Stanford University

CS 329P : Practical Machine Learning (2021 Fall)

5. Model Combination

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

So far…

- Data
- ML Models for different types of data
- Good models perform well on unseen data
  - Model specific metrics VS business metrics
  - Generalization error depends on model / data complexity
  - TODAY: Methods for reducing generalization error

Stanford University

CS 329P : Practical Machine Learning (2021 Fall)

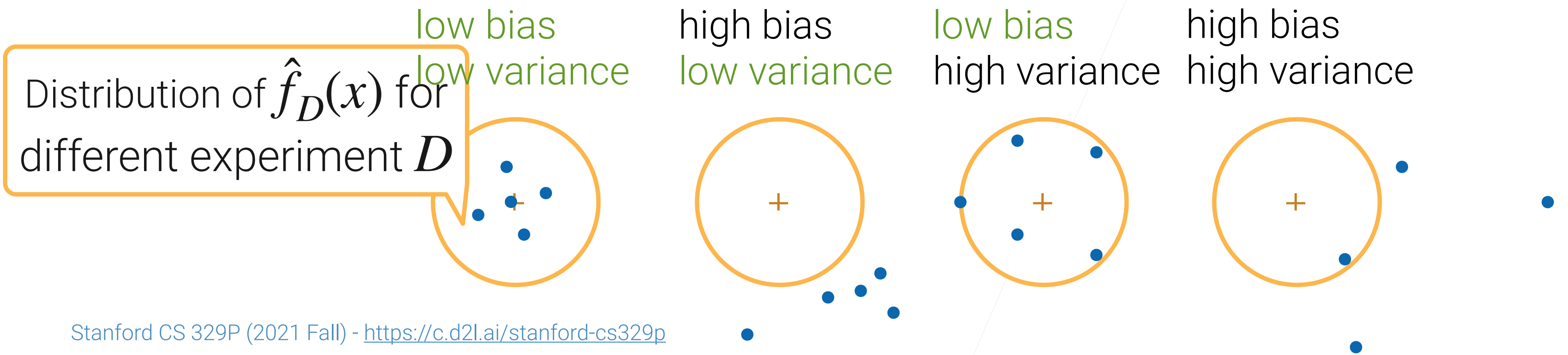
5.1 Bias & Variance

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

Bias & Variance

- Sample data $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ from $y = f(x) + \varepsilon$
- Learn $\hat{f}_D$ from data D by minimizing MSE: $\min_{\hat{f}_D} \sum_{(x_i, y_i) \in D} (y_i - \hat{f}_D(x_i))^2$
- We want $\hat{f}_D$ generalizes well to an unseen data point (x, y).



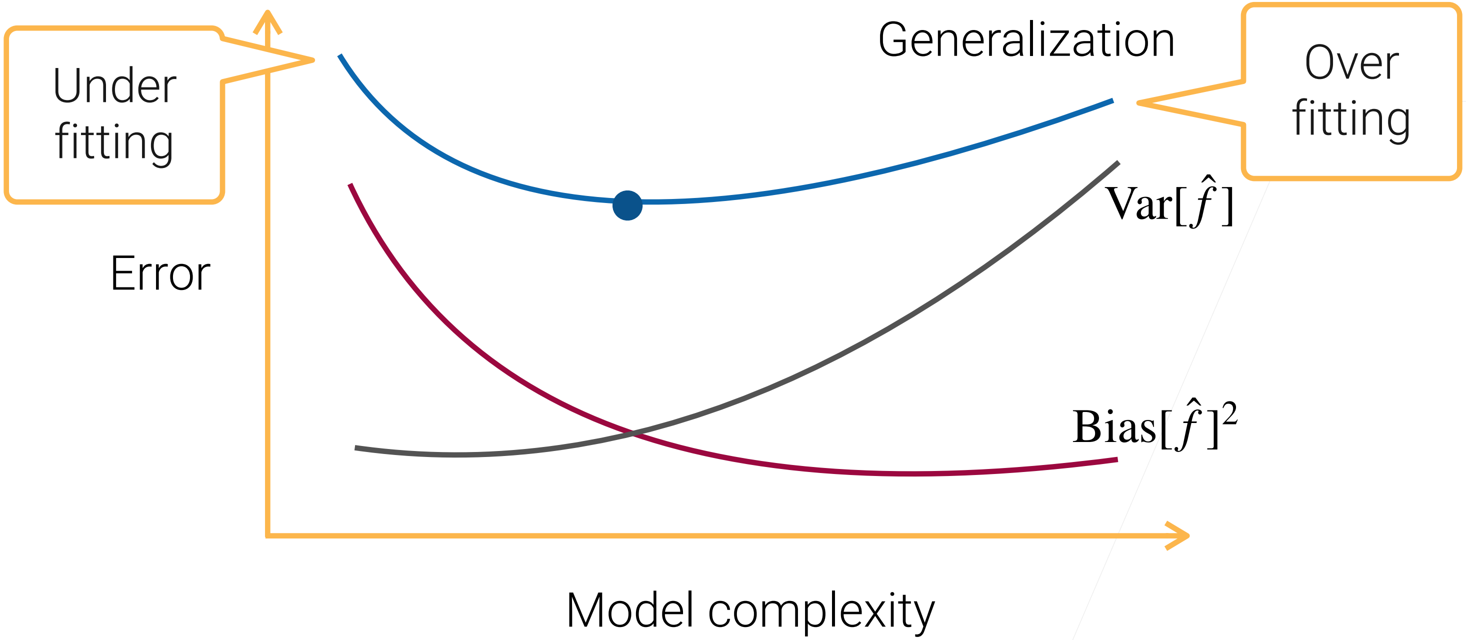
Bias-Variance Decomposition

- Learn $\hat{f}_D$ from dataset D sampled from $y = f(x) + \varepsilon$
- Evaluate generalization error $(y - \hat{f}_D(x))^2$ on a new data point (x, y)

$$
\begin{aligned}
\mathrm{E}_D\left[(y - \hat{f}_D(x))^2\right] &= \mathrm{E}_D\left[\left((f - \mathrm{E}_D[\hat{f}_D]) - (\hat{f}_D - \mathrm{E}_D[\hat{f}_D]) + \varepsilon\right)^2\right] \\
&= (f - \mathrm{E}_D[\hat{f}_D])^2 + \mathrm{E}_D\left[(\hat{f}_D - \mathrm{E}_D[\hat{f}_D]))^2\right] + \varepsilon^2 \\
&= \mathrm{Bias}[\hat{f}_D]^2 + \mathrm{Var}[\hat{f}_D] + \varepsilon^2
\end{aligned}
$$

Bias-Variance Tradeoff

$$E_D\left[(y-\hat{f}_D(x))^2\right] = \mathrm{Bias}[\hat{f}_D]^2 + \mathrm{Var}[\hat{f}_D] + \epsilon^2$$



Reduce Bias & Variance

$$E_D\left[(y-\hat{f}_D(x))^2\right] = \mathrm{Bias}[\hat{f}_D]^2 + \mathrm{Var}[\hat{f}_D] + \epsilon^2$$

- Reduce bias
  - A more complex model
    - e.g. increase #layers, #hidden units of MLP
  - Boosting
  - Stacking
- Reduce variance
  - A simpler model
    - e.g. regularization
  - Bagging
  - Stacking
- Reduce σ^2
  - Improve data

Ensemble learning: train and combine multiple models to improve predictive performance