Module 2: Statistical learning

TMA4268
15.01.2018

Topics for statistical learning:

i) aim: inference vs prediction

understand
interpret

ii) type of problem: regression vs classification

regression: quantitative output

- bus arrival: predict lateness
- power grid: predict consumption

classification: qualitative output

- prison: classify inmate as violate or not violate
- customer: good or bad (paying debt)

iii) type of set-up (data/aim/method)

supervised vs unsupervised learning

response available as in regression and classification

want to detect unknown patterns in data

iv) type of method

parametric vs non-parametric

- simple to use — flexible
- but constrained — can overfit
- computationally cheap ← need more data
← comp. more expensive

v) over- and underfitting

Ex: truth 2nd poly (black - and fitted = orange)
↓ 1st poly (red)
see figure below 20 poly (purple)

a) best? orange - purple - red ← students
2 20 1

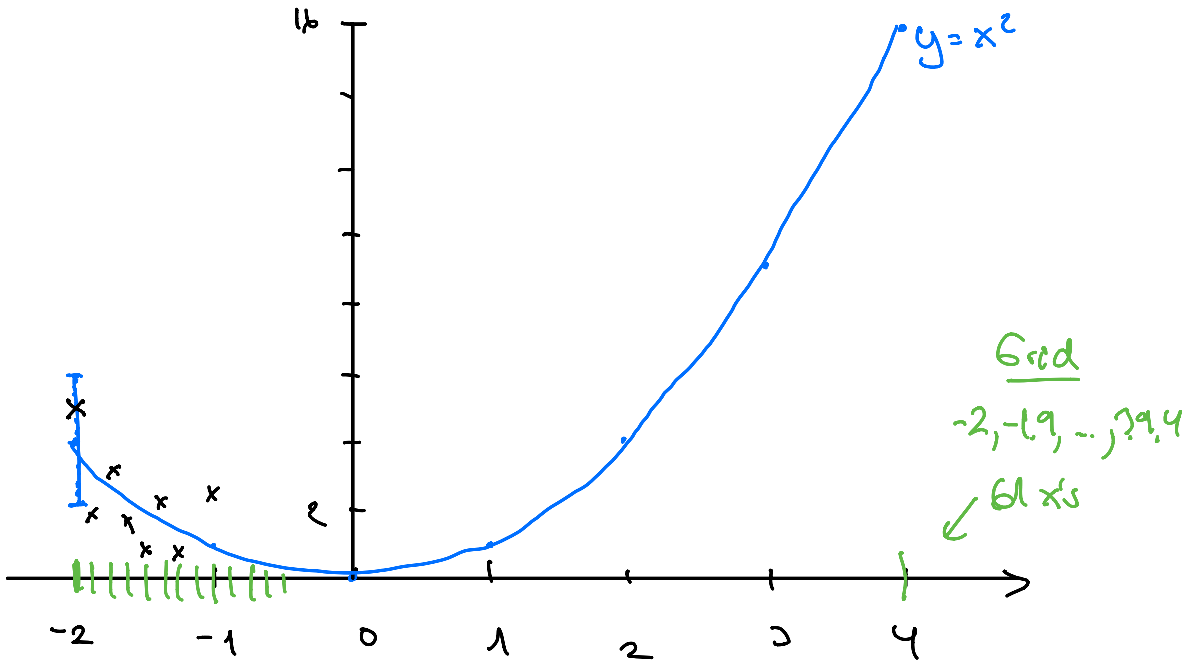
b) red: not hitting the "correct" curve, but less variable than 20
1

purple: on average hits correct, but rather variable
20

Training MSE vs Test MSE
if we choose the model with minimum on Training MSE this may lead to high test MSE if we have used a too flexible method.

Polynomial example:

$$Y = \underbrace{x^2}_{f(x)} + \varepsilon \qquad \varepsilon \sim N(0, 1^2)$$



our choice of parametric model

Fit: poly 1: $f(x) = \beta_0 + \beta_1 X$

2 $\quad f(x) = \beta_0 + \beta_1 X + \beta_2 x^2$

$\vdots$

20 $\quad f(x) = \beta_0 + \quad \cdots \quad + \beta_{20} x^{20}$

First: generate one data set
- → fit model
- → make prediction at each x (61)
- → plot predictions (line) ← compare with observed obs.
- → generate one test set → calc predictions & compare observation

Bias-variance trade-off

$$Y = f(x) + \varepsilon \quad \text{where} \quad E(\varepsilon) = 0$$

Training data (x_i, Y_i), independent pairs, used to fit model and give $\hat{f}$.

given model

Now, want to predict new observation at x_0. Use $\hat{f}(x_0)$ as estimator

function of (x_i, Y_i)

RV's

Quadratic loss at new observation Y at x_0

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

We want the "long term average" = the expected value ← over Y

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

$$= E(Y^2) - 2\,\underset{\text{new } Y}{E(Y)} \cdot \underset{\text{based on } (Y_1, \ldots, Y_n)}{E(\hat{f}(x_0))} + E(\hat{f}(x_0)^2)$$

$$\left\{\begin{array}{l} \text{Remember: } Var(Y) = E(Y^2) - E(Y)^2 \\ \qquad\qquad \Leftrightarrow E(Y^2) = Var(Y) + E(Y)^2 \\ \text{Same for } \hat{f}(x_0)\text{: } \quad E(\hat{f}(x_0)^2) = Var(\hat{f}(x_0)) + E(\hat{f}(x_0))^2 \end{array}\right.$$

$$= Var(Y) + E(Y)^2 + Var(\hat{f}(x_0)) + E(\hat{f}(x_0))^2$$

$$- 2E(Y) \cdot E(\hat{f}(x_0))$$

$Y = f(x_0) + \varepsilon$

$\parallel$

not random

$f(x_0)$

since

$E(\varepsilon) =$

$$= Var(\varepsilon) + f(x_0)^2 + Var(\hat{f}(x_0)) + E(\hat{f}(x_0))^2 - 2f(x_0) \cdot E(\hat{f}(x_0))^2$$

$$= \underset{\text{irreducible error}}{Var(\varepsilon)} + \underset{\text{venence of predicti}}{Var(\hat{f}(x_0))} + \underbrace{(\underset{\substack{\text{expected}\\ \text{value of}\\ \hat{f}(x_0)}}{E(\hat{f}(x_0))} - \underset{\text{true value at } x_0}{f(x_0)})^2}_{\text{squared bias of predicti}}$$

This is at one x_0, we migth aveerage over all x_0's.