación (versión de entrenamiento y prueba).

**Pérdida 0/1:** Las clasificaciones erróneas reciben la pérdida 1 y las clasificaciones correctas pérdida 0. (La pérdida cuadrática no se utiliza para la clasificación).

## Tasa de error de entrenamiento

Es la proporción de errores que se cometen si aplicamos nuestro estimador  $\hat{f}$  a las observaciones de entrenamiento, i.e.  $\hat{y}_i = \hat{f}(x_i)$ .

$$\frac{1}{n} \sum_{i=1}^n \mathbf{I}(y_i \neq \hat{y}_i).$$

Donde  $\mathbf{I}$  es la función indicadora (que da la pérdida 0/1) la cuál está definida como:

$$\mathbf{I}(a \neq \hat{a}) = \begin{cases} 1 & \text{if } a \neq \hat{a} \\ 0 & \text{else} \end{cases}$$

La función indicadora cuenta el número de veces que nuestro modelo ha hecho una clasificación incorrecta. La tasa de error de entrenamiento es la fracción de clasificaciones erróneas hechas en nuestro conjunto de entrenamiento. Una tasa de error de entrenamiento muy baja puede implicar un sobreajuste.

## Tasa de error de prueba

Aquí la fracción de clasificaciones erróneas se calcula cuando nuestro modelo se aplica en un conjunto de prueba. De lo que hemos aprendido sobre la regresión, podemos deducir que esto da una mejor indicación del verdadero rendimiento del clasificador (que el error de entrenamiento).

$$\text{Ave}(I(y_0 \neq \hat{y}_0))$$

donde el promedio se calcula sobre todas las observaciones de prueba  $(x_0, y_0)$ .

Se asume que un *buen* clasificador es aquel que tiene *menor* error de prueba.