RECONOCIMIENTO DE PATRONES

LABORATORIO AVANZADO DE PROCESAMIENTO DE IMÁGENES
FACULTAD DE INGENIERÍA

2020

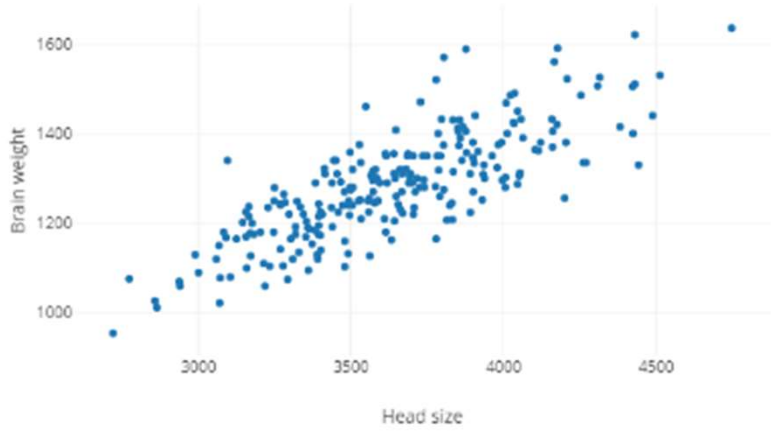
1

1

" Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? "

2

Datos



3

3

Vectores??

$f: X \rightarrow Y$

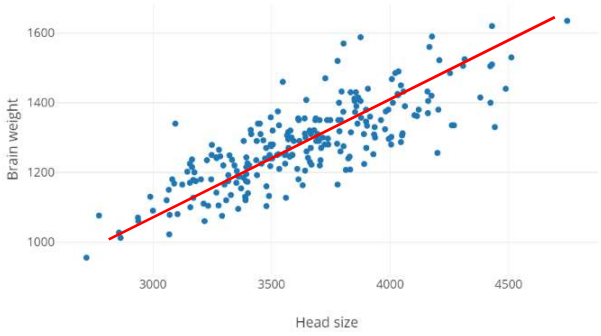
X, Y pueden ser vectores, matrices, tensor, grafo.

4

Ecuación de la recta

- $f: X \rightarrow Y$

$y=mx+b$ *reescribiendo* $f(x)=mx+b$



5

La línea como un modelo

$$y = mx + b$$

- Problema: estimar la línea a partir de los datos
- Relación entre x y y
- Dada una x identifico la y resultante

6

6

- Algo mas general

$$y=w_0x_0+w_1x_1$$

- $m=w_1$
- $b=w_0$
- $x=x_1$
- $1=x_0$

más general …

7

Regresión lineal y, sub y sobre ajuste

- La metodología

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \mathbb{E}(f_\theta(x_i), y_i)$$

8

8

Elementos de la línea

$$y = mx + b$$

- x variable independiente
- y variable dependiente
- m pendiente
- b intersección

9

9

Como modelo

$$f(x) = mx + b$$

10

10

Espacio de hipótesis

$$\mathbb{H} = \{mx + b, \forall m \in \mathbb{R}, b \in \mathbb{R}\}$$

11

11

Reformulando... (expandiendo x)

$$y = w_0 x_0 + w_1 x_1$$

- $m = w_1$
- $b = w_0$
- $x = x_1$
- $1 = x_0$

¿…pero por qué…?

12

12

Como operación de matrices

$$f(x) = W^T X = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} (x_0\ x_1) = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} (1\ \ x_1)$$

- Pero podemos hacerlo mejor

13

13

Pasado a m dimensiones

$$f(x) = W^T X = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} (x_0\ x_1) = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} (1\ \ x_1)$$

$$f(x) = W^T X = \begin{pmatrix} w_0 \\ w_1 \\ \cdot \\ . \\ w_m \end{pmatrix} (x_0\ x_1 \ldots x_m) = \begin{pmatrix} w_0 \\ w_1 \\ \cdot \\ . \\ w_m \end{pmatrix} (1\ x_1 \ldots x_m)$$

14

Redefiniendo regresión lineal

Para un conjunto $\boldsymbol{X}$ de *n* datos en $\mathbb{R}^m$, de la forma $\boldsymbol{x}_i^m$, $i = 1, \ldots, n$

para el cual agregamos un vector constante $x_0 = 1$

Y para un conjunto de y con n puntos relacionados con $\boldsymbol{X}$ en posición.

Suponemos:

$$f(x_i) = \boldsymbol{W}^T \boldsymbol{x}_i$$

¿Qué dimensiones tiene $\boldsymbol{W}$?

15

15

Evaluación: Error cuadrado

$$\boldsymbol{E}(\boldsymbol{f}(\boldsymbol{X}), \boldsymbol{y}) = \sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} (\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{x}_i)^2)$$

16

16

Más específico

$$E(f_W(X), y) = \sum_{i=0}^{n} \left(y_i - W^T x_i\right)^2$$

17

17

Mucho más específicos

$$E(f_W(X), y) = \sum_{i=0}^{n} (y_i - \sum_{j=0}^{m} w_j x_{ij})^2$$

18

18

Optimización: búsqueda

$$\underset{W}{\operatorname{argmin}} \sum_{i=0}^{n}(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij})^2$$

Descenso por gradiente

19

19

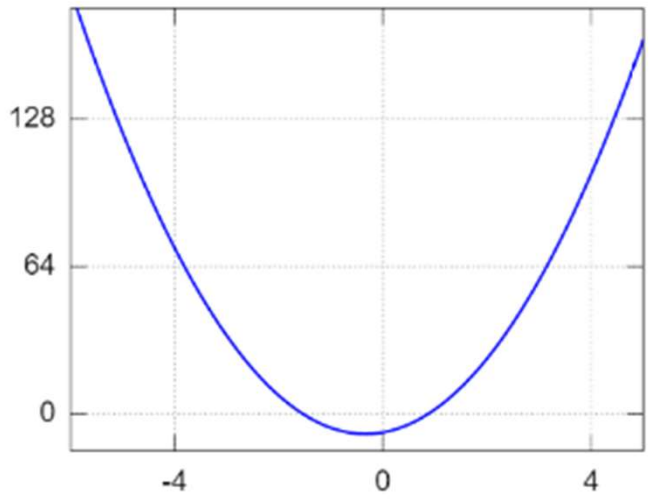
Descenso por Gradiente

- Método general de minimización
- Busca identificar un mínimo dentro de una función

20

20

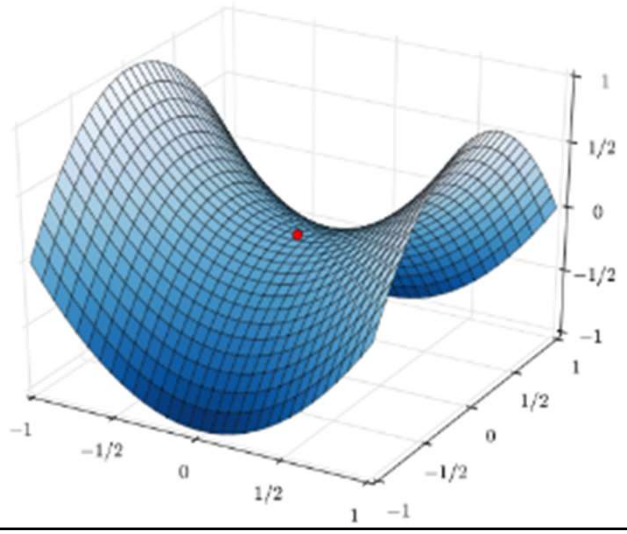
Algunas funciones son fáciles



21

21

Algunas funciones son difíciles



22

22

Nuestras opciones

- Analítica: 1ª derivada, igualar a cero, deducir valores
- Métodos numéricos: GD (1ª derivada), Newton's GD (1ª y 2ª derivada)
- Aproximativos: LBFS, PSO, AP

23

23

Imaginemos

24

24

El escenario naïve

- En algún punto de la montaña
- En cada dirección (S,N,O,P) medimos cuanto bajamos al dar un paso
- Dar un paso en la dirección que bajamos más

Avanzamos sólo un paso y en una dirección

25

25

¿Si sólo hubiera una forma de saber hacia donde decrece la montaña?

- Supongamos que estamos en algún punto de $f(\theta)$
- La pendiente en cada dimensión $\nabla = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_p}\right)$
- La pendiente indica cuanto crece la montaña
- Entonces, vayamos en dirección contraria (…)
- ¿Qué tan grande el paso? Entre más grande la pendiente un paso más grande, y viceversa
- Seamos inteligentes y pongamos un parámetro $\boldsymbol{k}$, *i.e*

Al final se tiene: $-\boldsymbol{k}\nabla$

26

26

Formulación

- Dado un punto, podemos encontrar un nuevo punto más bajo

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{k}\nabla f(\theta)$$

Método iterativo

27

27

Para identificar nuevos pesos que minimicen el error

$$\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{k}\nabla \mathbb{E}(f_w(x), y)$$

- Método iterativo

28

28

p1

Algoritmo

```
while abs(x_new - x_old) > thres:
    x_old  = x_new
    x_new = x_old - k * f_derivative(x_old)
```

Se ve fácil; la parte difícil es calcular ∇f

29

29

Regresando

- La derivada parcial me la da la cantidad por la dimensión
- $\left(\hat{\theta}_1, \ldots, \hat{\theta}_p\right) = \left(\theta_1, \ldots, \theta_p\right) - k\left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_p}\right)$

30

30

Slide 29

p1 personal, 1/28/2020

Regresando a nuestro problema

Buscando minimizar

$$\mathbb{E}(f_w(x), y)$$

$$(\widehat{w}_1, \dots, \widehat{w}_p) = (w_1, \dots, w_p) - k\left(\frac{\partial}{\partial w_1}, \dots, \frac{\partial}{\partial w_p}\right)$$

31

31

Recordemos

$$E(f_W(\boldsymbol{X}), y) = \sum_{i=0}^{n}(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij})^2$$

32

32

Derivando para dimensión d

$$\frac{dE(f_W(\boldsymbol{X}), y)}{dw_d} = \frac{d \quad \sum_{i=0}^{n}(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij})^2}{dw_d}$$

33

33

Derivadas de las sumas

$$\sum_{i=0}^{n} \frac{d(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij})^2}{dw_d}$$

34

34

Derivada de potencias

$$\sum_{i=0}^{n} 2(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij}) \frac{d(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij})}{dw_d}$$

35

35

Expandiendo sumatoria

$$\sum_{i=0}^{n} 2(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij}) \frac{d(y_i - w_0 x_{0i} \ldots - w_d x_{di} \ldots - w_j x_{ji})}{dw_d}$$

36

36

Expandiendo sumatoria

$$\sum_{i=0}^{n} 2(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij}) \frac{d(-w_d x_{di})}{dw_d}$$

37

37

Derivando elementos

$$\sum_{i=0}^{n} 2(y_i - \sum_{j=0}^{m} w_j \boldsymbol{x}_{ij})(-x_{di})$$

38

38

Simplificado

$$\frac{dE(f_W(\boldsymbol{X}), y)}{dw_d} = \sum_{i=0}^{n}(y_i - f(\boldsymbol{x_i}))(-x_{di})$$

39

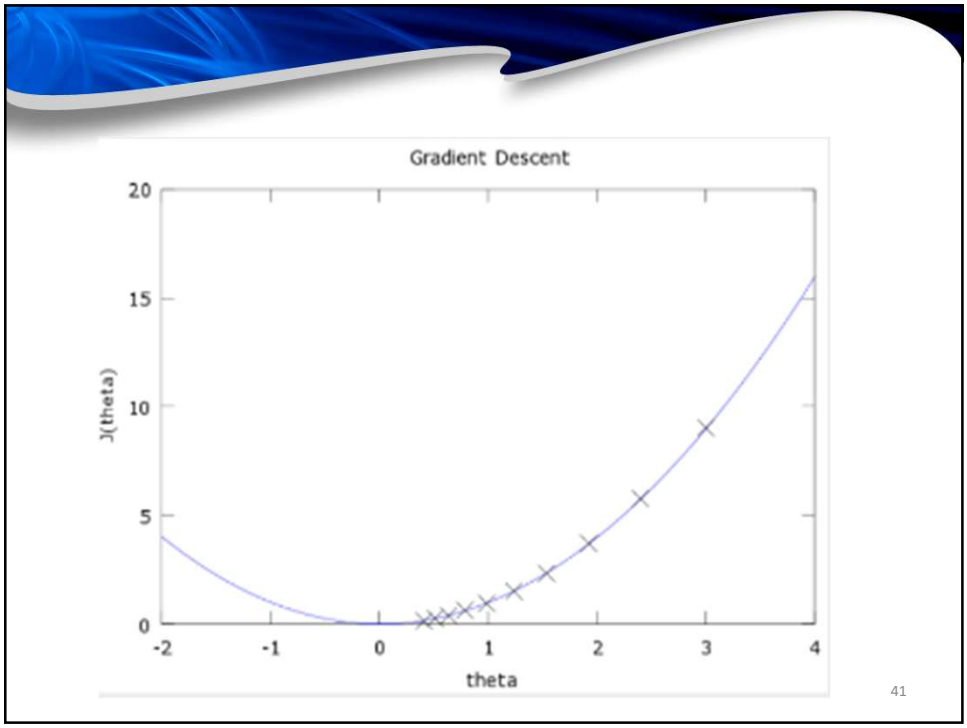
39

Resultado

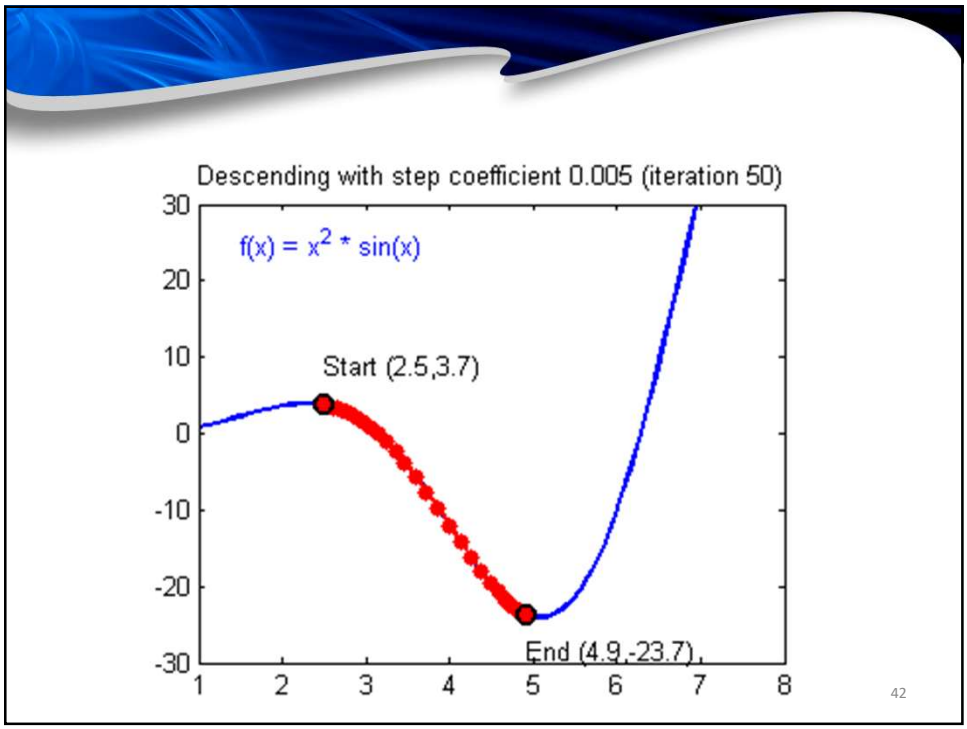
$$\hat{w}_d = w_d - k\sum_{i=0}^{n}(y_i - f(\boldsymbol{x_i}))(-x_{di})$$

40

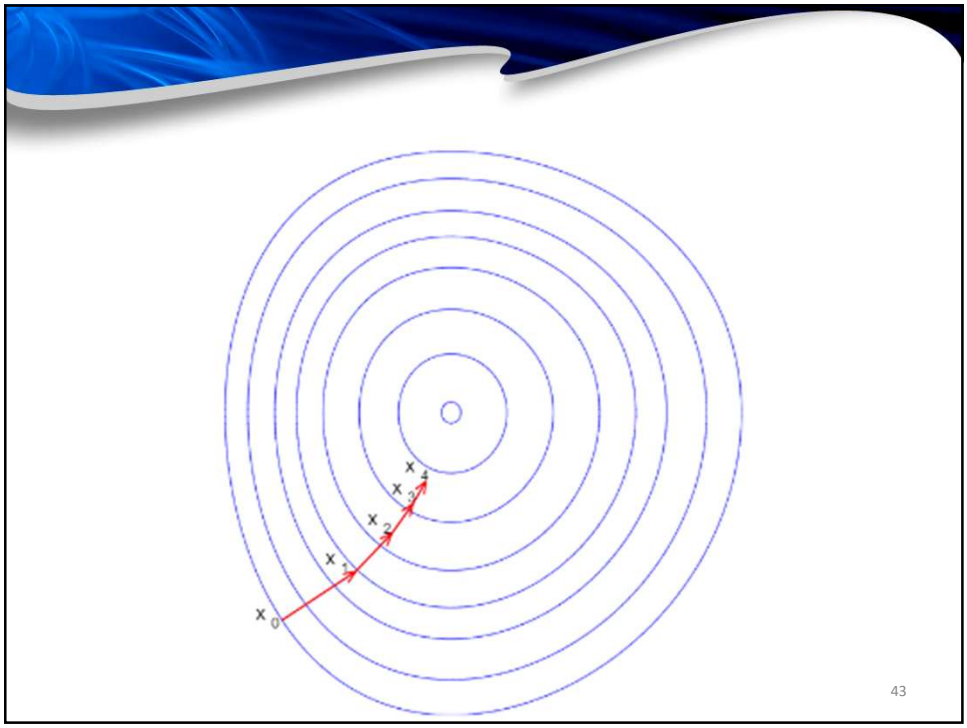
40


Gradient Descent
J(theta)
theta
20
15
10
5
0
-2
-1
0
1
2
3
4
41

41

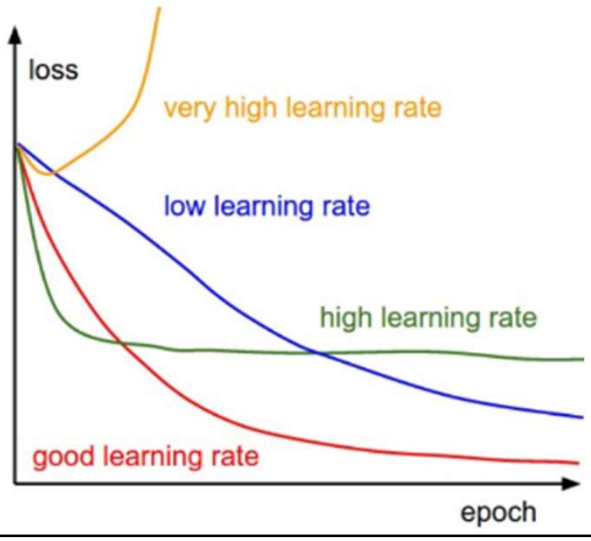

Descending with step coefficient 0.005 (iteration 50)
f(x) = x^2 * sin(x)
Start (2.5,3.7)
End (4.9,-23.7)
30
20
10
0
-10
-20
-30
1
2
3
4
5
6
7
8
42

42



43

¿Qué hay de k?



44

¿Pero con qué pesos comenzamos?

- Aleatorio
- Uniforme al número de pesos

Como es aproximativo, no siempre tenemos el mismo modelo

45

45

¿Qué tipo de errores tenemos?

- Podemos pensar que nuestro espacio de hipótesis $\mathbb{H}$ es un espacio de expertos
- Cada experto se desarrolló con una experiencia propia del problema, entonces difieren en como resolver el problema

46

46

¿Qué tipo de errores cometen?

- Se equivocan siempre de igual forma
- Se equivocan por todos lados

47

47

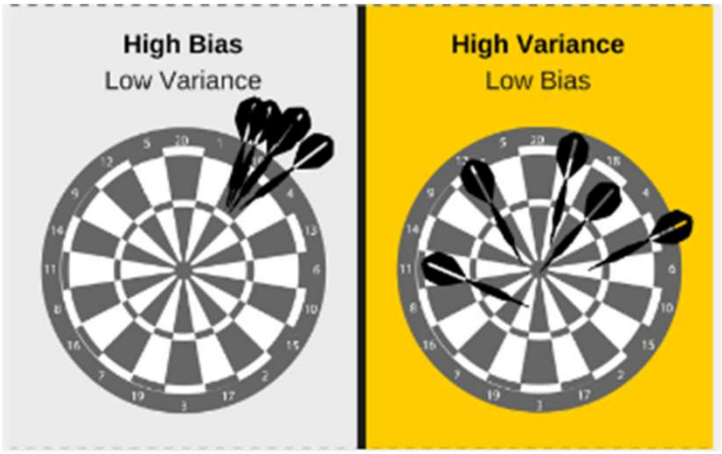
Es decir

- El experto tiene sesgo (Bias)
- El experto tiene varianza (Var)

48

48

Los errores son una combinación de ambos factores



Tomado de [WTF is the Bias-Variance Tradeoff?]

49

Falta un tipo de error

Intrínsecos del problema (ϵ)

50

50

¿Quién tiene la culpa del error?

¿El algoritmo o la hipótesis?

51

51

Valores esperados

Para variable discreta

$$\mathbb{E}[X] = \sum_{i=0}^{n} x_i p(x_i)$$

Para una distribución uniforme

$$\mathbb{E}[X] = \frac{1}{n} \sum_{i=0}^{n} x_i$$

52

52

Predicción esperada

$$\mathbb{E}[f_\theta(x)] = \frac{1}{n}\sum_{i=0}^{n} f_\theta(x_i)$$

53

53

Error cuadrático esperado

$$\mathbb{E}[(f_\theta(x) - y)^2] = \frac{1}{n}\sum_{i=0}^{n} (f_\theta(x) - y)^2$$

54

54

Sesgo

- Es la diferencia entre la predicción esperada y el valor verdadero

$$Bias[f_\theta(x)] = y - \mathbb{E}[f_\theta(\boldsymbol{X})]$$

- ¿Por qué son mala noticias para nosotros?

55

55

Varianza

- Es la diferencia entre la predicción al cuadrado esperada y el cuadrado de la predicción esperada

$$Var[f_\theta(x)] = \mathbb{E}[f_\theta(\boldsymbol{X})^2] - \mathbb{E}[f_\theta(\boldsymbol{X})]^2$$

56

56

Error cuadrático esperado

$$\mathbb{E}[(f_\theta(x) - y)^2] = (y - \mathbb{E}[(f_\theta(X)])^2 + \mathbb{E}[(f_\theta(X)^2] - \mathbb{E}[(f_\theta(X)]^2 + \epsilon$$

57

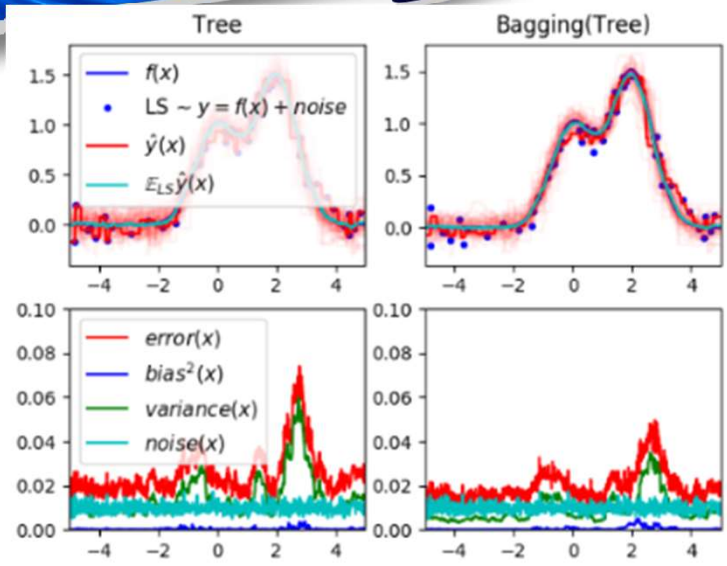
57

O en otras palabras

$$\mathbb{E}[(f_\theta(x) - y)^2] = Var[f_\theta(x)] + Bias[f_\theta(x)]^2 + \epsilon$$

58

58



Tomado de Single estimator versus bagging: bias-variance decomposition

59

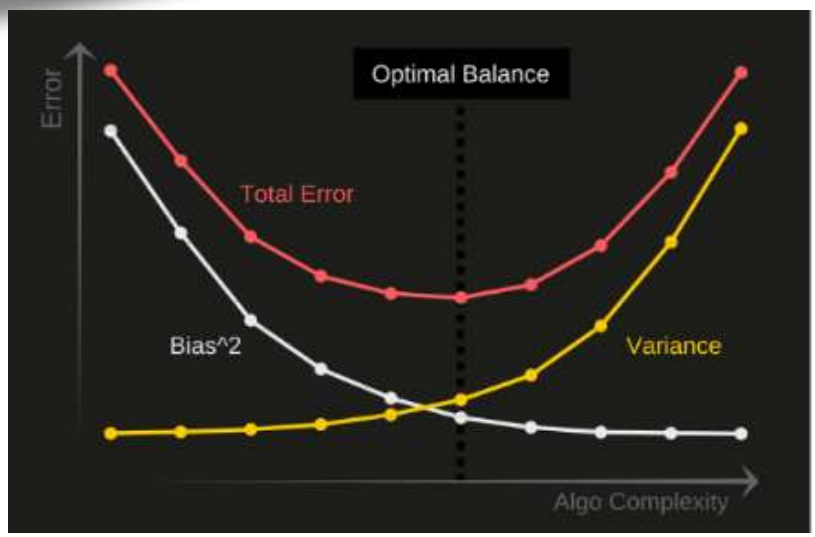
59

Compromiso entre sesgo y varianza

- Modelos más sencillos, producen alto sesgo
- Modelos más complejos, producen bajo sesgo
- Modelos más sencillos, producen una baja varianza en un espacio $\mathbb{H}$ más grande
- Modelos más complejos, producen una mayor varianza en un espacio $\mathbb{H}$ más grande

60

60



Tomado de WTF is the Bias-Variance Tradeoff

61

61

Sub ajuste

- Es cuando el modelo es tan simple que no es suficiente para modelar las relaciones de los datos de entrenamiento
- Imaginen el modelo $y = c$

62

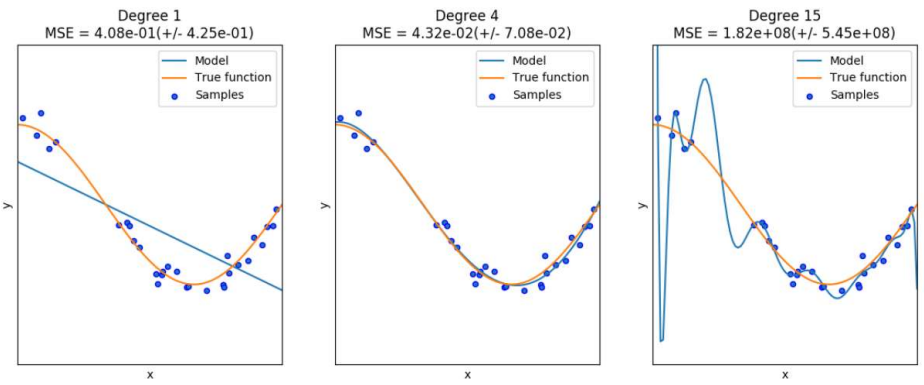
62

Sobre ajuste

- Es cuando el modelo es muy complejo, de tal forma que memoriza los datos de entrenamiento, lo único que modela es el ruido
- Imaginen el modelo $f = g + \epsilon$
- Ante nuevos datos falla más

63

63



Tomado de Underfitting vs. Overfitting

64

64