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.
- Note that the finiteness of $|\mathcal{H}|$ is crucial here. We have PAC learnability for a finite VC dimension in genral, though this is only a sufficient condition.

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. An algorithm that returns the smallest rectangle enclosing all positive examples in the training set is an ERM. The realizability assumption is crucial here in order to guarantee that 0-labeled training sets are located outside of this rectangle. We can show that for $m \geqslant \frac{4\log(4/\delta)}{\epsilon}$, the PAC condition holds.

- In order to prove this, we cannot directly use the above corollary though, since our hypothesis class is not finite; instead, we need some geometric tricks (consider $R(S) \subset R^*$ and $R_i, i = 1,2,3,4$ with $D(R_i) = \epsilon/4$) along with exponential bound (for interpretable formula) and realizability assumption.
- This is guaranteed by the fundamental theorem of statistical learning; VC dimension is $4 < \infty$.
- We can generalize to d-dimensional space: we simply replace 4 with 2d.
- For each dimension we find minimum and maximum value to construct the smallest enclosing rectangle, so the algorithm to return such a rectangle that ensures the above accuracy/probability from above takes $O(md) = O(d^2 \cdot \frac{1}{\epsilon} \cdot \log(1/\delta))$

Generalization

We propose three key ideas to further generalize the learning paradigm:

1. **Joint Distribution Assumption:** Instead of treating the input x in isolation with $x \sim \mathcal{D}$, both the input x and the output y are assumed to be drawn from a joint distribution, denoted as $(x, y) \sim \mathcal{D}$.

2. **Agnostic Learnability:** Moving beyond the traditional realizability assumption, we introduce the concept of *agnostic learnability*. The criterion for learnability in this context is:

$$L_{\mathcal{D}}(h) \leqslant \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

   Here, $L_{\mathcal{D}}(h)$ represents the expected loss of hypothesis h under distribution $\mathcal{D}$.

3. **Generalized Loss Function:** The loss function for our hypothesis can be extended as:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[l(h, z)]$$

   The loss function $l(h, z)$ in prior discussions was defined as a binary function: $l(h, z) = 1$ if $h(x) \neq y$ and 0 otherwise. This generalized approach permits a broader range of definitions. The empirical loss for a sample set S is:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$$

   As an exemplar, the quadratic loss function can be expressed as:

$$l(h, (x, y)) = (h(x) - y)^2$$

These ideas enable a more flexible framework, suitable for addressing a diverse set of problems and loss functions in machine learning.

$\diamond$ To further elaborate on the idea of 'relaxation' we can introduce the concept of uniform convergence to get a sufficient condition for agnostic learnability.

Proposition 4. *If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leqslant m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, in that case, the $ERM_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*