Empirical Risk Minimization

Andrew Nobel

October, 2021

Different Perspective on Classification

Background: Given a classification procedure ϕ_n, consider the family of all possible classification rules it can produce

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

- Procedure ϕ_n uses observations D_n to select a rule $\hat{\phi}_n \in \mathcal{F}$
- Selection process typically seeks rule in $\mathcal{F}$ that approximately minimizes training error $\hat{R}_n$

Idealization: Minimizing training error provides a useful theoretical framework for understanding classification procedures

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

Empirical Risk Minimization (ERM)

Ingredients

- Finite family $\mathcal{F} = \{\phi_1, \ldots, \phi_K\}$ of classification rules
- Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$ iid copies of (X, Y)

ERM: Select rule $\phi \in \mathcal{F}$ with smallest number of misclassifications

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

Downward bias of training error:

$$R(\hat{\phi}_n^{\text{ERM}}) \geq \mathbb{E}\hat{R}_n(\hat{\phi}_n^{\text{ERM}})$$

Estimation and Approximation Error

Question: In general, Bayes rule ϕ^* not in $\mathcal{F}$. How good is $\hat{\phi}_n^{\text{ERM}}$?

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

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

- [L] = *Estimation error:* $\hat{\phi}_n^{\text{ERM}}$ vs best rule in $\mathcal{F}$ (random)
- [R] = *Approximation error:* best rule in $\mathcal{F}$ vs Bayes rule (fixed)

Note: If $\mathcal{F}$ gets bigger estimation error increases while approximation error decreases

Bound on Estimation Error for ERM

Fact: If $\hat{\phi}_n^{\text{ERM}}$ is 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