OVERFITTING: BIAS-COMPLEXITY 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:*

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

Remark 1.

- Note that $\hat{f}$ follows the distribution of X since it's given by $x_1, \ldots, x_m$ for example. How $\hat{f}$ is independent of the new data is also self-explanatory, which is why the nice cancellation happens in the proof.
- 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.
- The mean square error loss here is usually used in regression problems; in general, we may use different loss functions and consequently different decompositions. This decomposition specifically focuses on the bias and variance in the context of mean squared error, but the principle applies more broadly to the trade-off between model flexibility and overfitting.
- This is NOT the formulation of the trade-off per se. As we vary the feature space, the value of $E((Y - f_k(x))^2)$ also changes, typically showing a U-shape pattern. The U-shape represents the trade-off between bias and variance: as the complexity of the model increases (moving from left to right on the U-shape), bias typically decreases and variance increases.

Proof. Trick is to reverse engineer, and the remainder term will cancel.

$$\begin{aligned}
E_{X\sim\mathcal{D}}\Big((Y - \hat{f}(x_0))^2 | x = x_0\Big) &= E_{X\sim\mathcal{D}}\Big((Y - f(x_0) + f(x_0) - \hat{f}(x_0))^2 | X = x_0\Big)\\
&= E_{X\sim\mathcal{D}}\Big((Y - f(x_0))^2 + (f(x_0) - \hat{f}(x_0))^2 | X = x_0\Big)\\
&= E_{X\sim\mathcal{D}}\Big((Y - f(x_0))^2 + (f(x_0) - E(\hat{f}(x_0)))^2 + (E(\hat{f}(x_0)) - \hat{f}(x_0))^2 | X = x_0\Big)\\
&= \sigma^2 + \text{Bias}(f_k(x))^2 + \text{Var}(f_k(x))
\end{aligned}$$

□

- As for the general formulation of the bias-variance trade-off (in the context of learning theory), we have $L_{\mathcal{D}}(h_S) = \epsilon_{app} + \epsilon_{est}$ where $\epsilon_{app} = \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ and $\epsilon_{est} = L_{\mathcal{D}}(h_S) - \epsilon_{app}$. In this decomposition:
  - **Bias (ϵ_{app})**: Measures the error due to the assumptions made by the model. It reflects how far off our model's predictions can be from the truth, even if we had infinite data.
  - **Variance (ϵ_{est})**: Captures the model's sensitivity to the specificities of the training set. It measures how much our predictions would vary if we re-trained the model on different training data.
- We can connect bias with approximation error (in the sense that it's more about our model selection, anlytical limitation) while variance with estimation error.

- More modern discussions involve the tradeoff between accuracy and interpretability. This topic is described in detail in this paper: `https://arxiv.org/pdf/2010.13764.pdf`

Definition 1. *We say that a hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with PAC property.*

Proposition 2. *Given the Empirical Risk Minimization (ERM) chosen hypothesis*

$$h_S = argmin_{h \in \mathcal{H}} L_S(h)$$

the following holds:

(1) $\mathcal{D}^m \left(S\,|_x \colon L_{\mathcal{D},f}(h_S) > \epsilon\right) \leqslant \mathcal{D}^m \left(\bigcup_{h \in \mathcal{H}_B} \{S\,|_x \colon L_S(h) = 0\}\right)$

(2) With the assumption of i.i.d.:

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

Remark 2.

- The first inequality is the consequence of the assumption that $L_S(h_S) = 0$ for $h_S \in \mathcal{H}_{\mathcal{B}}$. The book is a bit misleading about its assumption on "realizability." It actually means that the set in the left hand side is a "bad training set" in the sense that there exists $h \in \mathcal{H}_{\mathcal{B}}$ with $R_S(h) = 0$.
- 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.

$*$ It is important to distinguish the following two contexts of "realizability."

(1) There exists h such that $L_D(h) = 0$.
(2) For the set with its ERM h_S having non-zero error, it will always contain a hypothesis with zero empirical error.

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.
- As a quick reminder, notice the similarity to JL lemma, where we had $k \geqslant \Omega(\frac{\log N/\delta}{\epsilon^2})$

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.

- To establish the PAC condition, we cannot directly apply the previous corollary due to our hypothesis class being infinite. Instead, the proof requires a blend of geometric insights and probabilistic bounds. Our approach involves considering $R(S) \subset R^*$ and the outer rectangles R_i for $i = 1, 2, 3, 4$, each with a measure $D(R_i) = \frac{\epsilon}{4}$. While employ the exponential bounds, which gives an interpretable formula, and adhering to the realizability assumption; we focus on computing the probability that the squares R_i do not encompass any data points from the distribution.
- 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 $O(md) = O(d^2 \cdot \frac{1}{\epsilon} \cdot \log(1/\delta))$

Example 5. Consider the ERM rule defined such that whenever we encounter a value of 1 in our sample dataset, all subsequent values are assigned 0. To establish PAC learnability, assume $P(x_+) > \epsilon$ and $x_+ \neq S$. The probability that the learned hypothesis h_S misclassifies is given by

$$P(L_D(h_S)) \leqslant (1 - \epsilon)^m$$

Following from this, the sample complexity needed for PAC learning can be bounded as:

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$$

Example 6. As for minimizing the risk, we note that for binary classification problem, Bayes classifier is optimal. To show this, we need to prove that

$$P(f(X) \neq Y | X = x) - P(t(X) \neq Y | X = x) \geqslant 0$$

where $t(x) = 1$ if $P(Y = 1 | X = x) \geqslant 1/2$ and -1 otherwise. We can do this by showing that

$$P(f(X) \neq | X = x) = (1 - 2P(Y = 1 | X = x))1_{f(x)=1} + P(Y = 1 | X = x)$$

and

$$P(t(x) \neq Y | X = x) = min(P(Y = 1 | X = x), P(Y = -1 | X = x))$$

Example 7. We can define learnability of two-oracle model as a generalization by taking $\mathcal{D}^+(A) = \mathcal{D}(A)/\mathcal{D}(\mathcal{X}^+)$. Learnability of one-model oracle model implies the learnability of two-model oracle model. The key step is to define $\mathcal{D}'(E) = \frac{1}{2}\mathcal{D}^+(E) + \frac{1}{2}\mathcal{D}^-(E)$.

One can also show that if $h^+, h^- \in \mathcal{H}$, then learnability of two-oracle model implies that of one-oracle model.

$*$ What's the issue of ERM? The performance of Empirical Risk Minimization is highly dependent on the training set, so there might well be hypothesis in our hypothesis class that does extremely well on untrained data, even though it does not achieve 0 empirical loss.

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 allows for 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 $\mathcal{H}$.*

Proposition 5. *Uniform convergence is sufficient and necessary condition for PAC learnability (but only sufficient for agnostic learnability).*

Theorem 6. *(No-Free-Lunch) Let A be a learning algorithm for binary classification for 0-1 loss. Let m be any number smaller than $|\mathcal{X}|/2$. There exists a distribution $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$ such that*

(1) We have $L_{\mathcal{D}}(f) = 0$ for some classifying function $f : \mathcal{X} \to \{0,1\}$.
(2) With probability of at least 1/7, it holds that, over the choice of $S \sim \mathcal{D}^m$ we have $L_{\mathcal{D}}(A(S)) \geqslant 1/8$ (uncertainty on S).

Remark 8.
- In order to prove this, it suffices to show that

$$\max_{i\in[T]} E_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geqslant 1/4$$

- The idea of the proof is to use uniform distribution $\mathcal{D}_i$ induced by f_i.
- Since we are saying that no particular algorithm always works, it makes sense to use uniformity.
- The algebraic manipulation in the proof used the idea of max> *average¿*min. The NFL theorem makes the profound statment that while some algorithms may shine brilliantly in some scenarios and fail miserably in others, on average, they all even out; so the algebra goes along with this philosophy.
- We can generalize this by substituting $\frac{1}{4}$ with $\frac{1}{2}-\frac{1}{2k}$ and $2m$ with km (using $p \geqslant (k-1)m$ instead of $p \geqslant m$, we have $\frac{1}{km} \geqslant (1-\frac{1}{k})\frac{1}{p}$).

Corollary 7. *Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$. Then $\mathcal{H}$ is not PAC learnable.*

Dealing with infinite Hypothesis Even in cases where the hypothesis class is of infinite size, such as with rectangles or concentric circles, PAC learnability can still be achieved. This leads us to introduce the notion of the VC dimension, which pertains specifically to a given hypothesis class.

$*$ It's worth noting that in many instances, the VC dimension equates to the number of parameters that define the hypothesis class.

NFL theorem implies that a class of infinite VC-dimension is not PAC learnable. In other words, there is no universal algorithm that solves every problem ("different distribution, think how linear/nonlinear data affects our choice of model") when VC dimension is infinite.

Theorem 8 (Sauer's Lemma)**.** *Let $\mathcal{H}$ be a hypothesis class with VC dimension $\leqslant d < \infty$. Then for all m, $\tau_{\mathcal{H}}(m) \leqslant \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d + 1$ then $\tau_{\mathcal{H}}(m) \leqslant (em/d)^d$, where $\tau_{\mathcal{H}}$ is a growth function.*

Proof. It suffices to show that for every $\mathcal{H}$, $|\mathcal{H}_{\mathcal{C}}| \leqslant |\{B \subset \mathcal{C} : \mathcal{H} \text{ shatters } B\}|$ □

Theorem 9 (The fundamental theorem of PAC learning)**.** *Consider hypothesis class $\mathcal{H} : \mathcal{X} \to \{0, 1\}$. TFAE*

(1) $\mathcal{H}$ has the uniform converrgence property.
(2) Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.
(3) $\mathcal{H}$ is agnostic PAC learnable.
(4) $\mathcal{H}$ is PAC learnable.
(5) Any ERM rule is a successful PAC learner for $\mathcal{H}$.
(6) $\mathcal{H}$ has a finite VC-dimension.