Empirical Risk Minimization and Vapnik-Chervonenkis Theory

Andrew Nobel

September, 2023

Empirical Risk Minimization

Empirical Risk Minimization (ERM)

Setting

- Set $\mathcal{X}$ of features/predictors
- Set $\mathcal{Y}$ of responses/labels
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Recall

- A prediction rule is a map $h : \mathcal{X} \to \mathcal{Y}$
- Loss of h on feature-response pair (x, y) is $\ell(h(x), y)$
- Risk of h on random pair (X, Y) is given by $R(h) = \mathbb{E}\ell(h(X), Y)$

Empirical Risk Minimization (ERM)

Given

- Family $\mathcal{H}$ of prediction rules $h : \mathcal{X} \to \mathcal{Y}$ (possibly infinite)
- Jointly distributed pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$
- Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ iid copies of (X, Y)

Ideally: Find rule $h \in \mathcal{H}$ with minimum risk $R(h)$

ERM: Find rule in $\mathcal{H}$ minimizing empirical risk (proxy for true risk)

$$\hat{h}_n \;=\; \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_n(h) \;=\; \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

Example: Ordinary Least Squares

Setting

- Feature space $\mathcal{X} = \mathbb{R}^d$. Response $\mathcal{Y} = \mathbb{R}$
- Loss $\ell(y', y) = (y' - y)^2$
- Risk of rule $h : \mathbb{R}^d \to \mathbb{R}$ on pair (X, Y) is $R(h) = \mathbb{E}(h(X) - Y)^2$
- Let $\mathcal{H}$ = family of linear rules $h(x) = \langle x, \beta \rangle + \beta_0$

Upshot: ERM coincides with OLS

Example: Histogram Classification Rules

Setting

- Feature space $\mathcal{X}$ is general. Response $\mathcal{Y} = \{0, 1\}$
- Loss $\ell(y', y) = \mathbb{I}(y' \neq y)$
- Risk of rule $h : \mathcal{X} \to \{0, 1\}$ on pair (X, Y) is $R(h) = \mathbb{P}(h(X) \neq Y)$
- Let $\mathcal{H} =$ family of rules constant on the cells of a finite partition π of $\mathcal{X}$

Upshot: ERM coincides with the histogram classification rule

$$\hat{h}_n(x) \;=\; \text{maj-vote}\{Y_i : X_i \in \pi(x)\}$$

Binary Classification: General Case

Note: Every rule $h : \mathcal{X} \to \{0, 1\}$ corresponds to a subset of $\mathcal{X}$ and v.v.

- Let $\mathcal{C}$ = family of subsets of $\mathcal{X}$. For $C \in \mathcal{C}$ let $h_C(x) = \mathbb{I}(x \in C)$
- Define $\mathcal{H} = \{h_C : C \in \mathcal{C}\}$. Then ERM finds set $C \in \mathcal{C}$ minimizing

$$\hat{R}_n(h_C) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h_C(X_i) \neq Y_i) = \frac{1}{n} \sum_{i=1}^{n} |\mathbb{I}(X_i \in C) - Y_i|$$

Examples

- $\mathcal{C}$ = all half-spaces in $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{C}$ = all spheres in $\mathcal{X} = \mathbb{R}^d$

Caveat: Often, computationally efficient algorithms for ERM don't exist

Downward Bias of ERM Training Error

Idea: ERM rule $\hat{h}_n$ is defined by minimizing training error. Thus we expect the training error of $\hat{h}_n$ to be optimistic

Fact: Let $\hat{h}_n$ be ERM rule for a family $\mathcal{H}$ based on observations D_n. Then

$$\mathbb{E}\hat{R}_n(\hat{h}_n) \leq R(\hat{h}_n)$$

Assessing Performance of ERM

Ideal: For a given data generating distribution (X, Y) we would like to find the global optimal rule

$$h^* = \operatorname*{argmin}_h R(h)$$

where minimum is over all functions $h : \mathcal{X} \to \mathcal{Y}$. Note h^* depends on (X, Y)

In practice: Two issues

- (X, Y) unknown, accessible only through observations D_n
- Optimal rule h^* for (X, Y) need not be in $\mathcal{H}$

Estimation and Approximation Error

Easy to see: For any (X, Y) and any procedure h_n selecting rules in $\mathcal{H}$

$$R(\hat{h}_n) - R(h^*) \ = \ \left[R(\hat{h}_n) - \min_{h \in \mathcal{H}} R(h) \right] + \left[\min_{h \in \mathcal{H}} R(h) - R(h^*) \right]$$

- [L] = *Estimation error:* risk of $\hat{h}_n$ vs best rule in $\mathcal{H}$ [random]
- [R] = *Approximation error:* best rule in $\mathcal{H}$ vs optimal rule h^* [fixed]

For ERM, as we increase the size of the target family $\mathcal{H}$

- Estimation error tends to increase
- Approximation error decreases

Bound on Estimation Error for ERM

Once the family $\mathcal{H}$ and distribution (X, Y) are fixed, the approximation error is fixed. Focus on evaluation of the estimation error.

Fact: Let $\hat{h}_n$ be ERM estimator for family $\mathcal{H}$. For every distribution (X, Y)

$$0 \leq R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \;\leq\; 2 \sup_{h \in \mathcal{H}} |\hat{R}_n(h) - R(h)|$$

Analysis of Empirical Risk Minimization

Finite Family $\mathcal{H}$

Analysis of ERM: Finite $\mathcal{H}$, Bounded Loss, Zero Error

Fact: Suppose that $\mathcal{H}$ is finite, $\ell(y', y) \in [0, 1]$, and $\min_{h \in \mathcal{H}} R(h) = 0$. Then for every distribution (X, Y), sample size $n \geq 1$, and $t \geq 0$

$$\mathbb{P}\left(R(\hat{h}_n) > t\right) \;\leq\; |\mathcal{H}|\, e^{-nt}$$

Corollary: For every $\delta > 0$, with probability at least $1 - \delta$

$$R(\hat{h}_n) \;\leq\; \frac{1}{n} \log \frac{|\mathcal{H}|}{\delta}$$

and we can bound the expected risk as

$$\mathbb{E}R(\hat{h}_n) \;\leq\; \frac{(\log |\mathcal{H}| + 1)}{n}$$

Analysis of ERM: Finite $\mathcal{H}$, Bounded Loss

Fact: Suppose that $\mathcal{H}$ is finite and $\ell(y', y) \in [0, 1]$. Then for every distribution (X, Y), sample size n, and $t \geq 0$

$$\mathbb{P}\left(R(\hat{h}_n) - \min_{h \in \mathcal{H}} R(h) > t\right) \;\leq\; |\mathcal{H}|\, e^{-nt^2/2}$$

Corollary: For every $\delta > 0$, with probability at least $1 - \delta$

$$R(\hat{h}_n) \;\leq\; \min_{h \in \mathcal{H}} R(h) + \sqrt{\frac{2}{n} \log \frac{|\mathcal{H}|}{\delta}}$$

and we can bound the expected risk as

$$\mathbb{E}R(\hat{h}_n) \;\leq\; \min_{h \in \mathcal{H}} R(h) + \sqrt{\frac{2(\log |\mathcal{H}| + 1)}{n}}$$

Analysis of Empirical Risk Minimization

Infinite Family $\mathcal{H}$

ERM and Uniform Laws of Large Numbers

Analysis of ERM for infinite families leverages ideas from uniform laws of large numbers to bound estimation error

Fact: Upper bound on estimation error can be written as

$$\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| \;=\; \sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z) \right|$$

where $Z_i = (X_i, Y_i)$ are iid copies of $Z = (X, Y)$, and

$$\mathcal{G} \;=\; \{g(x,y) = \ell(h(x),y) : h \in \mathcal{H}\}$$

is the set of error functions associated with the prediction rules in $\mathcal{H}$

Note: If the loss function ℓ is bounded, so are the functions $g \in \mathcal{G}$

Uniform Laws of Large Numbers

Setting

- Measurable space $(\mathcal{X}, \mathcal{A})$
- Family $\mathcal{F}$ of bounded, measurable functions $f : \mathcal{X} \to [a, b]$
- $X_1, \ldots, X_n$ iid copies of $X \in \mathcal{X}$

Of interest: Worst-case difference between averages and expectations

$$\hat{\Delta}_n(\mathcal{F}) \;=\; \Delta_n(X_1^n : \mathcal{F}) \;=\; \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right|$$

Uniform Laws of Large Numbers, cont.

Recall: For each fixed $f \in \mathcal{F}$, each $n \geq 1$, and each $t > 0$, Hoeffding gives

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| > t \right) \;\leq\; 2 \,\exp\left\{ \frac{-2nt^2}{(b-a)^2} \right\}$$

Goal: Similar bound for uniform deviations of the form

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| > t \right) \;\leq\; \Gamma_n(t, X, \mathcal{F}) \,\exp\left\{ \frac{-2nt^2}{(b-a)^2} \right\}$$

where $\Gamma_n(\cdot)$ measures the complexity of $\mathcal{F}$ at resolution t on samples of size n drawn from the distribution of X

First Step: Concentration

Fact

1. Function $F(x_1^n) = \Delta_n(x_1^n : \mathcal{F})$ has difference coefficients $c_i = (b-a)/n$
2. By the bounded difference inequality, for all $t > 0$

$$\mathbb{P}(|\hat{\Delta}_n(\mathcal{F}) - \mathbb{E}\hat{\Delta}_n(\mathcal{F})| \;>\; t) \;\leq\; 2\, e^{-2nt^2/(b-a)^2}$$

So, it remains to analyze $\mathbb{E}\hat{\Delta}_n(\mathcal{F})$...

Second Step: Symmetrization

Fact: For every $n \geq 1$

$$\mathbb{E}\hat{\Delta}_n(\mathcal{F}) \;\leq\; 2\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right]$$

Here $\varepsilon_1, \ldots, \varepsilon_n \in \{-1, 1\}$ are iid with $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$, and are independent of $X_1, \ldots, X_n$

Note

- Random signs $\varepsilon_1, \ldots, \varepsilon_n$ referred to as Rademacher variables
- Upper bound called expected Rademacher complexity of $\mathcal{F}$ on X_1^n
- Complexity measures ability of functions $f \in \mathcal{F}$ to track noise

Focus on Families of Sets

Restriction: Assume that $\mathcal{F} = \{\mathbb{I}_C(x) : C \in \mathcal{C}\}$ is the family of indicator functions associated with a collection $\mathcal{C}$ of subsets of $\mathcal{X}$

Given $X_1, \ldots, X_n \in \mathcal{X}$ iid copies of X, define the discrepency of $\mathcal{C}$ by

$$\hat{\Delta}_n(\mathcal{C}) \;=\; \Delta_n(X_1^n : \mathcal{C}) \;=\; \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n I_C(X_i) - \mathbb{P}(X \in C) \right|$$

Opportunity: Measure the complexity of $\mathcal{C}$ using combinatorial ideas

- Shatter coefficient
- VC-dimension

Shatter Coefficient

Idea: Let $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be finite. Every set $C \in \mathcal{C}$ induces a subset $C \cap S$ of S. Number of induced subsets reflects complexity of $\mathcal{C}$

Definition: The *shatter coefficient* of $\mathcal{C}$ on $x_1, \ldots, x_n \in \mathcal{X}$ is the number of *distinct* subsets of $x_1, \ldots, x_n$ induced by sets in $\mathcal{C}$:

$$S(x_1^n : \mathcal{C}) \;=\; |\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}|$$

- Note that $1 \leq S(x_1^n : \mathcal{C}) \leq 2^n$
- If $S(x_1^n : \mathcal{C}) = 2^n$ then $\mathcal{C}$ induces every subset of $x_1, \ldots, x_n$ and we say that $\mathcal{C}$ *shatters* $x_1, \ldots, x_n$

Third Step: Bound on Expected Rademacher Complexity

Fact: If $\varepsilon \in \{-1, 1\}$ is Rademacher, then its MGF satisfies $M_\varepsilon(s) \leq e^{s^2/2}$

Fact: For every $n \geq 1$ we have

$$\mathbb{E}\left[\sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mathbb{I}(X_i \in C) \right| \right] \leq \sqrt{\frac{2 \log \mathbb{E} S(X_1^n : \mathcal{C})}{n}}$$

Vapnik-Chervonenkis Inequality

Fact: Let $\mathcal{C}$ be a family of subsets of $\mathcal{X}$, and let $X_1, X_2, \ldots, X \in \mathcal{X}$ be iid. For each $t > 0$,

$$\mathbb{P}\left(\sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \in C) - \mathbb{P}(X \in C) \right| \geq t \right) \leq \mathbb{E}S(X_1^n : \mathcal{C}) e^{-nt^2/8}$$

Better, but less attractive, upper bound is $(\mathbb{E}S(X_1^n : \mathcal{C}))^{16} \, e^{-2nt^2}$

Note: Complexity of $\mathcal{C}$ reflected in expected shatter coefficient $\mathbb{E}S(X_1^n : \mathcal{C})$

Application to Empirical Risk Minimization

Given class $\mathcal{H}$ of binary decision rules $h : \mathcal{X} \to \{0, 1\}$, consider associated family $\mathcal{A}$ of sets

$$A = (h^{-1}(1) \times \{0\}) \cup (h^{-1}(0) \times \{1\}) \subseteq \mathcal{X} \times \{0, 1\}$$

where h ranges over $\mathcal{H}$. Easy to see that $S((x, y)_1^n : \mathcal{A}) \leq S(x_1^n : \mathcal{H})^2$

Cor: Given observations D_n iid $\sim (X, Y)$ the ERM estimator $\hat{h}_n$ satisfies

$$\mathbb{P}\left(R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \geq t \right) \leq \mathbb{E} S(X_1^n : \mathcal{C})^2 \, e^{-nt^2/32}$$

The Vapnik-Chervonenkis Dimension

The VC Dimension

Recall: A family $\mathcal{C} \subseteq 2^{\mathcal{X}}$ shatters points $x_1, \ldots, x_n \in \mathcal{X}$ if $S(x_1^n : \mathcal{C}) = 2^n$

Definition: The *VC-dimension* of $\mathcal{C}$, denoted $\dim(\mathcal{C})$, is the largest k such that $\mathcal{C}$ shatters *some* set of k points in $\mathcal{X}$.

If $\mathcal{C}$ shatters arbitrarily large finite sets, then $\dim(\mathcal{C}) = +\infty$

First Examples

- $\mathcal{C} =$ all half-lines $(-\infty, t]$ in $\mathbb{R}$, $\dim(\mathcal{C}) = 1$
- $\mathcal{C} =$ all discs in $\mathbb{R}^2$, $\dim(\mathcal{C}) = 3$
- $\mathcal{C} =$ all convex sets in $\mathbb{R}^2$, $\dim(\mathcal{C}) = +\infty$

Sauer's Lemma

Sauer's Lemma establishes a connection between the VC-dimension of a family $\mathcal{C}$ and its shatter coefficients

Lemma: If $\mathcal{C}$ has VC-dimension d then for all $n \geq 1$ and all $x_1, \ldots, x_n \in \mathcal{X}$

$$S(x_1^n : \mathcal{C}) \;\leq\; \sum_{k=0}^{d} \binom{n}{k} \;\leq\; (n+1)^d$$

Upshot: If $\dim(\mathcal{C}) = d$ then the shatter coefficient of $\mathcal{C}$ grows at most polynomially with degree d

VC Dimension of Zero-Level Sets

Lemma: Let $\mathcal{G}$ be a v-dimensional vector space of functions $g : \mathcal{X} \to \mathbb{R}$. Let $\mathcal{C}$ be the family of sets

$$C = \{x : g(x) \geq 0\}$$

where g ranges over $\mathcal{G}$. Then $\dim(\mathcal{C}) \leq v$

Corollary

1. If $\mathcal{C} =$ all half-spaces in $\mathbb{R}^d$ then $\dim(\mathcal{C}) \leq d + 1$
2. If $\mathcal{C} =$ all closed balls in $\mathbb{R}^d$ then $\dim(\mathcal{C}) \leq d + 2$
3. If $\mathcal{C} =$ all ellipsoids $\{x : x^t Ax \leq 1\}$ where $A \in \mathbb{R}^{d \times d}$ and $A \geq 0$ then $\dim(\mathcal{C}) \leq (d + 1)d/2 + 1$

Lower Bounds

Canonical Problem: Identifying Direction of Bias

Idea: Given coin with $P(\text{heads})$ slightly above or below $1/2$. How difficult is it to determine the *direction* of the bias based on m flips?

DoB Model: Fix $\epsilon \in (0, 1)$

1. Sign variable $\sigma \in \{-1, 1\}$ with $\mathbb{P}(\sigma = 1) = \mathbb{P}(\sigma = 1) = 1/2$
2. Flips $Y_1, \ldots, Y_m \,|\, \sigma \sim$ iid Bern$(1/2 + \sigma\epsilon/2)$

DoB Problem: Size of bias is $\epsilon/2$, direction of bias determined by σ

- Decision rule $h : \{0, 1\}^m \to \{-1, 1\}$ has risk $R(h) = \mathbb{P}(h(Y_1^m) \neq \sigma)$
- Find lower bound on the risk $R(h)$ of any decision rule

Preliminaries

Fact 1: The conditional probability $\eta(y_1^m) = \mathbb{P}(\sigma = 1 \,|\, Y_1^m = y_1^m)$ can be written in the form

$$\eta(y_1^m) \;=\; \frac{1}{1 + c^{m_0 - m_1}}$$

where we have

$$m_0 = \sum_{i=1}^{m} \mathbb{I}(y_i = 0) \qquad m_1 = \sum_{i=1}^{m} \mathbb{I}(y_i = 1) \qquad c = \frac{1/2 + \epsilon/2}{1/2 - \epsilon/2} > 1$$

Fact 2: If $U \sim \text{Bin}(m, p)$ and $V \sim \text{Bin}(m, 1-p)$ then $U \stackrel{d}{=} m - V$. In particular,

$$|2U - m| \;\stackrel{d}{=}\; |2V - m|$$

First Step: Characterize Optimal Decision Rule

Fact: Let $\eta(y_1^m) = \mathbb{P}(\sigma = 1 | Y_1^m = y_1^m)$. The optimal decision rule h^* for the direction of bias problem is

$$h^*(y_1^m) = \begin{cases} 1 & \text{if } \eta(y_1^m) \geq 1/2 \\ -1 & \text{if } \eta(y_1^m) < 1/2 \end{cases}$$

Equivalently, $h^*(y_1^m) = 1$ iff $m_1 \geq m_0$. The risk of h^* is

$$\mathbb{P}(h^*(Y_1^m) \neq \sigma) = \mathbb{E}\min(\eta(Y_1^m), 1 - \eta(Y_1^m))$$

Cor: Any decision rule h for DoB has risk $R(h) \geq \mathbb{E}\min(\eta(Y_1^m), 1 - \eta(Y_1^m))$

Lower Bound for Risk in DoB

Prop'n: For each $\epsilon \in [0,1)$ and $m \geq 1$ the optimal risk for direction of bias

$$R^* \;\geq\; \frac{1}{2}\exp\left\{\frac{-2\epsilon(\sqrt{m}+m\epsilon)}{1-\epsilon}\right\} \;:=\; L(m:\epsilon)$$

Note that $L(\cdot:\epsilon)$ is monotone decreasing and convex

Fact: If $1 \leq Z \leq 1$ then for all $\gamma \in [0,1)$

$$\mathbb{P}(Z > \gamma) \;\geq\; \frac{\mathbb{E}Z - \gamma}{1-\gamma} \;>\; \mathbb{E}Z - \gamma$$

Lower Bound for Classification, General Case

Given: Family $\mathcal{H}$ of binary classification rules with VC-dim d, and a procedure h_n producing rules in $\mathcal{H}$

Theorem: If $n \geq 4d$ there is a distribution (X, Y) such that

$$\mathbb{E}\left[R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h)\right] \;\geq\; \frac{1}{2}\sqrt{\frac{d}{n}}e^{-8}$$

when $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ are iid $\sim (X, Y)$

Corollary: Under the same conditions

$$\mathbb{P}\left(R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \geq \frac{1}{2}\sqrt{\frac{d}{n}}\right) \;\geq\; e^{-8}$$

Lower Bound for Classification, Zero Error Case

Theorem: Let $\mathcal{H}$ be a family of binary classification rules with VC-dim d. Consider the family $\mathcal{P}$ of joint distributions (X, Y) for which

$$\min_{h \in \mathcal{H}} R(h) = 0.$$

For each $n \geq 1$ there exists a distribution $P \in \mathcal{P}$ such that

$$\mathbb{E}\left[R(\hat{h}_n)\right] \;\geq\; \frac{d-1}{2en}$$

when observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn iid from P