泛化理论

第二章 一致收敛

§2.2.1 VC dimension

@ 滕佳烨

[ref] Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz and Shai Ben-David (2014)
[ref] High-Dimensional Probability: An Introduction with Applications in Data Science, Roman Vershynin (2020).

Recall:

Uniform convergence: Decouple the dependency via "sup" over function class.

$$L(\hat{f}) - \hat{L}(\hat{f}) \leq \sup_{f \in \mathcal{F}} \left| L(f) - \hat{L}(f) \right| := \mathrm{UC}(\mathcal{F}).$$

Today's topic:

VC dimension: measures the complexity of the function class $\mathcal{F}$.

Note that $\mathrm{UC}(\mathcal{F})$ is closely related to $\mathcal{F}$'s ***complexity***.
For example, if $\mathcal{F} \subset \mathcal{G}$ ($\mathcal{G}$ is more complex than $\mathcal{F}$), then $\mathrm{UC}(\mathcal{F}) \leq \mathrm{UC}(\mathcal{G})$.

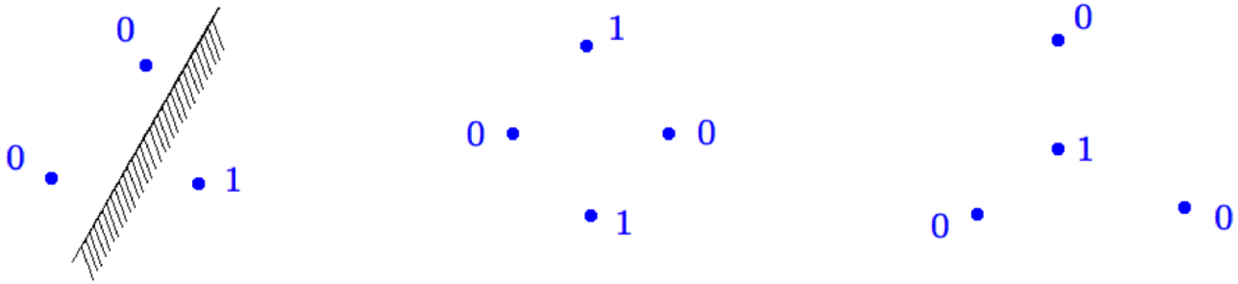
As we will show, VC dimension also measures the $\mathcal{F}$'s ***complexity.*** Therefore, it is natural to bound $\mathrm{UC}(\mathcal{F})$ using VC dimension…

VC dimension

Definition (VC dimension): Consider a class $\mathcal{F}$ of Boolean functions on some domain Ω. We say that a subset $\Lambda \subset \Omega$ is ***shattered*** by $\mathcal{F}$ if ***any*** function $g: \Lambda \to \{0, 1\}$ can be obtained by restricting some function $f \in \mathcal{F}$ onto Λ. The VC dimension of $\mathcal{F}$, denoted $vc(\mathcal{F})$, is the ***largest*** cardinality of a subset $\Lambda \subset \Omega$ shattered by $\mathcal{F}$.

Key point: how much points ***can*** the function class $\mathcal{F}$ ***completely*** fit?
can: there exist such a subset Λ
completely: for any labeling process (any g), there exist $f \in \mathcal{F}$ to fit it.

For example, a linear 2-d classifier class $\mathcal{F}_2$ has VC dim = 3 (see the following figure). When there are three points, there always exist a line to separate it (as long as the three points are not in a line).
When there are four points, $\mathcal{F}_2$ cannot fit it no matter how to set the four points.
Therefore, VC($\mathcal{F}_2$) = 3.



VC dimension and generalization (we will prove it in the later class)

Theorem 8.3.23 (Empirical processes via VC dimension). *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension* $\mathrm{vc}(\mathcal{F}) \geq 1$. *Let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω distributed according to the law μ. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq C \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}. \tag{8.29}$$

Connection to generalization:

Set $f(X_i)$ as the loss function $l(X_i, Y_i)$, e.g., 0-1 loss. We have that:

uniform convergence

$$\underbrace{\mathbb{E}L(\hat{f}) - \hat{L}(\hat{f})}_{\text{Generalization Gap}} \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left| L(f) - \hat{L}(f) \right| = \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \left(f(X_i) - \mathbb{E} f(X_i) \right) \right| \leq C \sqrt{\frac{vc(\mathcal{F})}{n}}.$$

f is loss

Take-away messages

(a) VC dimension and shattering: measuring the complexity of function class $\mathcal{F}$.

(b) how much points ***can*** the function class $\mathcal{F}$ ***completely*** fit?

Exist a pattern of points, $\mathcal{F}$ fit all the possible labels.

(a) VC dimension and generalization: $\sqrt{vc/n}$.

All the slides will be available at www.tengjiaye.com/generalization soon.

@ 滕佳烨

Thanks!