Excess Risk Decomposition

David S. Rosenberg

New York University

January 30, 2018
Excess Risk Decomposition
Error Decomposition

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

\[
f^* = \arg \min_{f} \mathbb{E} \ell(f(x), y)
\]

\[
f_{\mathcal{F}} = \arg \min_{f \in \mathcal{F}} \mathbb{E} \ell(f(x), y)
\]

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

Definition

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

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

Can excess risk ever be negative?
The excess risk of the ERM $\hat{f}_n$ can be decomposed:

\[
\text{Excess Risk}(\hat{f}_n) = R(\hat{f}_n) - R(f^*)
\]

\[
= R(\hat{f}_n) - R(f_{\mathcal{F}}) + R(f_{\mathcal{F}}) - R(f^*)
\]

- estimation error
- approximation error
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 considering all possible functions)

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

Concept check: Is approximation error a random or non-random variable?
Estimation error $R(\hat{f}_n) - R(f_F)$

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

With smaller $F$ we expect smaller estimation error.

*Under typical conditions:* “With infinite training data, estimation error goes to zero.”

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

- Given a loss function $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$.
- Choose hypothesis space $\mathcal{F}$.
- Use an optimization method to find ERM $\hat{f}_n \in \mathcal{F}$:
  \[
  \hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i).
  \]

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

- 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

$$\hat{R}(\tilde{f}_n) - \hat{R}(\hat{f}_n) \geq 0.$$
Error Decomposition in Practice

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

$$\text{Excess Risk}(\tilde{f}_n) = R(\tilde{f}_n) - R(f^*)$$
$$= R(\tilde{f}_n) - R(\hat{f}_n) + R(\hat{f}_n) - R(f_F) + R(f_F) - R(f^*)$$

- Optimization error
- Estimation error
- Approximation error

- Concept check: It would be nice to have a concrete example where we find an $\tilde{f}_n$ and look at its error decomposition. Why is this usually impossible?
- But we could construct an artificial example, where we know $P_{\mathcal{X} \times \mathcal{Y}}$ and $f^*$ and $f_F$...
Excess Risk Decomposition: Example
A Simple Classification Problem

\[ y = \{ \text{blue, orange} \} \]

\[ P_X = \text{Uniform}([0, 1]^2) \]

\[ P(\text{orange} | x_1 > x_2) = .9 \]

\[ P(\text{orange} | x_1 < x_2) = .1 \]

Bayes Error Rate = 0.1
Binary Decision Trees on $\mathbb{R}^2$

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

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

- $\mathcal{F} = \left\{ \text{all decision tree classifiers on } [0, 1]^2 \right\}$
- $\mathcal{F}_d = \left\{ \text{all decision tree classifiers on } [0, 1]^2 \text{ with DEPTH } \leq d \right\}$
- We’ll consider
  
  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{F}_4 \cdots \subset \mathcal{F}_{15}$

- Bayes error rate = 0.1
Theoretical Best in $\mathcal{F}_1$

- Risk Minimizer in $\mathcal{F}_1$ has Risk $= \mathbb{P}(\text{error}) = 0.3$.
- Approximation Error $= 0.3 - 0.1 = 0.2$. 
Theoretical Best in $\mathcal{F}_2$

- Risk Minimizer in $\mathcal{F}_2$ has Risk $= \mathbb{P}(\text{error}) = 0.2$.
- Approximation Error $= 0.2 - 0.1 = 0.1$
Theoretical Best in $\mathcal{F}_3$

- Risk Minimizer in $\mathcal{F}_3$ has Risk $= \mathbb{P}(\text{error}) = 0.15$.
- Approximation Error $= 0.15 - 0.1 = 0.05$
Theoretical Best in $\mathcal{F}_4$

- Risk Minimizer in $\mathcal{F}_4$ has Risk $= \mathbb{P}(\text{error}) = 0.125$.
- Approximation Error $= 0.125 - 0.1 = 0.025$
Decision Tree in $\mathcal{F}_3$ Estimated From Sample ($n = 1024$)

\[ R(\tilde{f}) = \mathbb{P}({\text{error}}) = 0.176 \pm 0.004 \]

\[
\text{Estimation Error + Optimization Error} = \underbrace{0.176 \pm 0.004 - 0.150}_{R(\tilde{f})} = 0.026 \pm 0.004
\]

\[ \min_{f \in \mathcal{F}_3} R(f) \]
Decision Tree in $\mathcal{F}_4$ Estimated From Sample ($n = 1024$)

$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.144 \pm 0.005$$

Estimation Error + Optimization Error = \underbrace{0.144 \pm 0.005}_{R(\tilde{f})} - \underbrace{0.125}_{\min_{f \in \mathcal{F}_4} R(f)} = 0.019 \pm 0.005
Decision Tree in $\mathcal{F}_6$ Estimated From Sample ($n = 1024$)

$$R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.148 \pm 0.007$$

Estimation Error + Optimization Error $= 0.148 \pm 0.007 - 0.106 = 0.042 \pm 0.007$
Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 1024$)

\[
R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.162 \pm 0.009
\]

\[
\text{Estimation Error + Optimization Error} = \underbrace{0.162 \pm 0.009}_{R(\tilde{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_8} R(f)} = 0.061 \pm 0.009
\]
Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 2048$)

\[
R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.146 \pm 0.006
\]

\[
\text{Estimation Error} + \text{Optimization Error} = \underbrace{0.146 \pm 0.006}_{R(\tilde{f})} - \underbrace{0.102}_{\min_{f \in \mathcal{F}_3} R(f)} = 0.045 \pm 0.006
\]
Decision Tree in $\mathcal{F}_8$ Estimated From Sample ($n = 8192$)

\[ R(\tilde{f}) = \mathbb{P}(\text{error}) = 0.121 \pm 0.002 \]

\[
\text{Estimation Error} + \text{Optimization Error} = \underbrace{0.121 \pm 0.002}_R(\tilde{f}) - \underbrace{0.102}_{\min_{f \in \mathcal{F}_3} R(f)} = 0.019 \pm 0.002
\]
Risk Summary

Why do some curves have confidence bands and others not?
Why do some curves have confidence bands and others not?