Bias-Complexity Tradeoff

Kontonis Vasilis

27 April, 2017

Corelab, NTUA

Contents

Introduction

The distribution on $\mathcal{X} \times \mathcal{Y}$

Let (X, Y) be a random pair taking values in $\mathcal{X} \times \{0, 1\}$.

- $\mu(A) = \mathbb{P}[X \in A]$
- $\eta(x) = \mathbb{P}[Y = 1 | X = x]$

Then the pair $(X, Y) \sim D$ is described by (μ, η).

Proof.
Write $C = C_0 \times \{0\} \bigcup C_1 \times \{1\}$, then

$$\begin{aligned} \mathbb{P}[(X, Y) \in C] &= \mathbb{P}[X \in C_0, Y = 0] + \mathbb{P}[X \in C_1, Y = 1] \\ &= \int_{C_0} (1 - \eta(x)) \mathrm{d}\mu + \int_{C_1} \eta(x) \mathrm{d}\mu \end{aligned}$$

□

Loss And Risk

- Loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$.
- True Risk
$$L_D(h) = \mathbf{E}_{z \sim D}[\ell(h, z)] = \int_Z \ell(h, z) \mathrm{d}D$$
- Empirical Risk
$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

.

Let D be known. Can you find a good hypothesis h^*?

Loss And Risk

- Loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$.
- True Risk
$$L_D(h) = \mathbf{E}_{z\sim D}[\ell(h,z)] = \int_Z \ell(h,z)\mathrm{d}D$$
- Empirical Risk
$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$$
.

Let D be known. Can you find a good hypothesis h^*?
Bayes Optimal Classifier

$$h^*(x) = \begin{cases} 1, & \text{if } \eta(x) > 1/2 \\ 0, & \text{otherwise} \end{cases}$$

Definition (Agnostic PAC Learnability)
A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$, if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution D over Z, when running the learning algorithm on $m \geqslant m_{\mathcal{H}}$ i.i.d examples generated by D, the algorithm returns $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta$

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

where $L_D(h) = \mathbf{E}_{z \sim D}[\ell(h, z)]$.

Uniform Convergence

Definition (Representative Sample)
A training set S is called ϵ-representative if

$$\forall h \in \mathcal{H},\ |L_S(h) - L_D(h)| \leqslant \epsilon$$

Definition (Uniform Convergence)
We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution D over Z, if S is a sample of $m \geqslant m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d according to D, then, with probability of at least $1 - \delta$, S is ϵ-representative.

Finite Hypothesis Classes

Theorem (PAC)
Every finite hypothesis class is PAC learnable with sample complexity

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

Theorem (APAC-UC)
Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0, 1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leqslant \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil$$

Bias-Complexity Tradeoff

No-Free-Lunch

Theorem (No-Free-Lunch)
Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let $m \leqslant |\mathcal{X}|/2$, represent a training set size. Then there exists a distribution D over $\mathcal{X} \times \{0,1\}$ such that:

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_D(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim D^m$ we have that $L_D(A(S)) \geqslant 1/8$.*

No-Free-Lunch

Proof. Let $C \subseteq \mathcal{X},\ |C| = 2m$.

- $T = 2^{2m}$ possible functions $f_1, \ldots, f_T,\ C \to \{0, 1\}$.
- For f_i define

$$D_i((x, y)) = \begin{cases} 1/|C|, & \text{if } y = f_i(x) \\ 0, & \text{otherwise} \end{cases}$$

It suffices to show that

$$\max_{i \in [T]} \mathbf{E}_{S \sim D_i^m}[L_{D_i}(A(S))] \geqslant 1/4$$

No-Free-Lunch

- Denote by $S_1, \dots S_k$, $k = (2m)^m$ the possible sequences of m examples from C.
- Let $S_j^i = ((x_1, f_i(x_1)), \dots, (x_m, f_i(x_m)))$.
- If the distribution is D_i then the possible training sets A can receive are $S_1^i, \dots, S_k^i$ which all have the same probability of being sampled. Therefore

$$\max_{i \in [T]} \mathbf{E}_{S \sim D_i^m}[L_{D_i}(A(S))] \geqslant \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$

$$\geqslant \min_{j \in [k]} \frac{1}{T} \sum_{j=1}^{T} L_{D_i}(A(S_j^i))$$

No-Free-Lunch

- Now, fix a $j \in [k]$. Denote $S_j = (x_1, \ldots, x_m)$ and let $v_i, \ldots, v_p$ be the examples in C that do not appear in S_j. It holds $p \geqslant m$. Therefore

$$
\begin{aligned}
L_{D_i}(h) &= \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[h(x) \neq f_i(x)]} \\
&\geqslant \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(u_r)]}.
\end{aligned}
$$

- Moreover,

$$
\begin{aligned}
\frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i)) &\geqslant \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}_{\left[A(S_j^i)(v_r) \neq f_i(v_r)\right]} \\
&\geqslant \frac{1}{2} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{\left[A(S_j^i)(v_r) \neq f_i(v_r)\right]}.
\end{aligned}
$$

No-Free-Lunch

Fix $r \in [p]$. Partition the $T = 2^{2m}$ functions $f_1, \ldots, f_T$ into $T/2$ disjoint pairs, such that for a pair (f_i, f_i') it holds

$$\forall c \in C, f_i(c) \neq f_{i'}(c) \iff c = v_r.$$

For these pairs it holds that $S_j^i = S_j^{i'}$ and therefore

$$\mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} + \mathbb{1}_{[A(S_j^{i'})(v_r) \neq f_{i'}(v_r)]} = 1$$

which yields

$$\frac{1}{T} \sum_{i=1}^{T} \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} = \frac{1}{2}$$

Error Decomposition

Let h_S be an $ERM_{\mathcal{H}}$ hypothesis. Then

$$L_D(h_S) = \epsilon_{app} + \epsilon_{est}$$

where: $\epsilon_{app} = \min_{h \in \mathcal{H}} L_D(h), \quad \epsilon_{est} = L_D(h_s) - \epsilon_{app}$.

- **Approximation Error**: The minimum risk achievable by a predictor in the hypothesis class.
  - Enlarging the hypothesis class **can** decrease the approximation error.
- **Estimation Error**: The difference between the approximation error the error achieved by the ERM predictor.
  - The estimation error results because the empirical risk is only an estimate of the true risk.
  - The estimation error depends on the **training set size**, and the **complexity** of the hypothesis class.

Bias-Variance Decomposition

- Training Set $((x_1, y_1), \ldots, (x_m, y_m)) \sim D^m$.
- Data come from a function with noise $y = f(x) + \epsilon$.
- $\mathbf{E}[\epsilon] = 0, \ \mathbf{V}[\epsilon] = \sigma^2$.
- $\mathrm{Bias}[\hat{f}] = \mathbf{E}[\hat{f} - f]$
- $\mathbf{V}[\hat{f}] = \mathbf{E}[\hat{f}^2] - \mathbf{E}[\hat{f}]^2$

The generalization error decomposes

$$\mathbf{E}[(y - \hat{f})^2] = \sigma^2 + \mathbf{V}[\hat{f}] + \mathrm{Bias}[\hat{f}]^2$$

Bias-Variance Decomposition

Proof.

$$\begin{aligned}
\mathbf{E}[(y-\hat{f})^2] &= \mathbf{E}[y^2+\hat{f}^2-2y\hat{f}] \\
&= \mathbf{E}[y^2]+\mathbf{E}[\hat{f}^2]-\mathbf{E}[2y\hat{f}] \\
&= \mathbf{V}[y]+\mathbf{E}[y]^2+\mathbf{V}[\hat{f}]+\mathbf{E}[\hat{f}]^2-2f\mathbf{E}[\hat{f}] \\
&= \mathbf{V}[y]+\mathbf{V}[\hat{f}]+(f^2-2f\mathbf{E}[\hat{f}]+\mathbf{E}[\hat{f}]^2) \\
&= \mathbf{V}[y]+\mathbf{V}[\hat{f}]+\mathbf{E}[f-\hat{f}]^2 \\
&= \sigma^2+\mathbf{V}[\hat{f}]+\mathrm{Bias}[\hat{f}]^2
\end{aligned}$$

□

- Very rich $\mathcal{H} \rightarrow$ small bias - overfitting - large estimation error.
- Very small $\mathcal{H} \rightarrow$ large bias - underfitting - large approximation error.

Questions?

References i

M. J. Kearns and U. Vazirani.
An Introduction to Computational Learning Theory.
The MIT Press, Cambridge, Mass, Aug. 1994.

M. Mohri, A. Rostamizadeh, and A. Talwalkar.
Foundations of Machine Learning.
The MIT Press, Cambridge, MA, Aug. 2012.

S. Shalev-Shwartz and S. Ben-David.
Understanding Machine Learning: From Theory to Algorithms.
Cambridge University Press, New York, NY, USA, 1 edition edition, May 2014.