CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Structural Risk Minimization

Lecturer: Michael I. Jordan *Scribes: Omid Shakernia*

Recall from last time, the following bounds of the growth function in terms of the VC dimention:

$$N^{\Lambda}(l) \leq \sum_{i=0}^{h} \binom{l}{i} \leq \left(\frac{el}{h}\right)^{h}$$

$$H^{\Lambda}(l) \leq H^{\Lambda}_{ann}(l) \leq G^{\Lambda}(l) \leq h\left(1 + \ln\frac{l}{h}\right)$$

1 Main results from VC theory

For a given $\eta \in [0, 1]$, the following "risk bound" in terms of the empirical risk and the VC dimension holds with probability $1 - \eta$:

$$R(\alpha_l) \leq R_{\text{emp}}(\alpha_l) + \sqrt{\frac{h(\ln(2l/h) + 1) - \ln(\eta/4)}{l}} + \frac{1}{l}. \tag{1}$$

Also, the following risk bound in terms of the minimal risk and the VC dimension holds with probability $1 - 2\eta$:

$$R(\alpha_l) \leq R(\alpha_0) + \sqrt{\frac{h(\ln(2l/h) + 1) - \ln(\eta/4)}{l}} + \frac{1}{l} - \sqrt{-\frac{\ln\eta}{2l}}. \tag{2}$$

2 Structural Risk Minimization

Although the bounds may not be tight, one can exploit the *shape* of these bounds for purposes of *model selection*. Vapnik's version of model selection is called "Structural Risk Minimization."

A **structure** is a nested set classes of function S_i such that $S_1 \subseteq S_2 \subseteq \cdots$. An **admissible structure** is a structure for which:

- S_k has finite VC dimension h_k for each k,
- $S = \bigcup_k S_k$ is *dense* in $\{Q(z, \alpha)\}$ in the $L_1(F)$ metric.[1]

The latter is a technical definition that ensures we are considering a sufficiently rich class of functions. That is, for any possible $Q(z, \alpha)$, and for any $\epsilon > 0$, there exists a function in S that is within an ϵ ball of $Q(z, \alpha)$. The truth need not be in the class of functions, but we can get arbitrarily close to it.

A natural question is the following: How do the bounds given in equations 1 and 2 behave as we range over the structure? Consider the following observations:

[1] The $L_1(F)$ metric is the usual L_1 metric weighted by the distribution F: $\|f - g\|_{L_1(F)} = \int |f(z) - g(z)| dF(z)$

- Since classes of functions are nested, the empirical risk R_{emp} decreases as the complexity of classes of functions (and hence VC dimension h) increases.
- The confidence bound (the second term in equations 1 and 2) increases as h increases.

The risk bound is the sum of the confidence bound and the empirical risk over the entire structure. This suggests a principled way of choosing the optimal class of functions by choosing that which minimizes the risk bound over the entire structure. This is really a form of model selection. The familiar graph relating the risk bound to the empirical risk and confidence bound is shown in Figure 1.

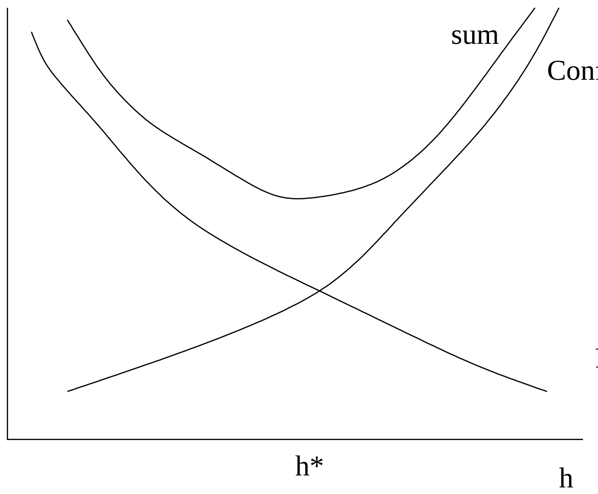

Figure 1: Structural Risk Minimization is the model selection technique of minimizing the sum of the confidence bound and empirical risk.

3 A Consistency Theorem

As is typical with consistency theorems, the following result shows that the method of structural risk minimization works asymptotically:

Theorem 3.1 *Consider an* a priori *rule $n = n(l)$ that selects a structure S_n based on l data points. Define the following:*

$$\alpha_l^{n(l)} \triangleq \arg\min_{S_{n(l)}} R_{emp}(\alpha), \qquad \alpha_0^{n(l)} \triangleq \arg\min_{S_{n(l)}} R(\alpha), \qquad \alpha_0 \triangleq \arg\min R(\alpha).$$

Then

$$R(\alpha_l^{n(l)}) \to R(\alpha_0) \quad \textit{if} \quad \frac{h_{n(l)} \ln l}{l} \to 0 \quad \textit{as} \quad n(l) \to \infty$$

with asymptotic rate of convergence

$$V(l) = r_{n(l)} + \sqrt{\frac{h_{n(l)} \ln l}{l}}$$

where

$$r_{n(l)} \triangleq \int |Q(z, \alpha_0^{n(l)}) - Q(z, \alpha_0)| dF(z).$$

The term $r_{n(l)}$ can be considered the "bias" or "approximation error." One can apply the vast literature on mathematical approximation theory here. The approximation error does not depend on the data and is purely a measure of how well one can approximate a function $Q(z, \alpha_0)$ with a function $Q(z, \alpha_0^{n(l)})$ in the class $S_{n(l)}$.

The term $\sqrt{(h_{n(l)} \ln(l))/l}$ can be considered the "variance" or "estimation error." The estimation error is a statistical term that depends on the complexity of the classes of functions.