Empirical Risk Minimization

Andrew Nobel

March, 2020

Alternative View of Classification Procedures

Given a classification procedure ϕ_n, let

$$\mathcal{F} = \{\phi_n(x : d_n) : d_n \in (\mathcal{X} \times \{0, 1\})^n\}$$

be the family of all possible classification rules it can produce. Note that

- ϕ_n selects rule $\hat{\phi}_n \in \mathcal{F}$ based on observations D_n
- selection involves fitting rules to observations D_n via indirect, approximate minimization of training error $\hat{R}_n$

Exact minimization of training error not computationally feasible, but provides a useful theoretical framework for understanding

- Role of family $\mathcal{F}$
- Tradeoff between performance and complexity

Empirical Risk Minimization (ERM)

Given: Large finite family of fixed classification rules

$$\mathcal{F} \; = \; \{\phi_1, \ldots, \phi_K\}$$

ERM: Given $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ select rule $\phi \in \mathcal{F}$ with smallest number of misclassifications. Formally, let

$$\hat{\phi}_n \; = \; \operatorname*{argmin}_{\phi \in \mathcal{F}} \hat{R}_n(\phi) \; = \; \operatorname*{argmin}_{\phi \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\phi(X_i) \neq Y_i)$$

Fact: (Training error bias) The ERM rule $\hat{\phi}_n$ satisfies the inequality

$$R(\hat{\phi}_n) \; \geq \; \mathbb{E}\hat{R}_n(\hat{\phi}_n)$$

Estimation and Approximation Error

Given: Family of rules $\mathcal{F}$, joint distribution (X, Y). How good is $\hat{\phi}_n$?

- Note: Bayes rule $\phi^*(x)$ for (X, Y) probably not in $\mathcal{F}$

Compare conditional risk $R(\hat{\phi}_n)$ and Bayes risk $R(\phi^*)$. Easy to see that

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

- [1] = *Estimation error:* $\hat{\phi}_n$ vs best rule in $\mathcal{F}$ (random)
- [2] = *Approximation error:* best rule in $\mathcal{F}$ vs Bayes rule (fixed)

In general: If $\mathcal{F}$ gets bigger EstE increases while AppE decreases

Bound on Estimation Error for ERM

Fact: If $\hat{\phi}_n$ is the ERM rule derived from a family $\mathcal{F}$ then the estimation error

$$0 \leq R(\hat{\phi}_n) - \min_{\phi \in \mathcal{F}} R(\phi) \ \leq \ 2 \max_{\phi \in \mathcal{F}} |R(\phi) - \hat{R}_n(\phi)|$$

Upshot

- For finite families $\mathcal{F}$ we can control the estimation error using Chebyshev's or Hoeffding's inequalities plus the union bound
- For infinite families $\mathcal{F}$ we can control the estimation error using Vapnik-Chervonenkis inequalities and uniform LLNs