CS group meeting 02/10/2017

The Nature of Statistical Learning Theory V. Vapnik.

Lecture 2

Two sided empirical process

$$\xi^{\ell} = \sup_{\alpha \in \Lambda} \left| \int Q(z,\alpha)\, dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i,\alpha) \right|$$

Uniform two-sided convergence $\ell = 1, 2, \ldots$

$$\lim_{\ell \to \infty} \mathbb{P}\left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z,\alpha)\, dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i,\alpha) \right| > \varepsilon \right\} = 0$$

$\forall \varepsilon > 0$

One-sided empirical process

$$\xi_{+}^{\ell} = \sup_{\alpha \in \Lambda} \left( \int Q(z,\alpha)\, dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i,\alpha) \right)_{+}$$

Uniform one-sided convergence

$$\lim_{\ell \to \infty} \mathbb{P}\left( \sup_{\alpha \in \Lambda} \left( \int Q(\alpha)\, dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i,\alpha) \right) > \varepsilon \right) = 0$$

$\forall \varepsilon > 0$

"Law of Large Numbers"

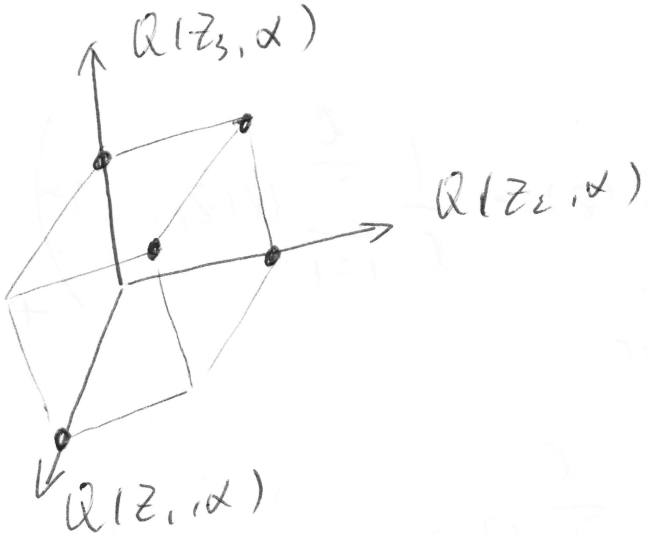
Entropy of the set of functions $Q(z,\alpha)$, $\alpha \in \Lambda$ on a sample of size $\ell$.

- The case of a set of indicator functions

If $Q(z,\alpha)$ $\alpha \in \Lambda$ is a set of indicator functions and $z_1, \ldots, z_\ell$ is a training sample

$$q(\alpha) = (Q(z_1,\alpha) \ldots . Q(z_\ell,\alpha)) \quad \alpha \in \Lambda$$

will be a binary vector that belongs to the vertices of an $\ell$-dimensional cube




$N^\wedge(z_1, \ldots, z_\ell)$ is the number of vertices that are different and can be obtained by the training sample $z_1 \ldots z_\ell$ and the set of functions $Q(z,\alpha)$, $\alpha \in \Lambda$

Random entropy $H^{\wedge}(z_1, \ldots, z_\ell) = \ln N^{\wedge}(z_1 \ldots z_\ell)$

Entropy of the set of indicator functions $Q(z, \alpha)$ $\alpha \in \Lambda$ on samples of size $\ell$

$$H^{\wedge}(\ell) = \mathbb{E} \ln N^{\wedge}(z_1 \ldots z_\ell).$$

- The general case and VC-entropy

Definition Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ be a set of bounded loss functions. Using this set of functions and the training set $z_1 \ldots z_\ell$ one can construct the following set of $\ell$-dimensional vectors

$$q(\alpha) = (Q(z_1, \alpha), \ldots, Q(z_\ell, \alpha)) \quad \alpha \in \Lambda$$

This set of vectors belongs to the $\ell$-dimensional cube and has a finite minimal $\varepsilon$-net in the metric $C^0$ (or $L_p$). Let $N = N^{\wedge}(\varepsilon; z_1 \ldots z_\ell)$ be the number of elements of the minimal $\varepsilon$-net of this set of vectors $q(\alpha)$, $\alpha \in \Lambda$.

Random VC entropy

$$H^{\wedge}(\varepsilon; z_1, \dots z_\ell) = \ln N^{\wedge}(\varepsilon; z_1, \dots z_\ell)$$

VC entropy

$$H^{\wedge}(\varepsilon; \ell) = \overline{\mathbb{E}}\, H^{\wedge}(\varepsilon; z_1, \dots z_\ell)$$

Conditions for uniform two-sided convergence

Theorem For uniform two-sided convergence it is necessary and sufficient that

$$\lim_{\ell \to \infty} \frac{H^{\wedge}(\varepsilon; \ell)}{\ell} = 0 \qquad \forall \varepsilon > 0$$

Conditions for uniform one-sided convergence

Theorem In order for uniform one-sided convergence of empirical means to their expectations to hold for the set of totally bounded functions $Q(z, \alpha)$, it is necessary and sufficient that for any positive $\delta$, $\eta$ and $\varepsilon$ there exists a set of functions $Q^*(z, \alpha^*)$, $\alpha^* \in \Lambda^*$ satisfying

$$Q(z, \alpha) - Q^*(z, \alpha^*) \geq 0 \quad \forall z$$

$$\int (Q(z, \alpha) - Q^*(z, \alpha^*))\, dF(z) \leq \delta$$

such that

$$\lim_{\ell \to \infty} \frac{H^{\wedge *}(\varepsilon, \ell)}{\ell} < \eta$$

holds for the $\varepsilon$-entropy of the set $Q^*(z, \alpha), \alpha \in \Lambda^*$ on samples of size $\ell$.

Question What of $\lim_{\ell \to \infty} \frac{H^{\wedge}(\varepsilon_0, \ell)}{\ell} \neq 0$ for some $\varepsilon_0$?

ERM does not hold $\Rightarrow$ non-consistent

But Why non-consistent?

Theory of Non-falsifiability in Philosophy

Kant's problem of Demarcation

(i). Deductive : General $\longrightarrow$ Particular

(ii). Inductive : Particular $\longrightarrow$ General

Kant's problem: What is the difference between the cases of a justified inductive step, and those for which the inductive step is not justified?

All Natural Science is a result of inductive inference.

So Kant's problem = Is there a formal way to distinguish true theories and false theories?

## Popper's solution

A necessary condition for justifiability of a theory is the feasibility of its falsification.

falsification = existence of a collection of particular assertions which cannot be explained by the given theory although they fall into its domain

If a given theory can be falsified → scientific
If a given theory cannot be falsified → non-scientific

▷ Complete Nonfalsifiability

We know for indicator functions

$$H^{\wedge}(\ell) = \mathbb{E} \ln N^{\wedge}(z_1, \dots z_\ell), \quad N^{\wedge}(z_1, \dots z_\ell) \leq 2^{\ell}$$

If $\lim_{\ell \to \infty} \frac{H^{\wedge}(\ell)}{\ell} = \ln 2$ (maximal entropy)

Then for almost all samples $z_1, \dots z_\ell$ we have

$$N^{\wedge}(z_1, \dots z_\ell) = 2^{\ell}$$

That is, almost all sample $z_1, \dots z_\ell$ of arbitrary size $\ell$ can be separated in all possible ways by functions

of this set, so minimum of empirical risk for this machine is 0

Such a machine can give a general explanation for almost any data.

Non-falsifiable machine ⟶ Not scientific

Minimal value of the empirical risk is 0 independent of the expected risk,

That is why we need variance and bias!

▷ Partial Non-falsifiability

Theorem For the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$ let the convergence

$$\lim_{l \to \infty} \frac{H^{\wedge}(l)}{l} = c > 0$$

be valid. Then there exists a subset $Z^*$ of the set $Z$ for which $\mathbb{P}(Z^*) = a(c) \neq 0$ and for almost any training set $z_1 \dots z_l$ we have $z_1^*, \dots, z_k^* = \{z_1 \dots z_l\} \cap Z^*$ and any sequence of binary values $\delta_1 \dots \delta_k$ $\delta_i \in \{0, 1\}$

there exists a function $Q(z, \alpha^*)$ for which

$$\delta_i = Q(z_i^*, \alpha^*) \quad i = 1, 2, \ldots, k$$

holds true.

So if the conditions of uniform two-sided convergence fail, then the exists some subspace of the input space where the learning machine is non-falsifiable

▷ There is also a more soplisticated "potentially nonfalsifiable" related to general VC entropy that I will skip here.

Three milestones in learning theory.

Consider the case of indicator functions $Q(z, \alpha)$

VC entropy $\quad H^\wedge(\ell) = \overline{E} \ln N^\wedge(z_1 \ldots z_\ell)$

Annealed VC entropy $\quad H^\wedge_{ann}(\ell) = \ln \overline{E} N^\wedge(z_1 \ldots z_\ell)$

Growth function $\quad G^\wedge(\ell) = \ln \sup_{z_1 \ldots z_\ell} N^\wedge(z_1 \ldots z_\ell)$

We have $\quad H^\wedge(\ell) \leq H^\wedge_{ann}(\ell) \leq G^\wedge(\ell)$

First milestone in learning theory

$$\lim_{\ell \to \infty} \frac{H^{\wedge}(\ell)}{\ell} = 0 \Longrightarrow$$ Consistency of the ERM principle

Second milestone in learning theory

$$\lim_{\ell \to \infty} \frac{H^{\wedge}_{ann}(\ell)}{\ell} = 0 \Longrightarrow$$ fast convergence of $R(\alpha_\ell)$ towards $R(\alpha_0)$

$$\mathbb{P}\left( R(\alpha_\ell) - R(\alpha_0) > \varepsilon \right) < e^{-c\varepsilon^2 \ell} \quad (c > 0)$$

Third milestone in learning theory

$$\lim_{\ell \to \infty} \frac{G^{\wedge}(\ell)}{\ell} = 0 \Longrightarrow$$ fast convergence of $R(\alpha_\ell)$ towards $R(\alpha_0)$ independent of the choice of the probability measure