UPPSALA UNIVERSITET

Introduction to seminar course

The unreasonable effectiveness of overparameterized machine learning models (3 hp)

Dave Zachariah

Machine Learning Arena IT-CIM & Div. Systems and Control

dave.zachariah@it.uu.se

ML journal clubs

1. Introduction to Machine Learning (2017)
2. Learning Complex Models via Approximate Inference (2018)
3. Adversarial Machine Learning (2019)

dave.zachariah@it.uu.se

ML journal clubs

1. Introduction to Machine Learning (2017)
2. Learning Complex Models via Approximate Inference (2018)
3. Adversarial Machine Learning (2019)

New effort: Seminar course!

dave.zachariah@it.uu.se

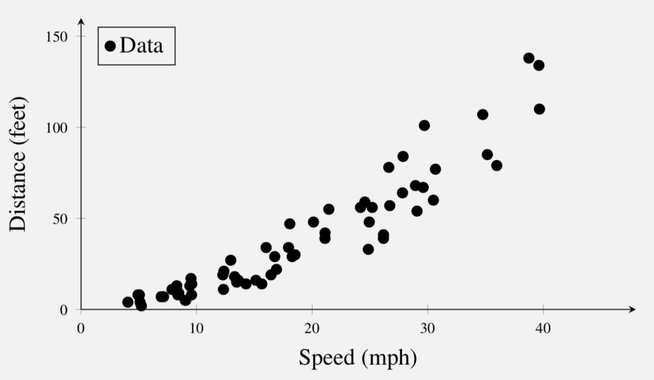
UPPSALA UNIVERSITET

Background

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Textbook example of prediction

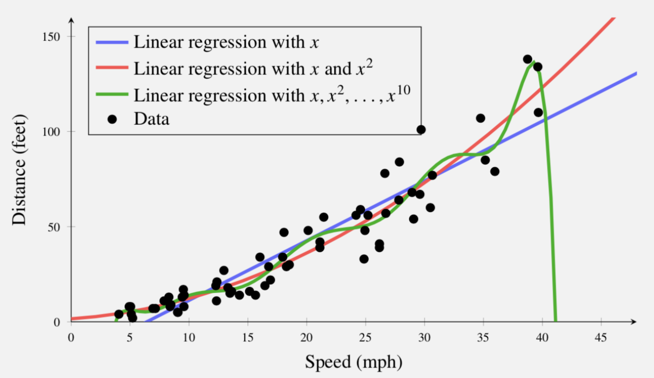


Data on outcome y and covariate x

Lindholm, A. et al.: *Machine Learning - A First Course for Engineers and Scientists*, Cambridge University Press, 2021.

dave.zachariah@it.uu.se

Textbook example of prediction

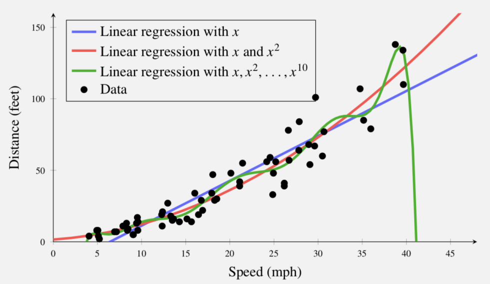


Example of predictor model parameterized by $\boldsymbol{\theta}$:

$$\widehat{y}_{\boldsymbol{\theta}}(x) = \boldsymbol{\theta}^\top \begin{bmatrix} 1 \\ x \end{bmatrix} \qquad \dim(\boldsymbol{\theta}) = 2$$

dave.zachariah@it.uu.se

Increasing 'model complexity'

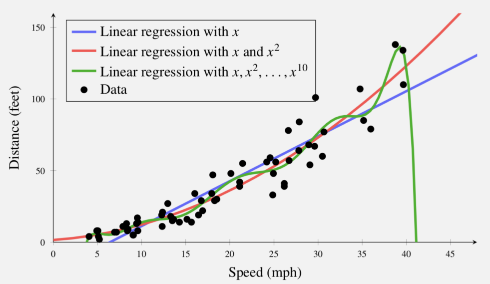


Example of predictor model parameterized by $\boldsymbol{\theta}$:

$$\widehat{y}_{\boldsymbol{\theta}}(x) = \boldsymbol{\theta}^{\top} \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{10} \end{bmatrix} \qquad \dim(\boldsymbol{\theta}) = 11$$

dave.zachariah@it.uu.se

Increasing ‘model complexity’

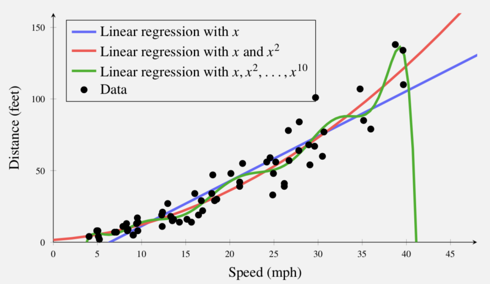


If $\dim(\boldsymbol{\theta}) \leq$ #training samples, then (typically) there exists unique predictor $\widehat{y}_{\boldsymbol{\theta}}(x)$ that minimizes training error, e.g.,

$$E_{\text{train}} = \widehat{\mathbb{E}}[(y - \widehat{y}_{\boldsymbol{\theta}}(x))^2]$$

dave.zachariah@it.uu.se

Increasing 'model complexity'



If $\dim(\boldsymbol{\theta}) > \#$training samples, then many predictors $\widehat{y}_{\boldsymbol{\theta}}(x)$ can reduce the training error to zero, e.g.,

$$E_{\text{train}} = \widehat{\mathbb{E}}[(y - \widehat{y}_{\boldsymbol{\theta}}(x))^2] = 0$$

dave.zachariah@it.uu.se

UPPSALA UNIVERSITET

Overfit and overparameterization

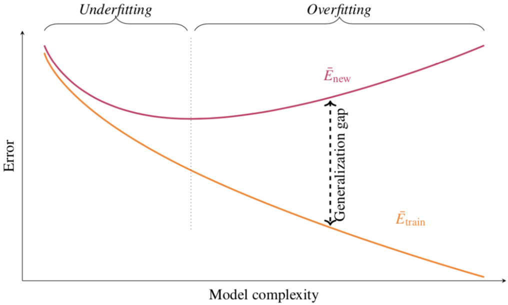


Figure 4.3: Behavior of $\bar{E}_{\text{train}}$ and $\bar{E}_{\text{new}}$ for many supervised machine learning methods, as a function of model complexity. We have not made a formal definition of complexity, but a rough proxy is the number of parameters that are learned from the data. The difference between the two curves is the generalization gap. The training error $\bar{E}_{\text{train}}$

Training error vs. test error

dave.zachariah@it.uu.se

Overfit and overparameterization

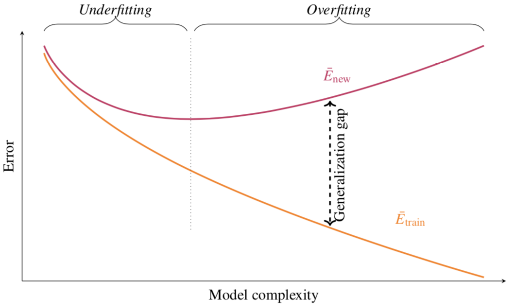


Figure 4.3: Behavior of $\bar{E}_{\text{train}}$ and $\bar{E}_{\text{new}}$ for many supervised machine learning methods, as a function of model complexity. We have not made a formal definition of complexity, but a rough proxy is the number of parameters that are learned from the data. The difference between the two curves is the generalization gap. The training error $\bar{E}_{\text{train}}$

Puzzle: Modern ML methods using

- models with $\dim(\boldsymbol{\theta}) \gg \#$training samples
- stochastic gradient search methods

achieve state-of-the-art test error E_{test}!

dave.zachariah@it.uu.se