CS group meeting  02/10/2017

The Nature of Statistical Learning Theory  V. Vapnik

Lecture 2

Two sided empirical process

\[ \xi^l = \sup_{z \in \Lambda} \left| \int Q(z, x) dF(z) - \frac{1}{l} \sum_{i=1}^{l} Q(z_i, x) \right| \]

Uniform two-sided convergence  \( l = 1, 2, \ldots \)

\[ \lim_{l \to \infty} P \left( \sup_{z \in \Lambda} \left| \int Q(z, x) dF(z) - \frac{1}{l} \sum_{i=1}^{l} Q(z_i, x) \right| > \varepsilon \right) = 0 \quad \forall \varepsilon > 0 \]

One-sided empirical process

\[ \xi_{\pm} = \sup_{z \in \Lambda} \left( \int Q(z, x) dF(z) - \frac{1}{l} \sum_{i=1}^{l} Q(z_i, x) \right) \]

Uniform one-sided convergence

\[ \lim_{l \to \infty} P \left( \sup_{z \in \Lambda} \left( \int Q(z, x) dF(z) - \frac{1}{l} \sum_{i=1}^{l} Q(z_i, x) \right) > \varepsilon \right) = 0 \quad \forall \varepsilon > 0 \]
"Law of Large Numbers"

Entropy of the set of functions \( \mathcal{Q}(z, x) \), \( x \in \Lambda \) on a sample of size \( d \).

- The case of a set of indicator functions

If \( \mathcal{Q}(z, x) \) \( x \in \Lambda \) is a set of indicator functions and \( z_1, \ldots, z_d \) is a training sample

\[
q(x) = (\mathcal{Q}(z_1, x), \ldots, \mathcal{Q}(z_d, x)) \times \Lambda
\]

will be a binary vector that belongs to the vertices of an \( d \)-dimensional cube

\[
\mathcal{N}^d(z_1, \ldots, z_d)
\]

is the number of vertices that are different and can be obtained by the training sample \( z_1, \ldots, z_d \) and the set of functions \( \mathcal{Q}(z, x) \), \( x \in \Lambda \)
Random entropy \( H^\wedge(\epsilon, \ldots, \epsilon) = \ln N^\wedge(\epsilon, \ldots, \epsilon) \)

Entropy of the set of indicator functions \( \chi(\epsilon, \alpha) \)
\( \alpha \in \Lambda \) on samples of size \( \epsilon \)

\( H^\wedge(\epsilon) = \mathbb{E} \ln N^\wedge(\epsilon, \ldots, \epsilon) \).

* The general case and VC-entropy

**Definition** Let \( A \leq \chi(\epsilon, \alpha) \leq B, \alpha \in \Lambda \) be a set of bounded loss functions. Using this set of functions and the training set \( \epsilon, \ldots, \epsilon \) one can construct the following set of \( d \)-dimensional vectors

\( q(\alpha) = (\chi(\epsilon, \alpha), \ldots, \chi(\epsilon, \alpha)) \alpha \in \Lambda \)

This set of vectors belongs to the \( d \)-dimensional cube and has a finite minimal \( \epsilon \)-net in the metric \( C^0 \) (or \( L_p \)). Let \( N = N^\wedge(\epsilon; \epsilon, \ldots, \epsilon) \) be the number of elements of the minimal \( \epsilon \)-net of this set of vectors \( q(\alpha), \alpha \in \Lambda \).
Random VC entropy

\[ H^\wedge(\varepsilon; z, \ldots, z_\varepsilon) = \ln N^\wedge(\varepsilon; z, \ldots, z_\varepsilon) \]

VC entropy

\[ H^\wedge(\varepsilon; \ell) = \mathbb{E} H^\wedge(\varepsilon; z, \ldots, z_\varepsilon) \]

Conditions for uniform two-sided convergence

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

\[ \lim_{\ell \to \infty} \frac{H^\wedge(\varepsilon; \ell)}{\ell} = 0 \quad \forall \varepsilon > 0 \]

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 \( \Omega(\varepsilon, x) \), it is necessary and sufficient that for any positive \( S, \eta \) and \( \varepsilon \) there exists a set of functions \( \Omega^\wedge(\varepsilon, x^\wedge) \), \( x^\wedge \in \Lambda^\wedge \) satisfying

\[ \Omega(\varepsilon, x) - \Omega^\wedge(\varepsilon, x^\wedge) \geq 0 \quad \forall \varepsilon \]

\[ \int (\Omega(\varepsilon, x) - \Omega^\wedge(\varepsilon, x^\wedge)) dF(x) \leq S \]
such that
\[
\lim_{l \to \infty} \frac{H^\ast(\varepsilon, l)}{l} < \eta
\]
holds for the \( \varepsilon \)-entropy of the set \( \omega^\ast(\varepsilon, x), x \in \lambda^\ast \)
on samples of size \( l \).

**Question**

What if \( \lim_{l \to \infty} \frac{H^\ast(\varepsilon_0, l)}{l} \neq 0 \) 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 Remarcation

(i). Deductive: General \( \Rightarrow \) Particular

(ii). Inductive: Particular \( \Rightarrow \) 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?
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 $\rightarrow$ scientific
- If a given theory cannot be falsified $\rightarrow$ non-scientific

**Complete Non-falsifiability**

We know for indicator functions

$$H^l(l) = E \ln N^l(z_1, \ldots, z_e), \quad N^l(z_1, \ldots, z_e) \leq 2^l$$

If

$$\lim_{l \to \infty} \frac{H^l(l)}{l} = \ln 2 \quad \text{(maximal entropy)}$$

Then for almost all samples $z_1, \ldots, z_e$ we have

$$N^l(z_1, \ldots, z_e) = 2^l$$

That is, almost all sample $z_1, \ldots, z_e$ of arbitrary size $l$ 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 \(\implies\) Not scientific

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

That is why we need variance and bias!

\[ \text{Partial Non-falsifiability} \]

\[ \text{Theorem} \quad \text{For the set of indicator functions} \]

\[ d(x, y), x \in \Lambda \text{ let the convergence} \]

\[ \lim_{k \to \infty} \frac{1}{H^k} = C > 0 \]

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 \( z_1, \ldots, z_e \)

we have \( z_1^*, \ldots, z_k^* = \{z, \ldots, z_e\} \cap z^* \)

and any sequence of binary values \( s_1, \ldots, s_k \quad s_i \in \{0, 1\} \)
there exists a function \( Q(\varepsilon, x^*) \) for which

\[ S_i = Q(\varepsilon_i^*, x^*) \quad i = 1, 2, \ldots, k \]

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 sophisticated "potentially non-falsifiable" related to general VC entropy that I will skip here.

Three milestones in learning theory.

Consider the case of indicator functions \( Q(\varepsilon, x) \)

**VC entropy**

\[ H^\wedge(l) = \mathbb{E} \ln N^\wedge(\varepsilon_1, \ldots, \varepsilon_l) \]

**Annealed VC entropy**

\[ H_{\text{ann}}^\wedge(l) = \ln \mathbb{E} N^\wedge(\varepsilon_1, \ldots, \varepsilon_l) \]

**Growth function**

\[ G^\wedge(l) = \ln \sup_{\varepsilon_1, \ldots, \varepsilon_l} N^\wedge(\varepsilon_1, \ldots, \varepsilon_l) \]

We have

\[ H^\wedge(l) \leq H_{\text{ann}}^\wedge(l) \leq G^\wedge(l) \]
First milestone in learning theory
\[
\lim_{d \to \infty} \frac{H^d(l)}{l} = 0 \implies \text{Consistency of the ERM principle}
\]

Second milestone in learning theory
\[
\lim_{d \to \infty} \frac{H_{\text{ann}}^d(l)}{l} = 0 \implies \text{fast convergence of } R(x_l) \text{ towards } R(x_0)
\]

\[
P \left( R(x_l) - R(x_0) > \varepsilon \right) \leq e^{-c\varepsilon^2 l} \quad (c > 0)
\]

Third milestone in learning theory
\[
\lim_{d \to \infty} \frac{G^d(l)}{l} = 0 \implies \text{fast convergence of } R(x_l) \text{ towards } R(x_0) \text{ independent of the choice of the probability measure.}
\]