Bias-Variance Tradeoff

ML Basics

ML process

Machine Learning task

↓

Get and integrate data

↓

Explore and pre-process data

↓

Features

↓

Training and tuning

↓

Evaluation

↓

Deploy, maintain and update

Training and test error

Training error

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

- Prediction error based on **training data**
- with e.g. squared error loss L

Test error

$$\text{Err}_{\mathcal{T}} = \text{E}(L(Y, \hat{f}(X))|\mathcal{T})$$

- Prediction error using **test data** (given training data $\mathcal{T}$)

Training and test error

Training error

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

- Prediction error based on **training data**
- with e.g. squared error loss L

Test error

$$\text{Err}_{\mathcal{T}} = \text{E}(L(Y, \hat{f}(X))|\mathcal{T})$$

- Prediction error using **test data** (given training data $\mathcal{T}$)

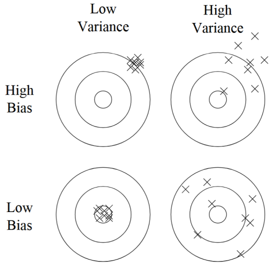
Training and test error

Expected test error decomposition

$$\text{Err}(x_0) = \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) + \text{Var}(\varepsilon)$$

- Minimizing the (expected) test error requires
  - Low bias ($[E\hat{f}(x_0) - f(x_0)]^2$) **and**
  - Low variance ($E[\hat{f}(x_0) - E\hat{f}(x_0)]^2$)

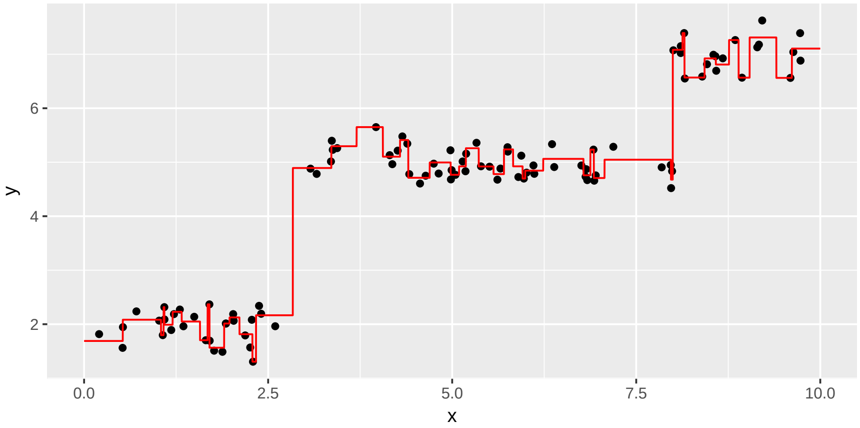
Figure: Bias and variance illustration



Domingos (2012)

Bias-Variance Trade-Off

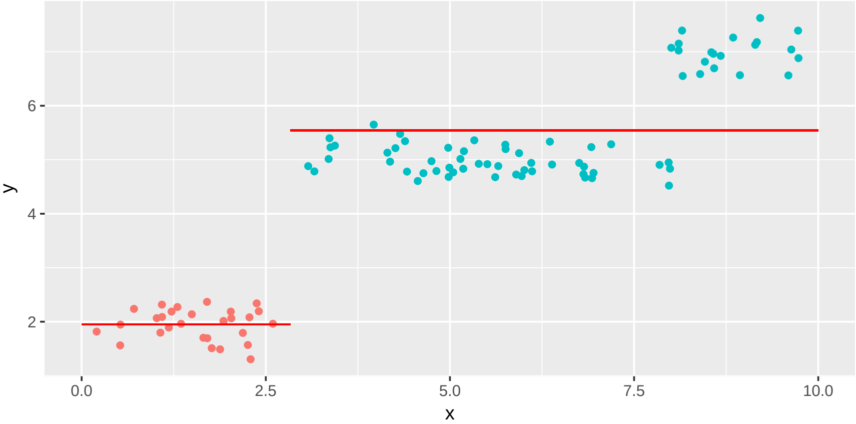
Figure: High Variance in Trees



- High Variance = Different data would lead to a different function
- Overfitting = Poor generalization to new data

Bias-Variance Trade-Off

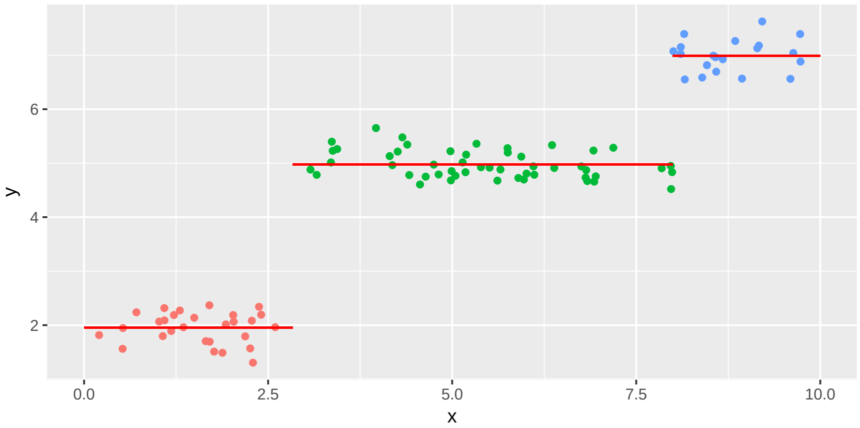
Figure: High Bias in Trees



- High Bias = Blue points are poorly predicted
- Underfitting = Function should adapt better to the data

Bias-Variance Trade-Off

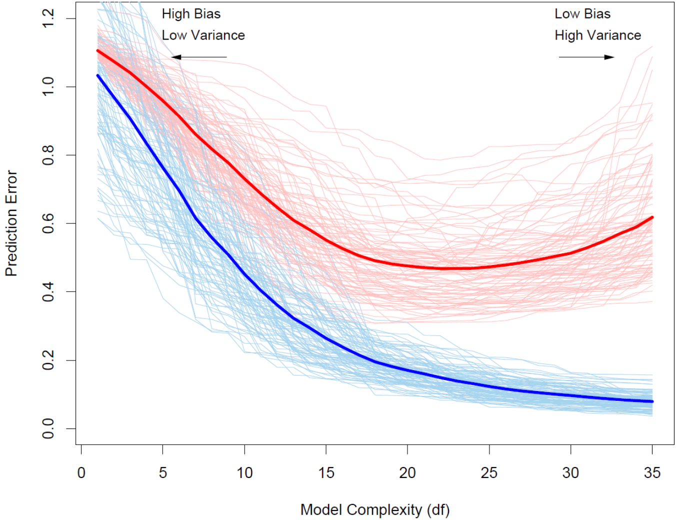
Figure: Optimal Solution



- Goal: Find optimal compromise between bias and variance

Bias-Variance Trade-Off

Figure: Training error and test error by model complexity



Hastie et al. (2009)

Bias-Variance Trade-Off

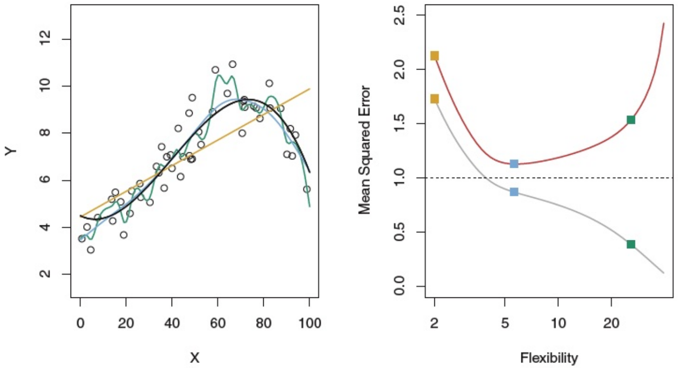


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Introduction to Statistical Learning (James et al. 2013)

Bias-Variance Trade-Off

Goal of prediction (again): Minimal **test error**

$$\arg\min_{f \in \mathcal{F}} \mathbb{E}(L(f(x), y))$$

but we cannot simply minimize its empirical analogue in **training data**

$$\arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i)$$

because this would overfit if the capacity of f is high enough.

Bias-Variance Trade-Off

Solution: Solve (as before)

$$\arg\min_{f \in \mathcal{F}_K} \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i)$$

but f must come from a restricted hypothesis space (limited capacity)

- Tree with at most K leaves
- Regression with $\sum |\beta_j| < K$
- General form: Penalty$(f) < K$

This is **regularization** – in general form:

$$\arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i) + \lambda \cdot \text{Penalty}(f)$$

Quiz

If we have a high bias problem (underfitting), what can be done?

- Add more predictors (= collect more variables or transform existing ones)?
- Allow higher function capacity (= reduce regularization parameter)?
- Use more flexible algorithms (e.g., a tree instead of linear regression)?

If we have a high variance problem (overfitting), what can be done?

- Add more predictors (= collect more variables or transform existing ones)?
- Allow higher function capacity (= reduce regularization parameter)?
- Use more flexible algorithms (e.g., a tree instead of linear regression)?
- Collect more training data?

References

Buskirk, T. D., Kirchner, A., Eck, A., Signorino, C. S. (2018). An Introduction to Machine Learning Methods for Survey Researchers. *Survey Practice 11*(1).

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM 55*(10), 78–87.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.