Relation Between the Growth Function and the Shattering Coefficient

Institute of Mathematics and Computer Sciences – USP

December 2, 2019

Let $Q(z, \alpha)$ (z defined to be (x, y)) be the short notation for $Q(y, f_\alpha(x))$ where α is an element from the space of admissible functions Λ. Q is an arbitrary loss function, an example being the squared error loss:

$$Q(y, \hat{x}) = (y - \hat{x})^2.$$

Define the vector of loss outcomes on a given sample $z_1, \ldots, z_n$ as $q(\alpha, n) = (Q(z_1, \alpha), \ldots, Q(z_n, \alpha))$, which instead of being a function of α, n can be seen as a function of Y and $F_\alpha(X)$ as defined below:

$$\begin{aligned} q(\alpha, n) &= (Q(y_1, f_\alpha(x_1)), \ldots, Q(y_n, f_\alpha(x_n)) \\ &= q((y_1, \ldots, y_n), (f_\alpha(x_1), \ldots, f_\alpha(x_n))) \\ &= q(Y, F_\alpha(X)) \end{aligned}$$

where Y is the sequence of associated labels $y_1, \ldots, y_n$ and $F_\alpha(X)$ is the sequence of outputs $(f_\alpha(x_1), \ldots, f_\alpha(x_n))$ of the classifier f_α over the items $X = (x_1, \ldots, x_n)$.

Now let us define the total number of different ways the specified sample can be classified by all admissible functions

$$\Pi^\Lambda(X) = |\{F_\alpha(X) : \alpha \in \Lambda\}|$$

In order to calculate the total number of different loss outcomes, we need to find

$$\Sigma^\Lambda(Y, X) = |\{q(Y, F_\alpha(X)) : \alpha \in \Lambda\}| \tag{1}$$

but since Y and X are fixed, we know that

$$F_{\alpha_1}(X) = F_{\alpha_2}(X) \implies q(Y, F_{\alpha_1}(X)) = q(Y, F_{\alpha_2}(X)) \tag{2}$$

so to each element of $\Pi^\Lambda(X)$ can be associated one (possibly repeated) element of $\Sigma^\Lambda(Y, X)$. Note that the relation in (2) is one-sided because there can be equal loss outcomes for two unequal values of $F_\alpha(X)$ (in which case F is not injective). In other words, there exists a surjection from $\Pi^\Lambda(X)$ into $\Sigma^\Lambda(Y, X)$, which means (due to properties of cardinality from set theory) that $\Sigma^\Lambda(Y, X)$ is bounded as follows

$$\Sigma^\Lambda(Y, X) = |\{q(Y, F_\alpha(X)) : \alpha \in \Lambda\}| \leq |\{F_\alpha(X) : \alpha \in \Lambda\}| = \Pi^\Lambda(X)$$

where equality holds only if it is bijective. Finally, since the above holds for any fixed sample (X, Y) the growth function is bounded by

$$G^{\Lambda}(n) = \ln \max_{X,Y} \Sigma^{\Lambda}(X, Y) \leq \ln \max_{X,Y} \Pi^{\Lambda}(X)$$

but the term $\max_{X,Y} \Pi^{\Lambda}(X)$ is the maximum number of different classifications that the set of admissible functions can give to any sample, which was defined as the shattering coefficient $\mathcal{N}^{\Lambda}(n)$. Thus we have

$$G^{\Lambda}(n) = \ln \max_{X,Y} \Sigma^{\Lambda}(X, Y) \leq \ln \max_{X,Y} \mathcal{N}^{\Lambda}(n) = \max_{X,Y} \ln \mathcal{N}^{\Lambda}(n)$$

where the last equality holds due to the logarithm being monotonically increasing. Now we see the link between the works of Von Luxburg and Schölkopf (2011) and Vapnik (1998), because the latter proved that learning happens if and only if

$$\lim_{n \to \infty} \frac{G^{\Lambda}(n)}{n} \to 0 \tag{3}$$

whereas Von Luxburg and Schölkopf (2011) says that it occurs if

$$\lim_{n \to \infty} \frac{\ln \mathcal{N}^{\Lambda}(n)}{n} \to 0 \tag{4}$$

which is true. To see this, assume (4), then we have

$$0 \leq \frac{G^{\Lambda}(n)}{n} \leq \frac{\ln \mathcal{N}^{\Lambda}(n)}{n}$$

so by the squeeze theorem of calculus all terms go to zero when taking the limit $n \to \infty$.

References

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley and Sons.

Von Luxburg, U. and Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pages 651–706. Elsevier.