Excess Risk Decomposition

David Rosenberg

New York University

January 29, 2017

Review: Statistical Learning Theory

Statistical Learning Theory Framework

The Spaces

- $\mathcal{X}$: input space
- $\mathcal{Y}$: output space
- $\mathcal{A}$: action space

Decision Function

A **decision function** produces an action $a \in \mathcal{A}$ for any input $x \in \mathcal{X}$:

$$\begin{aligned} f: \quad \mathcal{X} &\to \mathcal{A} \\ x &\mapsto f(x) \end{aligned}$$

Loss Function

A **loss function** evaluates an action in the context of the output y.

$$\begin{aligned} \ell: \quad \mathcal{A} \times \mathcal{Y} &\to \mathbf{R} \\ (a, y) &\mapsto \ell(a, y) \end{aligned}$$

The Gold Standard: Bayes Decision Function

Definition

The **expected loss** or **"risk"** of a decision function $f : \mathcal{X} \to \mathcal{A}$ is

$$R(f) = \mathbb{E}\ell(f(x), y),$$

where the expectation taken is over $(x, y) \sim P_{\mathcal{X}\times\mathcal{Y}}$.

Definition

A **Bayes decision function** $f^* : \mathcal{X} \to \mathcal{A}$ is a function that achieves the *minimal risk* among all possible functions:

$$R(f^*) = \inf_f \mathbb{E}\ell(f(x), y).$$

- But risk function cannot be computed because we don't know $P_{\mathcal{X}\times\mathcal{Y}}$.

Empirical Risk Minimization

- Let $\mathcal{D}_n = ((x_1, y_1), \ldots, (x_n, y_n))$ be drawn i.i.d. from $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.

Definition

The **empirical risk** of $f : \mathcal{X} \to \mathcal{A}$ with respect to $\mathcal{D}_n$ is

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i).$$

- Minimizing empirical risk over all functions leads to overfitting.

Constrain to a Hypothesis Space

- Hypothesis space $\mathcal{F}$, a set of functions mapping $\mathcal{X} \to \mathcal{A}$
  - Example hypothesis spaces?
- **Empirical risk minimizer** (ERM) **in** $\mathcal{F}$ is

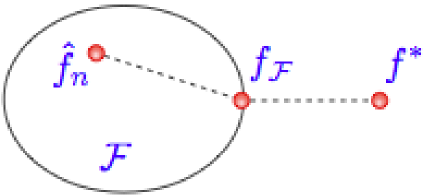
$$\hat{f}_n = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

- **Risk minimizer in** $\mathcal{F}$ is

$$f_{\mathcal{F}} = \underset{f \in \mathcal{F}}{\arg\min} \mathbb{E}\ell(f(x), y).$$

Excess Risk Decomposition

Error Decomposition



$$f^* = \arg\min_f \mathbb{E}\ell(f(X), Y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(X), Y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- **Approximation Error** (of $\mathcal{F}$) $= R(f_{\mathcal{F}}) - R(f^*)$
- **Estimation error** (of $\hat{f}_n$ in $\mathcal{F}$) $= R(\hat{f}_n) - R(f_{\mathcal{F}})$

Figure from Sasha Rakhlin's MLSS Lectures (2012): `http://yosinski.com/mlss12/MLSS-2012-Rakhlin-Statistical-Learning-Theory/`

Excess Risk

Definition

The **excess risk** compares the risk of f to the Bayes optimal f^*:

$$\textbf{Excess Risk}(f) = R(f) - R(f^*)$$

- Can excess risk ever be negative?

Excess Risk Decomposition for ERM

- The excess risk of the ERM $\hat{f}_n$ can be decomposed:

$$\begin{aligned}\textbf{Excess Risk}(\hat{f}_n) &= R(\hat{f}_n) - R(f^*) \\ &= \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}}.\end{aligned}$$

Approximation Error

Approximation error $R(f_{\mathcal{F}}) - R(f^*)$ is

- a property of the class $\mathcal{F}$
- the penalty for restricting to $\mathcal{F}$ rather than all possible functions

Bigger $\mathcal{F}$ mean *smaller* approximation error.

Concept check: Is approximation error a random or non-random variable?

Estimation Error

Estimation error $R(\hat{f}_n) - R(f_{\mathcal{F}})$

- is the performance hit for choosing f using finite training data
- is the performance hit for using empirical risk rather than true risk

With *smaller* $\mathcal{F}$ we expect *smaller* estimation error.

Under typical conditions: "With infinite training data, estimation error goes to zero."

- [Infinite training data solves the *statistical* problem, which is not knowing the true risk.]

Concept check: Is estimation error a random or non-random variable?

ERM Overview

- Given a loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbf{R}$.
- Choose hypothesis space $\mathcal{F}$.
- Use an optimization method to find ERM $\hat{f}_n \in \mathcal{F}$:

$$\hat{f}_n = \underset{f \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

- Data scientist's job:
  - choose $\mathcal{F}$ to balance between approximation and estimation error.
  - as we get more training data, use a bigger $\mathcal{F}$

ERM in Practice

- We've been cheating a bit by writing "argmin".
- In practice, we need a method to find $\hat{f}_n \in \mathcal{F}$.
- For nice choices of loss functions and classes $\mathcal{F}$, the algorithmic problem can be solved to any desired accuracy
  - But takes time – is it worth it?
- For some hypothesis spaces (e.g. neural networks), we don't know how to find $\hat{f}_n \in \mathcal{F}$.

Optimization Error

- In practice, we don't find the ERM $\hat{f}_n \in \mathcal{F}$.
- We find $\tilde{f}_n \in \mathcal{F}$ that we hope is good enough.
- **Optimization error:** If $\tilde{f}_n$ is the function our optimization method returns, and $\hat{f}_n$ is the empirical risk minimizer, then

$$\text{Optimization Error} \ = \ R(\tilde{f}_n) - R(\hat{f}_n).$$

- Can optimization error be negative? Yes!
- But

$$\hat{R}(\tilde{f}_n) - \hat{R}(\hat{f}(n)) \geqslant 0.$$

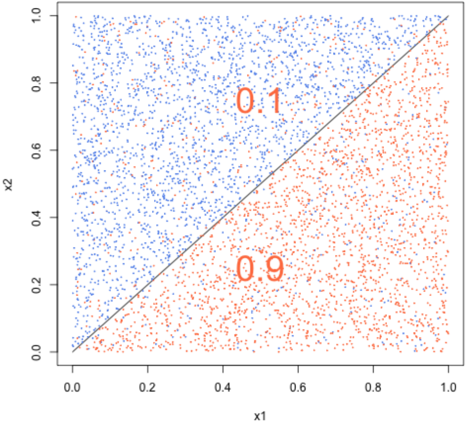
Error Decomposition in Practice

Excess risk decomposition for function $\tilde{f}_n$ returned by algorithm:

$$\begin{aligned}\textbf{Excess Risk}(\tilde{f}_n) &= R(\tilde{f}_n) - R(f^*) \\ &= \underbrace{R(\tilde{f}_n) - R(\hat{f}_n)}_{\text{optimization error}} + \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}}\end{aligned}$$

Excess Risk Decomposition: Example

A Simple Classification Problem



$$\mathcal{Y} = \{\text{blue}, \text{orange}\}$$

$$P_{\mathcal{X}} = \text{Uniform}([0,1]^2)$$

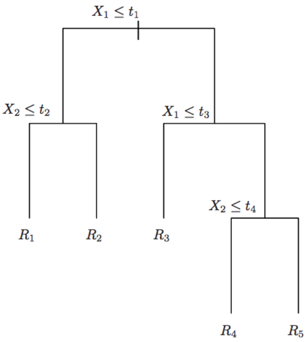
$$\mathbb{P}(\text{orange} \mid x_1 > x_2) = .9$$

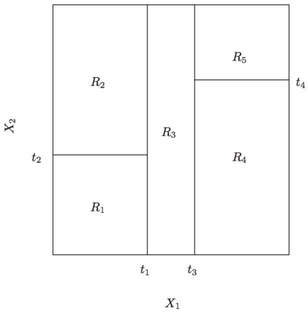
$$\mathbb{P}(\text{orange} \mid x_1 < x_2) = .1$$

Bayes Error Rate $= 0.1$

Binary Decision Trees on $\mathbf{R}^2$

- Consider a binary tree on $\{(X_1, X_2) \mid X_1, X_2 \in \mathbf{R}\}$





From *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

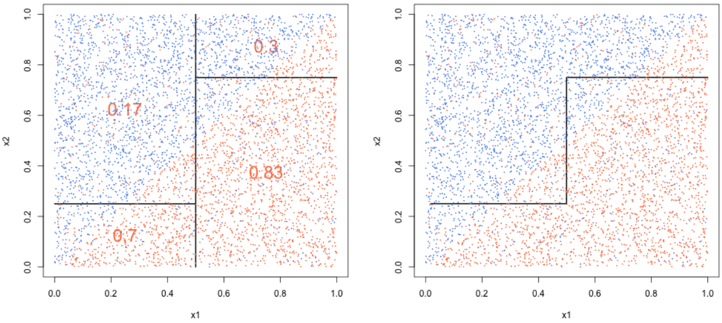
Hypothesis Space: Decision Tree

- $\mathcal{F} = \left\{\text{all decision tree classifiers on } [0,1]^2\right\}$
- $\mathcal{F}_d = \left\{\text{all decision tree classifiers on } [0,1]^2 \text{ with DEPTH} \leqslant d\right\}$
- We'll consider

$$\mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{F}_4 \cdots \subset \mathcal{F}_{15}$$

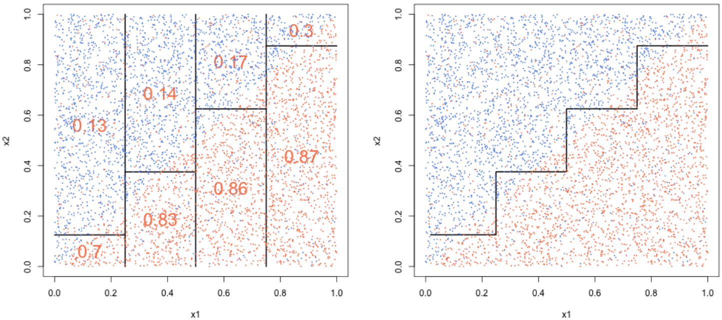
- Bayes error rate $= 0.1$

Theoretical Best in $\mathcal{F}_2$



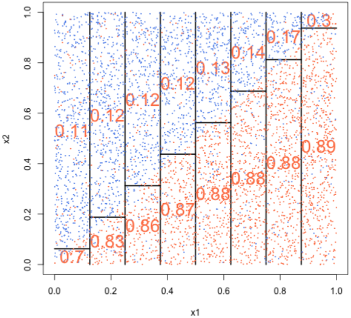
- Risk Minimizer in $\mathcal{F}_2$ (e.g. assuming **infinite training data**); Risk = P(error) = 0.2
- Approximation Error = 0.2 - 0.1 = 0.1

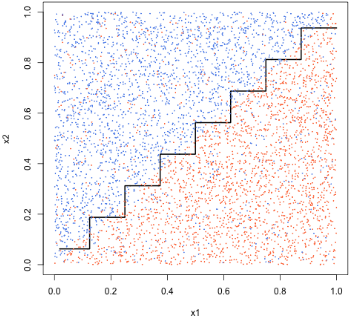
Theoretical Best in $\mathcal{F}_3$



- Risk Minimizer in $\mathcal{F}_3$ (e.g. assuming **infinite training data**); Risk = P(error) = 0.15
- Approximation Error = 0.15 - 0.1 = 0.05

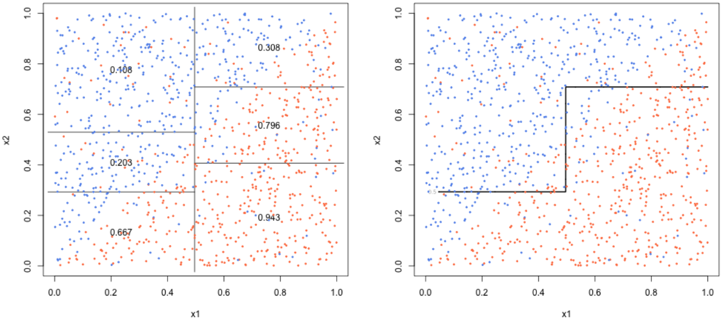
Theoretical Best in $\mathcal{F}_4$





- Risk Minimizer (e.g. assuming **infinite training data**); Risk = P(error) = 0.125
- Approximation Error = 0.125 - 0.1 = 0.025

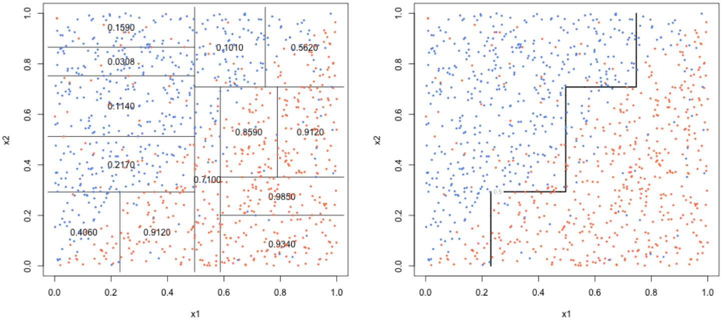
Decision Tree in $\mathcal{F}_3$ Estimated From Sample ($n = 1024$)



$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.176 \pm .004$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.176 \pm .004}_{R(\tilde{f})} - \underbrace{0.150}_{\min_{f \in \mathcal{F}_3} R(f)} = .026 \pm .004$$

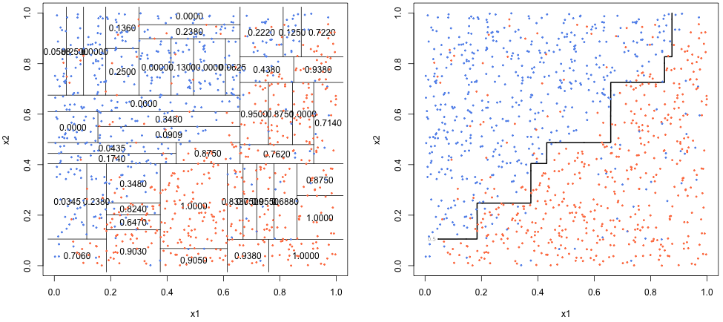
Decision Tree in $\mathcal{F}_4$ Estimated From Sample ($n = 1024$)



$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.144 \pm .005$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.144 \pm .005}_{R(\tilde{f})} - \underbrace{0.125}_{\min_{f \in \mathcal{F}_4} R(f)} = .019 \pm .005$$

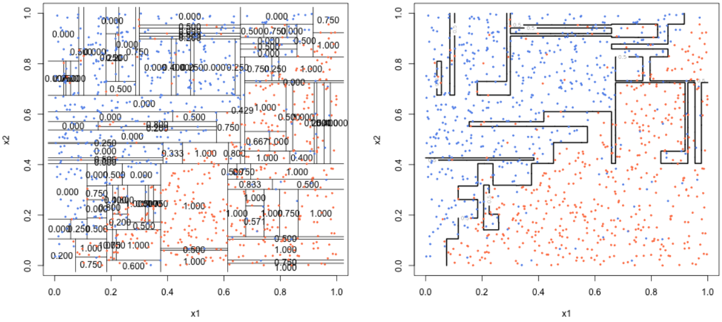
Decision Tree in $\mathcal{F}_6$ Estimated From Sample ($n = 1024$)



$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.148 \pm .007$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.148 \pm .007}_{R(\tilde{f})} - \underbrace{0.106}_{\min_{f \in \mathcal{F}_6} R(f)} = .042 \pm .007$$

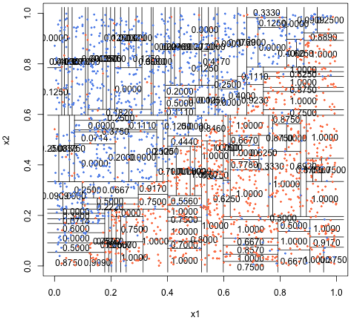
Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 1024$)

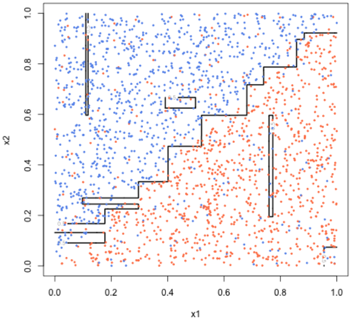


$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.162 \pm .009$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.162 \pm .009}_{R(\tilde{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_8} R(f)} = .061 \pm .009$$

Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 2048$)

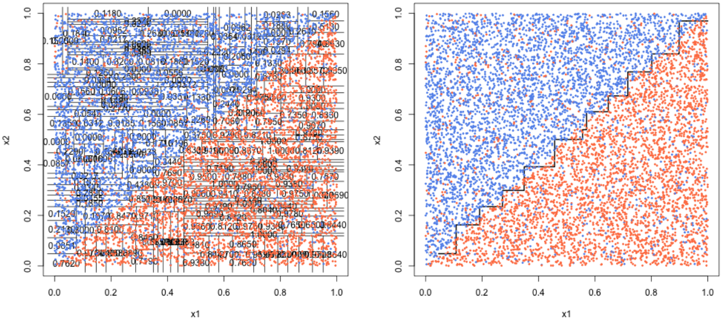




$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.146 \pm .006$$

$$\text{Estimation Error+Optimization Error} = \underbrace{0.146 \pm .006}_{R(\tilde{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_3} R(f)} = .045 \pm .006$$

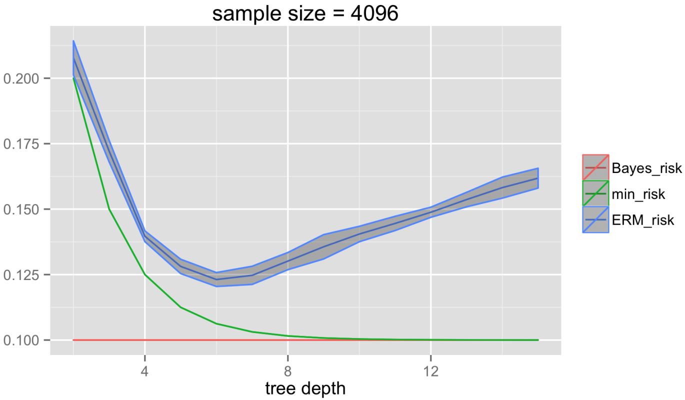
Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 8192$)



$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.121 \pm .002$$

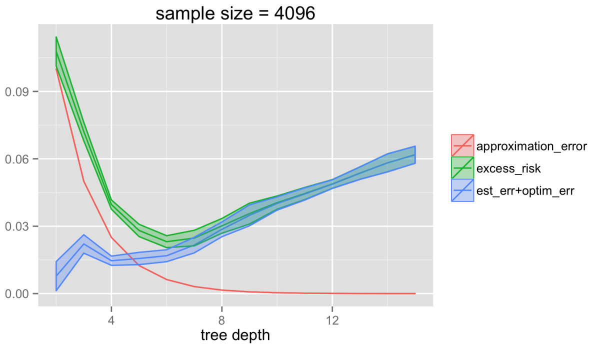
$$\text{Estimation Error+Optimization Error} = \underbrace{0.121 \pm .002}_{R(\tilde{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_3} R(f)} = .019 \pm .002$$

Risk Summary



Why do some curves have confidence bands and others not?

Excess Risk Decomposition



Why do some curves have confidence bands and others not?