Machine learning algorithms

Overfitting & underfitting

2020–09-29

CSCI 471 / 571, Fall 2020

Kameron Decker Harris

What is ML?

d P polynomial Kernel

- Data + Optimization + Statistics → Predictions

$X, \vec{y}$

least squares normal eqns
- linear system
- solution $\vec{\beta}^*$

overfitting/under effects of noise

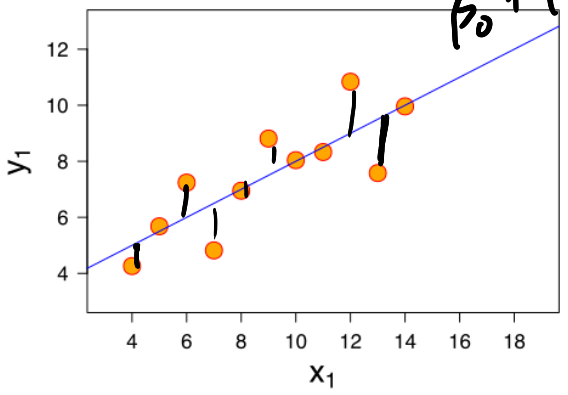
fitting

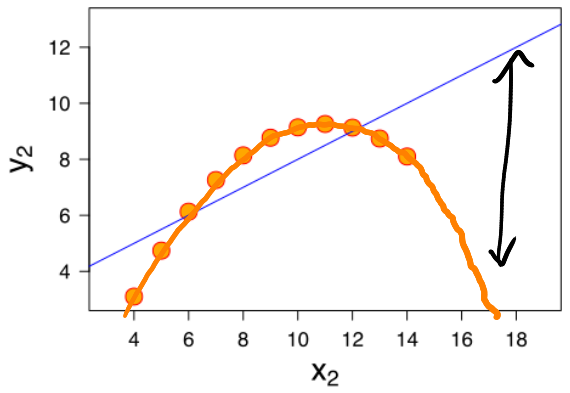
☆ Ref 2. Chapter 3, James, Witten et al.

Linear models are simple

- fits are very close

linear looks good →



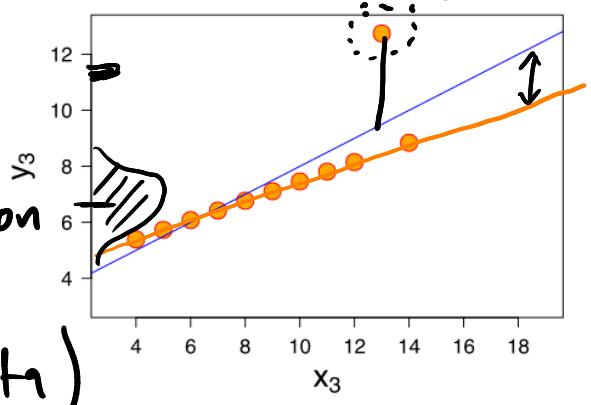


☆ some hope for squared err.

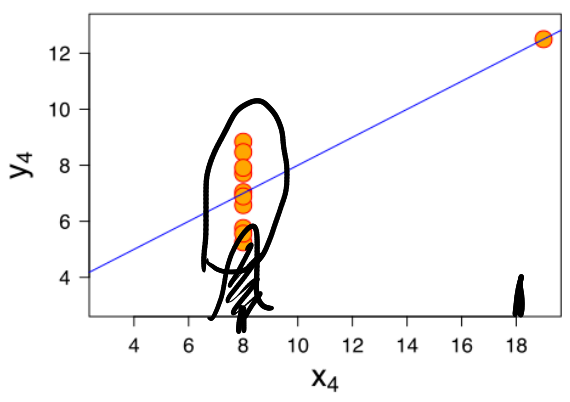
looks like polynomial

- same sq. err.
- same means $\bar{x}$ $\bar{y}$
- same std's

outlier



poor generalization (predictions outside training data)



high leverage atypical x

Anscombe, 1973

Least-squares in nonlinear basis

$\underline{X} \longrightarrow$ add polynomial features

$$\min_{\vec{\beta}} \; \| \underline{\Phi}(\underline{X})\,\vec{\beta} - \vec{y} \|^2$$

$\underline{\Phi}$: $n \times p$

e.g. $\vec{\phi}(\vec{x})$, $d=2$ $= \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^3 \\ \cos x_2 \end{bmatrix}$

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^3 + \beta_4 \cos x_2$$

Beware extrapolation

STAT Topics Coronavirus Opinion Podcast Newsletters Reports Events

HEALTH

Influential Covid-19 model uses flawed methods and shouldn't guide U.S. policies, critics say

By SHARON BEGLEY @sxbegle / APRIL 17, 2020 Reprints