Clase 1: 11 de junio de 2019

Fundamentos de Machine Learning.

Parte 1 de 2

EAE 253 B

C Dagnino. cdagnino@gmail.com

Find the next number of the sequence

1, 3, 5, 7, ?

Correct solution

217341

because when

$$f(x)=\frac{18111}{2}x^4-90555\,x^3+\frac{633885}{2}x^2-452773\,x+217331$$

$f(1)=1$

$f(2)=3$

$f(3)=5$

$f(4)=7$

$f(5)=217341$

much solution

very logic

wow

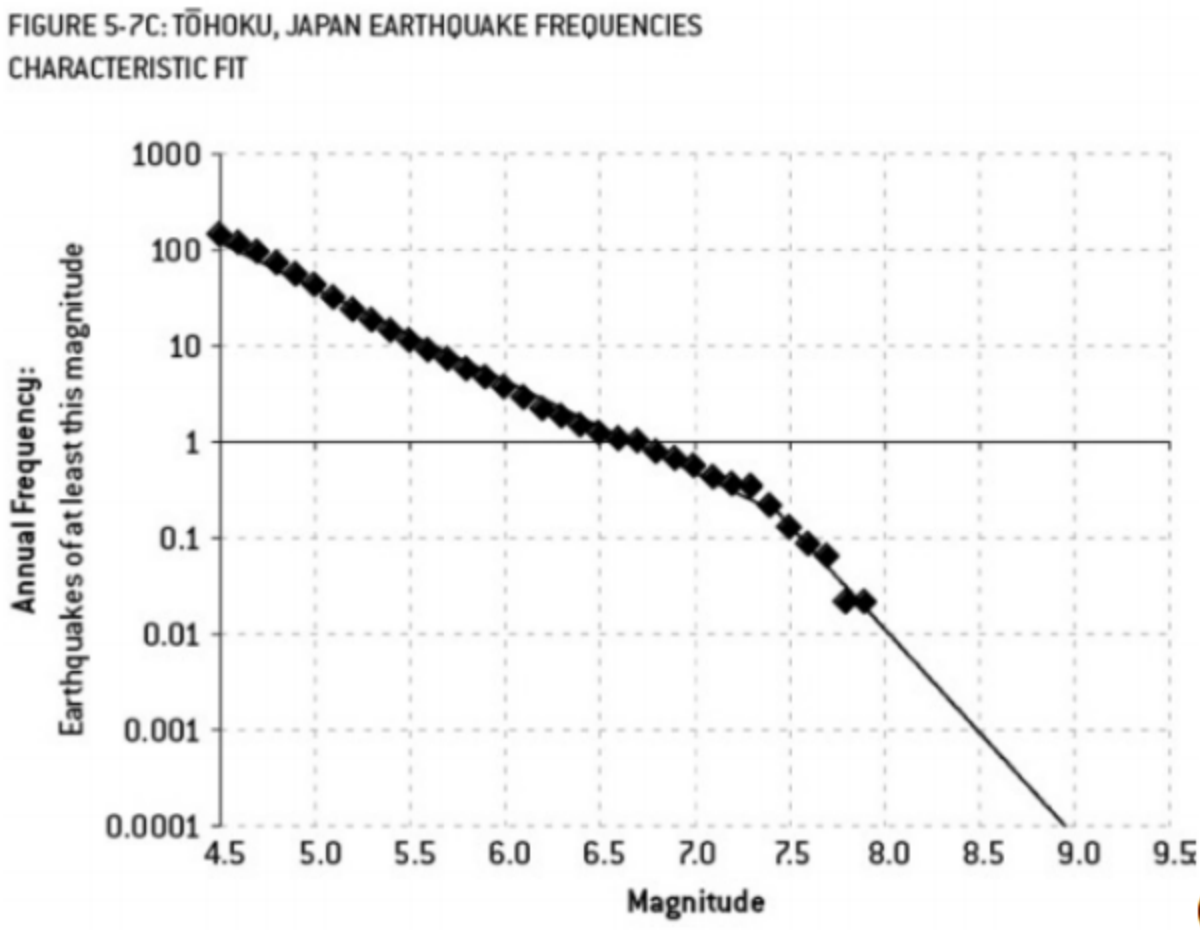
such function

many maths

wow

Fukushima: predecir frecuencia de terremotos

- Entrenaron un modelo de regresión de los últimos 400 años
- Diamantes son datos
- Línea es el modelo (y predicción)

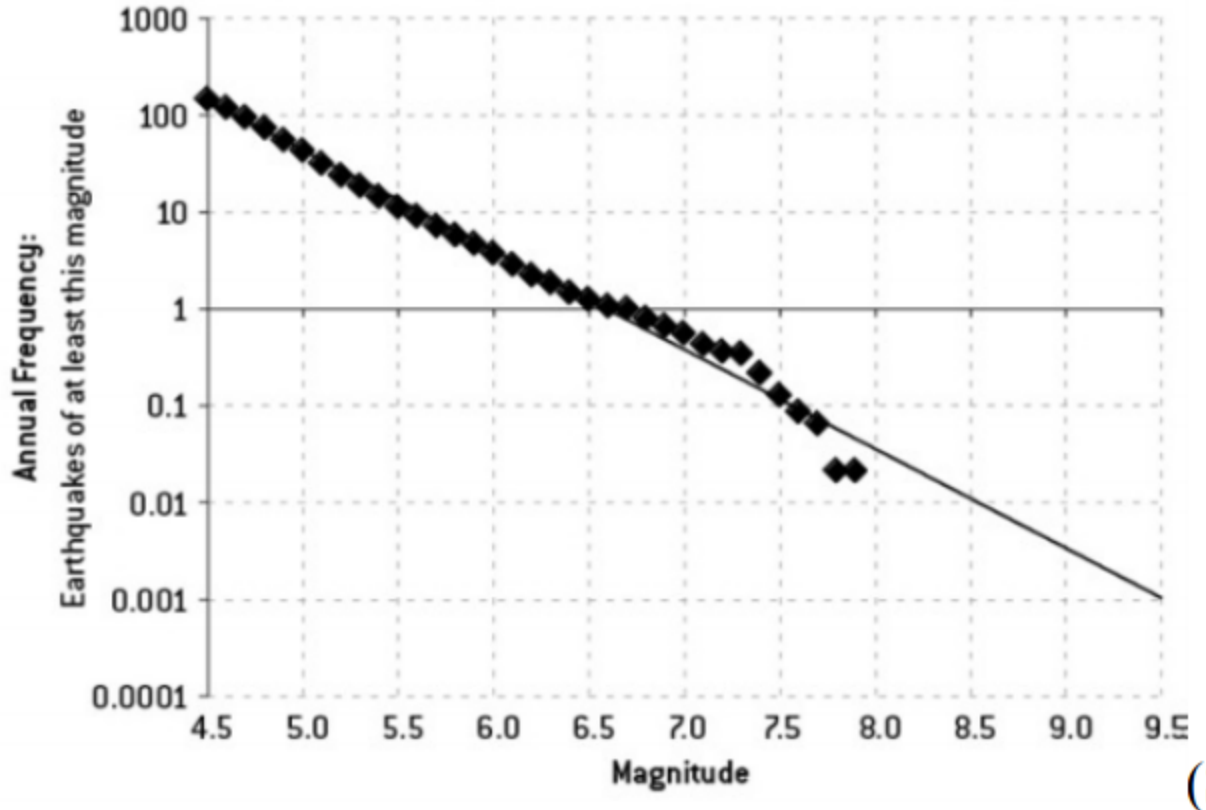


(Silver, N, 2012)

¿Qué es el sobreajuste?

Otro ajuste de los terremotos en Japón

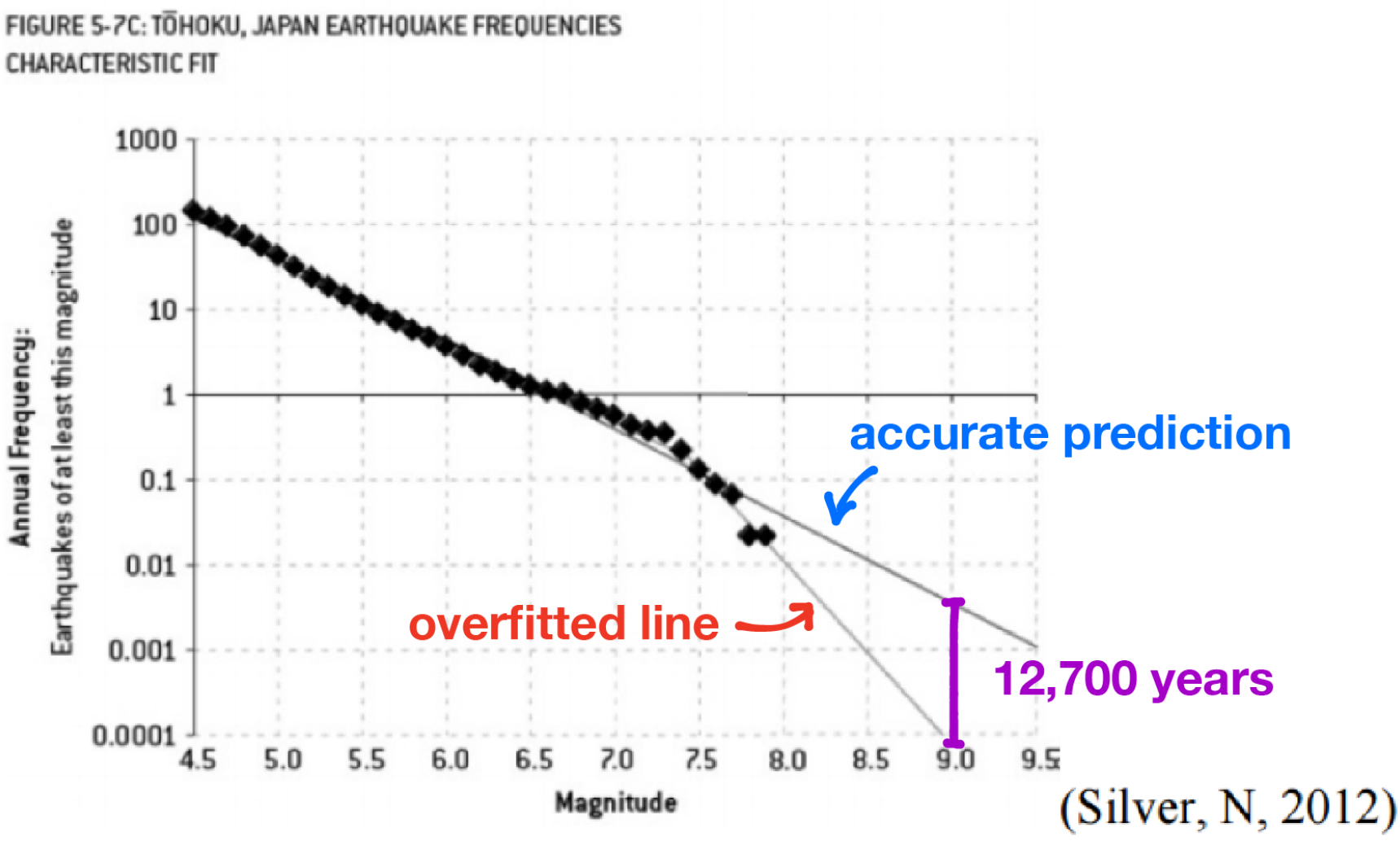
FIGURE 5-7B: TŌHOKU, JAPAN EARTHQUAKE FREQUENCIES
GUTENBERG-RICHTER FIT



(Silver, N, 2012)

El terremoto de 2011 fue de 9

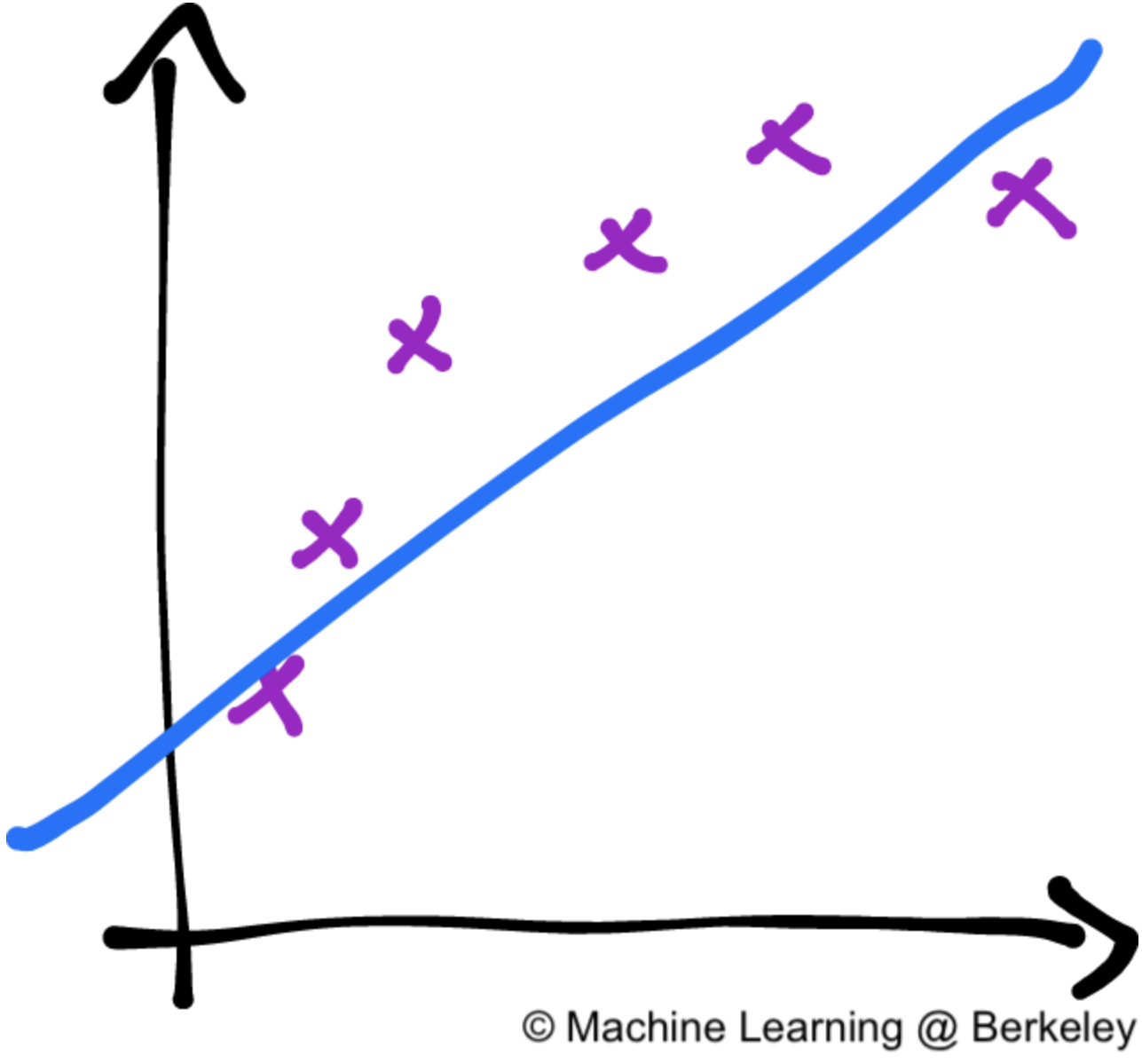
- Fukushima fue construída solamente para aguantar hasta 8.6



(Silver, N, 2012)

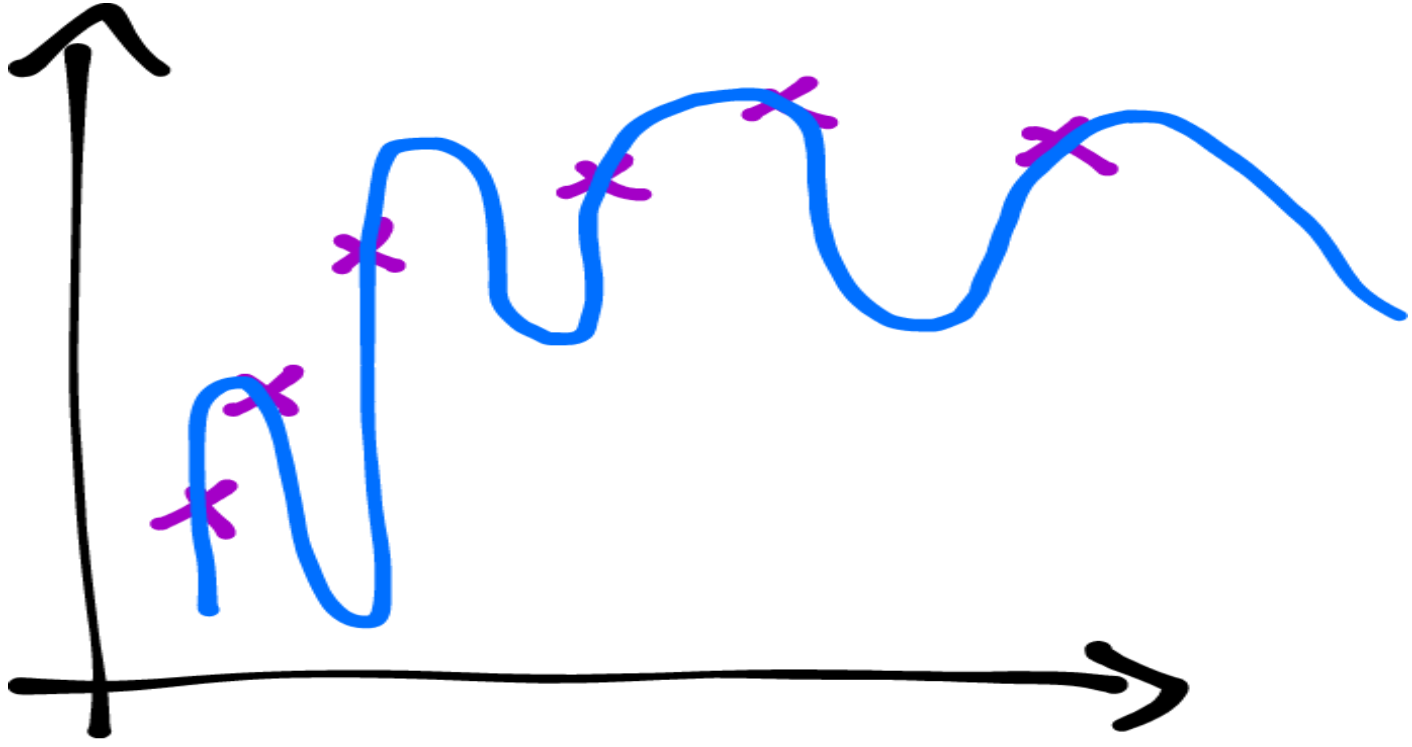
Sobre o subajuste?

Ingreso vs edad



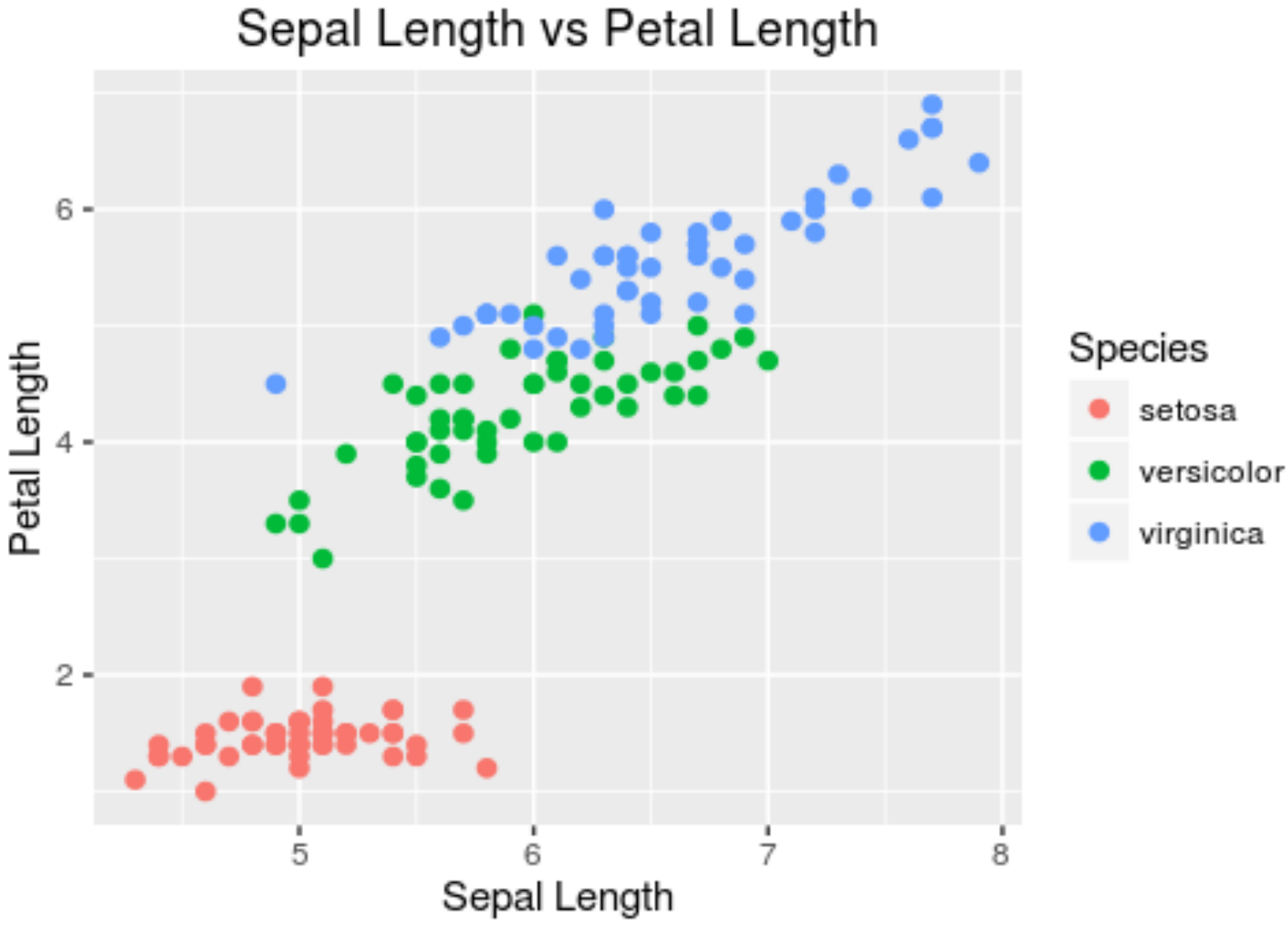
Sobre o subajuste?

Ingreso vs edad

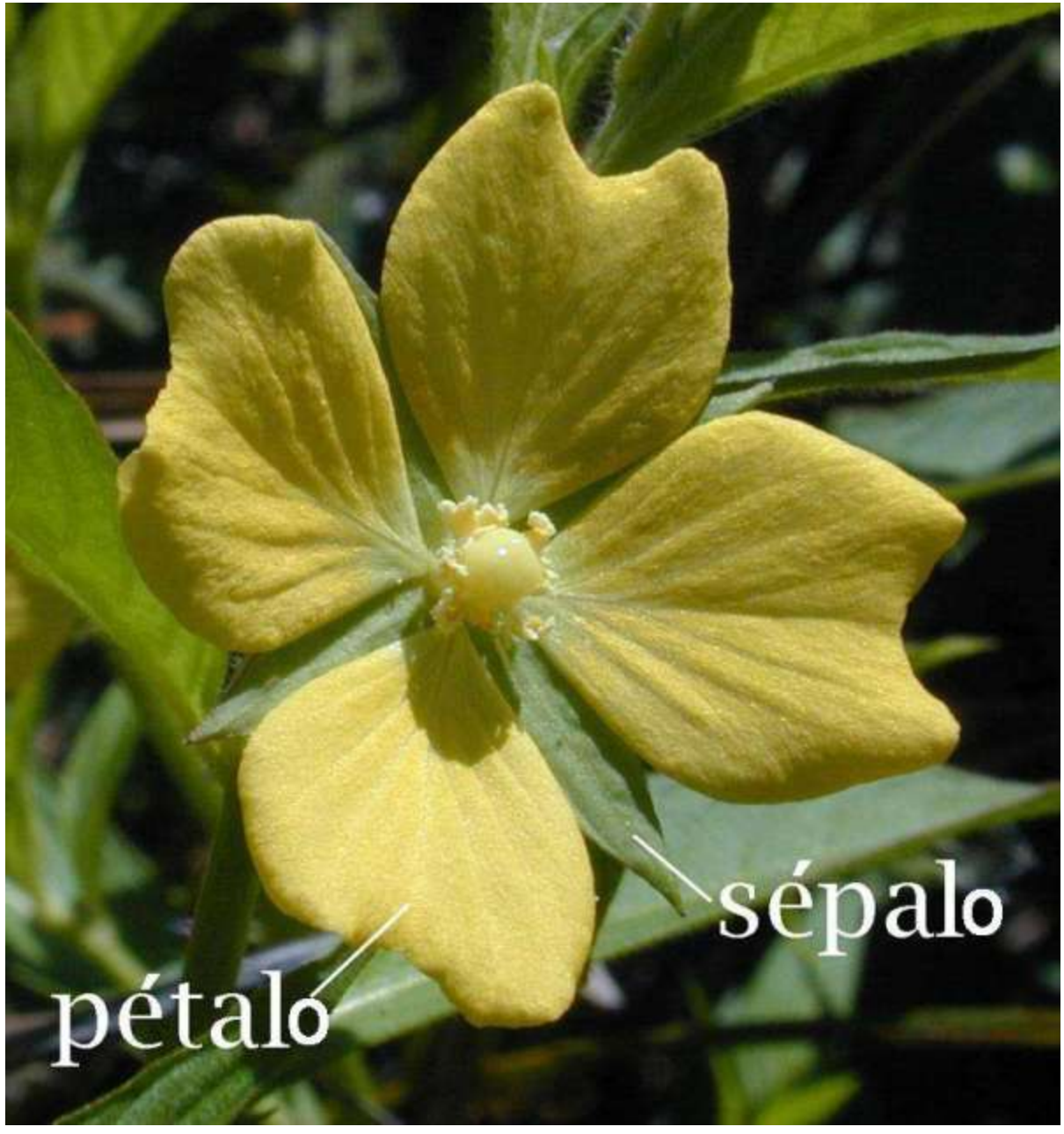


Problema de clasificación

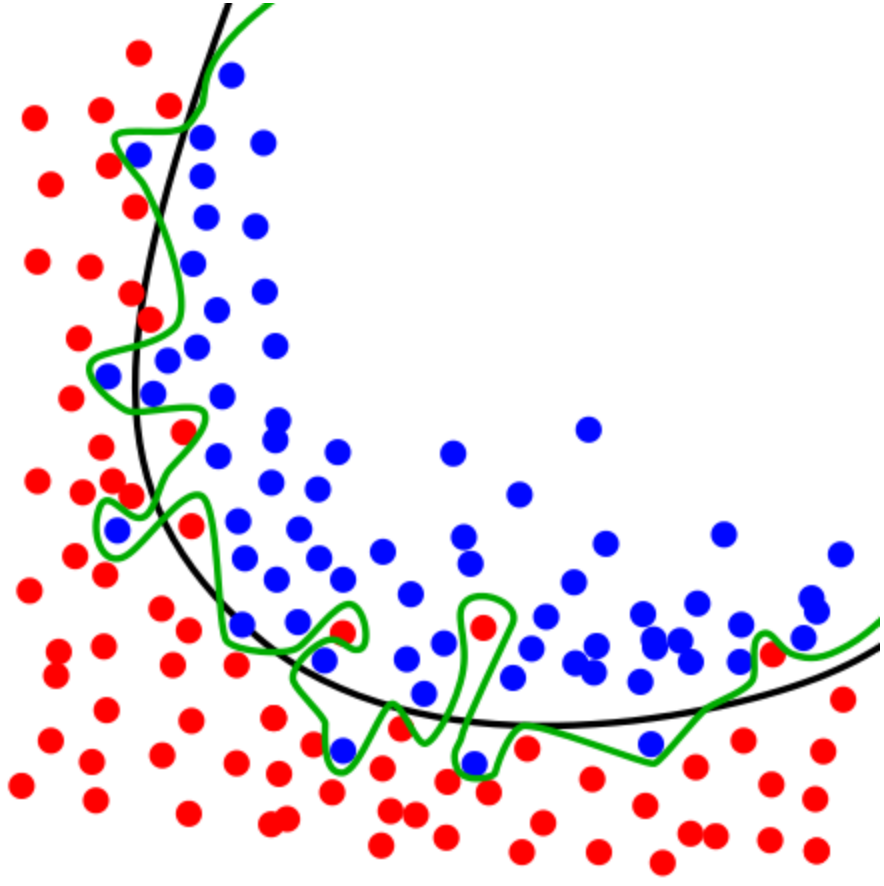
- Observamos tamaños de pétalos y sépalos



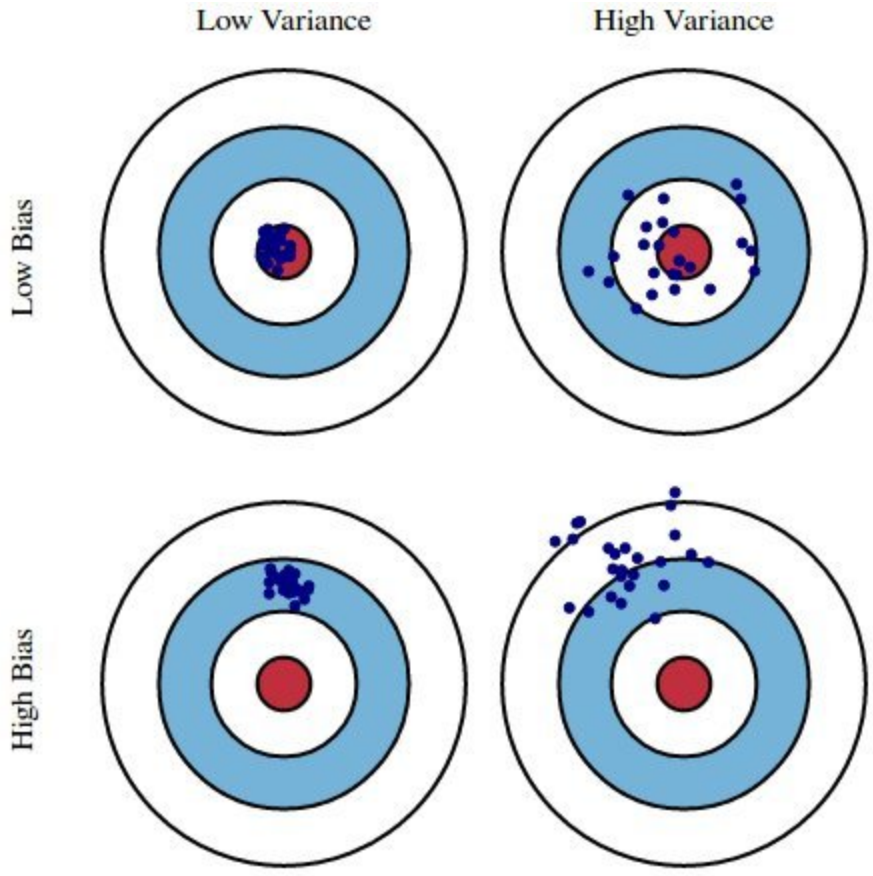
Bonus (yo tampoco sabía lo que era un sépalo)



Sobreajuste en problemas de clasificación



Sesgo y Varianza



- Sobreajuste / *overfitting*: alta varianza, alta (demasiada) complejidad
- Subajuste / *underfitting*: alto sesgo, baja (muy poca) complejidad

Dos ideas:

1. Sesgo vs Varianza
2. Complejidad: ajuste en la muestra y fuera de la muestra

Modelo (o cómo aproximarnos a este problema)

Observamos una respuesta Y y diferentes predictores $X_1, X_2, ..., X_k$. Los ponemos todos en X. Entonces, de manera bastante general:

$$Y = f(X) + \varepsilon$$

$f(X)$ es desconocido. Si asumimos que es lineal, $f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, pero aún así no conocemos los $(\beta_0, \beta_1, ..., \beta_k)$

Idea 1: Sesgo vs Varianza

Sesgo vs Varianza

Son dos maneras de fallar en nuestra predicción. Idealmente, tendríamos sesgo y varianza igual a cero.

- $Sesgo[\hat{f}(X)] = E_x[f(X) - \hat{f}(X)]$. Tomando distintas muestras de X, ¿qué tan lejos está $\hat{f}$ de f?
- $Var[\hat{f}(X)]$. ¿Qué tanto varía $\hat{f}$ cuando se aplica a distintas muestras de X?

Descomposición fundamental I: Sesgo vs Varianza

- Realidad: $Y = f(X) + \varepsilon$
- Omitimos las x: $\hat{f} = \hat{f}(X)$

$$E[(y - \hat{f})^2] = Sesgo[\hat{f}]^2 + Var[\hat{f}] + Var[\varepsilon]$$

- $Var[\varepsilon]$ se llama error irreducible
- Generalmente: disminuir el sesgo implica aumentar la varianza y vice-versa

Idea 2: Ajuste en la muestra, fuera de muestra y complejidad

Posibles Objetivos

Predicción

Obtener una predicción de Y, llamémosla $\hat{Y}$.

¿De dónde sacarla? $\rightarrow$ $\hat{f}$

$\hat{f}$ representa nuestra mejor aproximación a f

Inferencia (causal)

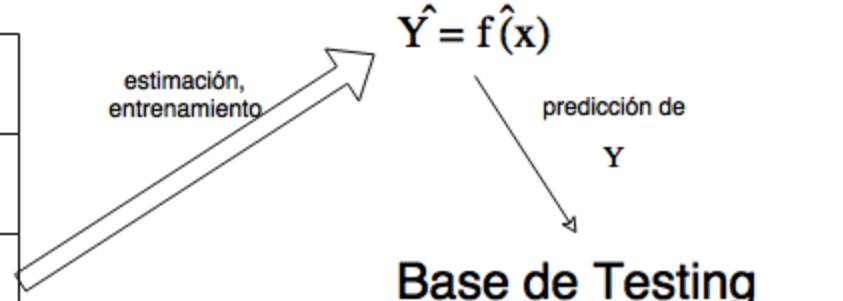
- Interés en los parámetros de $\hat{f}$
- Qué variables están asociados con Y?
- ¿Cuál es el efecto causal de X_2 en Y?

$$E[Y|do(X_2 = 1)] - E[Y|do(X_2 = 0)]$$

Proceso de predicción

Base de Training

id	edad	salud	Y
Juan	23	3	10
Jacinta	45	1	7
Andrea	29	5	9
Esteban	78	1	2
María	81	3	5



Base de Testing

id	edad	salud	Y
Lucrecio	55	2	?
Adán	22	1	?
Valeria	41	3	?

Evaluación de la predicción

Una comparación entre Y e $\hat{Y}$

La forma depende de si la variable Y es cardinal o categórica

Cardinal

- Índice de salud de 1 a 10
- Cantidad demandada de chocolitos
- Nivel de polución en Santiago mañana

Categórica:

- Mortalidad
- Tipo de flor
- Spam de e-mail

Evaluación de la predicción - Y cardinal

- Error cuadrático medio (MSE)

$$MSE(Y, \hat{Y}) = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

- Error absoluto medio (MAE)

$$MAE(Y, \hat{Y}) = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

Evaluación de la predicción - Y categórica

- Queremos evaluar si un implante de cadera funciona (F) o no (NF). Las columnas son las predicciones (P), las filas los datos reales de la *base de test*.

	P: F	**P: NF**
F	1004	21
NF	7	2122

- Suma de Falsos positivos y falsos negativos / total de observaciones
- Muchas otras métricas: AUC, *specificity* , *recall*

Resumen I

$$Y = f(X) + \varepsilon$$

- $f(X)$ puede ser una función muy compleja (no tiene por qué ser lineal)
- El objetivo es predecir Y (en el futuro).
- ε son factores desconocidos, así que lo más razonable es aproximar $f(X)$
- A la estimación la llamamos $\hat{Y} = \hat{f}$
- La evaluación se realiza mirando Y vs $\hat{Y}$

Resumen II

- Un buen modelo predictivo encuentra el compromiso preciso entre Sobreajuste (overfit) y subajuste (underfit)

Subajuste	Sobreajuste
mucho Sesgo	mucha varianza
muy poca complejidad	demasiada complejidad

Descomposición fundamental II: en-muestra vs fuera-de-muestra

- Error Real de Predicción = error en-muestra + complejidad del modelo
- Error Real de Predicción = error en-muestra + "optimismo"