Overfitting: Bias-Variance Tradeoff and PAC Learnability

MinSeok Song

Overfitting is a central challenge in machine learning and statistical modeling. This phenomenon can be viewed from multiple perspectives.

Proposition 1. *For a fixed x, assume there exists a distribution for $f_k(x)$, representing a distribution over all training sets. Given that the data arises from the model $Y = f(x) + \epsilon$, where the expected value of ϵ is 0, we have:*

$$E(Y - f_k(x)^2) = \sigma^2 + Bias(f_k)^2 + Var(f_k(x))$$

Remark 1.
- σ is an irreducible error: this is the noise inherent in any real-world data collection process, which cannot be removed or reduced.
- The expected value is taken for a distribution over all training sets.

Proof. The detailed proof involves algebraic manipulations, available at: `https://stats.stackexchange.com/questions/204115/understanding-bias-variance-tradeoff-derivation/354284#354284` □

Proposition 2. *Given the Empirical Risk Minimization (ERM) chosen hypothesis $h_S = argmin_{h \in \mathcal{H}} L_S(h)$, the following holds:*

1. $\mathcal{D}^m \left(S \mid_x : L_{\mathcal{D},f}(h_S) > \epsilon\right) \leqslant \mathcal{D}^m \left(\bigcup_{h \in \mathcal{H}_B} \{S \mid_x : L_S(h) = 0\}\right)$
2. *With the assumption of I.I.D. data:*

$$\mathcal{D}^m \left(S \mid_x : L_{\mathcal{D},f}(h_S) > \epsilon\right) \leqslant |\mathcal{H}_\mathcal{B}| e^{-\epsilon m} \leqslant |\mathcal{H}| e^{-\epsilon m}$$

Remark 2.
- The first inequality illustrates that, given the realizability assumption, we obtain $L_S(h_S) = 0$. Since ERM operates on the set $\mathcal{H}_B = \{L_{\mathcal{D},f}(h_S) > \epsilon\}$, there exists some $h \in \mathcal{H}_B$ such that $L_S(h) = 0$.
- The second inequality employs the inequality $1 - \epsilon \leqslant e^{-\epsilon}$, which is tight for smaller ϵ, and has an analytic advantage by using exponential.
- The cardinality of $\mathcal{H}$ is used instead of $\mathcal{H}_B$ because we do not know the size of $\mathcal{H}_\mathcal{B}$ a priori.

Corollary 3. *For a finite hypothesis class $\mathcal{H}$, if $\delta \in (0,1)$, $\epsilon > 0$, and $m \geqslant \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$, then:*

$$L_{\mathcal{D},f}(h_S) \leqslant \epsilon$$

with probability at least $1-\delta$ over an i.i.d. sample S of size m, given the realizability assumption.

Remark 3.

- A smaller ϵ or δ necessitates a larger m (ϵ has a stronger effect), which makes sense.
- A larger hypothesis class also increases the value of m, demonstrating the problem of overfitting. This can be traced back to the necessity for $L_S(h)$ to hold for every possible S. However, it's crucial to note that this represents a worst-case scenario and might not be tight.

Example 4. Let h be defined as:

$$h_{(a_1,b_1,a_2,b_2)}(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \leqslant x_1 \leqslant b_1 \text{ and } a_2 \leqslant x_2 \leqslant b_2 \\ 0 & \text{otherwise} \end{cases}$$

Consider the hypothesis class consisting of all axis-aligned rectangles in the plane. When presented with positive samples in this plane, the ERM corresponds to the rectangle that encompasses all these positive samples. Smallest such is our reason hypothesis of interest.