Statistical Learning Theory

Consistency and bounds on the rate of convergence for ERM methods

Miguel A. Veganzones

Grupo Inteligencia Computacional
Universidad del País Vasco

Outline

Consistency of learning processes

- Consistency: convergence in probability to the best possible result.
- Consistency of learning processes:
  - To explain when a learning machine that minimizes empirical risk can achive a small value of actual risk (to generalize) and when it can not.
  - Equivalently, to describe necessary and sufficient conditions for the consistency of learning processes that minimize the empirical risk.
- This guarantees that the constructed theory is general and cannot be improved from the conceptual point of view.

Theory of non-falsiability

- Kant's problem of demarcation (s. XVIII): is there a formal way to distinguish true theories from false theories?
    - One of the main questions of modern philosophy.
- Popper's theory of non-falsiability (s. XX): criterion for demarcation between true and false theories.
- Strongly related to what happens if the ERM method is not consistent.

Bounds on the rate of convergence

- It is required for any machine minimizing empirical risk to satisfy consistency conditions.
- But, consistency conditions say nothing about the rate of convergence of the obtained risk $R(\alpha_l)$ to the minimal one $R(\alpha_0)$.
- It is possible to construct examples where the ERM principle is consistent, but where the risks have an arbitrary slow asymptotic rate of convergence.
- The theory of bounds on the rate of convergence tries to answer the following question:
    - Under what conditions is the asymptotic rate of convergence fast?

Outline

Notation

- Let $Q(\mathbf{z}, \alpha_l)$ be a function that minimizes the empirical risk functional

$$R_{emp} = \frac{1}{l} \sum_{i=1}^{l} Q(\mathbf{z_i}, \alpha)$$

for a given set of i.i.d. observations $\mathbf{z_1}, \ldots, \mathbf{z_l}$.

Classical definition of consistency

- The ERM principle is consistent for the set of functions $Q(\mathbf{z},\alpha), \alpha \in \Lambda$, and for the p.d.f. $F(\mathbf{z})$ if the following two sequences converge in probability to the same limit:

$$R(\alpha_l) \xrightarrow[l \to \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha) \quad (1)$$

$$R_{emp}(\alpha_l) \xrightarrow[l \to \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha) \quad (2)$$

- Equation (1) asserts that the values of achieved risks converge to the best possible.
- Equation (2) asserts that one can estimate on the basis of the values of empirical risk the minimal possible value of the risk.

Classical definition of consistency

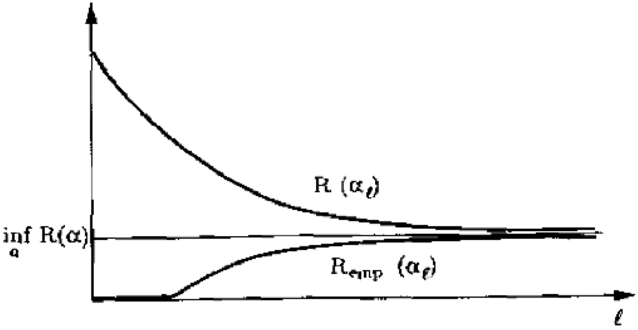


Figure: The learning process is consistent if both the expected risks $R(\alpha_l)$ and the empirical risks $R_{emp}(\alpha_l)$ converge to the minimal possible value of the risk $\inf_{\alpha\in\Lambda} R(\alpha)$.

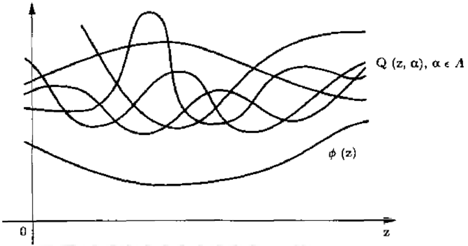
Goal

- To obtain conditions of consistency for the ERM method in terms of general characteristics of the set of functions and the probability measure.
- This is an impossible task because the classical definition of consistency includes cases of *trivial consistency*.

Trivial consistency

- Suppose that for some set of functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, the ERM method is not consistent.
- Consider an extended set of functions including this set of functions and the additinal function $\phi(\mathbf{z})$ that satisfies the following inequality

$$\inf_{\alpha\in\Lambda} Q(\mathbf{z},\alpha) > \phi(\mathbf{z}), \qquad \forall \mathbf{z}$$



Trivial consistency

- For the extended set of functions (containing $\phi(\mathbf{z})$) the ERM method will be consistent.
- For any distribution function and number of observations, the minimum of the empirical risk will be attained on the function $\phi(\mathbf{z})$ that also gives the minimum of the expected risk.
- This example shows that there exist trivial cases of consistency that depend on wether the given set of functions contains a minorizing function.

ERM consistency

- In order to create a theory of consistency of the ERM method depending only on the general properties (capacity) of the set of functions, a consistency definition excluding trivial consistency cases is needed.
- This is done by non-trivial (strict) consistency definition.

Non-trivial consistency

- The ERM principle is nontrivially consistent for the set of functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, and the probability distribution function $F(\mathbf{z})$ if for any nonempty subset $\Lambda(c)$, $c \in (-\infty,\infty)$ defined as

$$\Lambda(c) = \left\{ \alpha : \int Q(\mathbf{z},\alpha)\, dF(\mathbf{z}) > c, \quad \alpha \in \Lambda \right\}$$

the convergence

$$\inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \xrightarrow[l \to \infty]{P} \inf_{\alpha \in \Lambda(c)} R(\alpha) \tag{3}$$

is valid.

Key theorem of learning theory

- Vapnik and Chervonenkins, 1989.

Theorem

Let $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, be a set of functions that satisfy the condition

$$A \leq \int Q(\mathbf{z},\alpha)\, dF(\mathbf{z}) \leq B \quad (A \leq R(\alpha) \leq B)$$

then for the ERM principle to be consistent, it is necessary and sufficient that the empirical risk $R_{emp}(\alpha)$ converges uniformly to the actual risk $R(\alpha)$ over the set $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, in the following sense:

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left(R(\alpha)-R_{emp}(\alpha)\right) > \varepsilon\right\} = 0, \quad \forall\varepsilon > 0 \qquad (4)$$

Consistency of the ERM principle

- According to the key theorem, the uniform one-sided convergence (4) is a necessary and sufficient condition for (non-trivial) consistency of the ERM method.
- Conceptually, the conditions for consistency of the ERM principle are necessarily and sufficiently determined by the "worst" function of the set of functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$.

Outline

Introduction

- The key theorem expresses that consistency of the ERM principle is equivalent to existence of uniform one-sided convergence.
- Conditions for uniform two-sided convergence play an important role in constructing conditions for uniform two-sided convergence.
- Necessary and sufffficient conditions for both uniform one-sided and two-sided convergence are obtained on the basis of the VC entropy concept.

Empirical process

- An empirical process is an stochastic process in the form of a sequence of random variables

$$\xi^l = \sup_{\alpha\in\Lambda}\left|\int Q(\mathbf{z},\alpha)\,dF(\mathbf{z}) - \frac{1}{l}\sum_{i=1}^{l} Q(\mathbf{z_i},\alpha)\right|, \quad l = 1,2,\dots \quad (5)$$

  that depend on both, the probability measure $F(\mathbf{z})$ and the set of functions $Q(\mathbf{z},\alpha)$, $\alpha\in\Lambda$.
- The problem is to describe conditions under which this empirical process converges in probability to zero.

Consistency of an empirical process

- The necessary and sufficient conditions for an empirical process to converge in probability to zero imply that the equality

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(\mathbf{z},\alpha)\,dF(\mathbf{z}) - \frac{1}{l}\sum_{i=1}^{l} Q(\mathbf{z_i},\alpha)\right| > \varepsilon\right\} = 0, \quad \forall\varepsilon > 0 \tag{6}$$

holds true.

Law of large numbers and its generalization

- If the set of functions contains only one element, then the sequence of random variables ξ^l always converges in probability to zero: law of large numbers.
- Generalization of the law of large numbers for the case where a set of functions has a finite number of elements:

Definition

The sequence of random variables ξ^l converges in probability to zero if the set of functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, contains a finite number N of elements.

Law of large numbers and its generalization

- When $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, has an infinite number of elements, the sequence of random variables ξ^l does not necessarily converges in probability to zero.
- Problem of the existence of a law of large numbers in functional space (uniform two-sided convergence of the means to their probabilities): generalization of the classical law of large numbers.

VC Entropy

- Necessary and sufficient conditions for both uniform one-sided convergence and uniform two-sided convergence are obtained on the basis of a concept called *the VC entropy of a set of functions* $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, for a sample of size l.

VC Entropy of the set of indicator functions

Diversity

- Lets characterize the *diversity* of a set of indicator functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, on the given set of data by the quantity $N^{\wedge}(\mathbf{z_1},\ldots,\mathbf{z_l})$ that evaluates how many different separations of the given sample can be clone using functions from the set of indicator functions.
- Consider the set of l-dimensional binary vectors:

$$q(\alpha) = (Q(\mathbf{z_1},\alpha),\ldots,Q(\mathbf{z_l},\alpha)), \quad \alpha \in \Lambda$$

Geometrically, the diversity is the number of different vertices of the l-dimensional cube that can be obtained on the basis of the sample $\mathbf{z_1},\ldots,\mathbf{z_l}$ and the set of functions.

VC Entropy of the set of indicator functions

Diversity (geometrics)

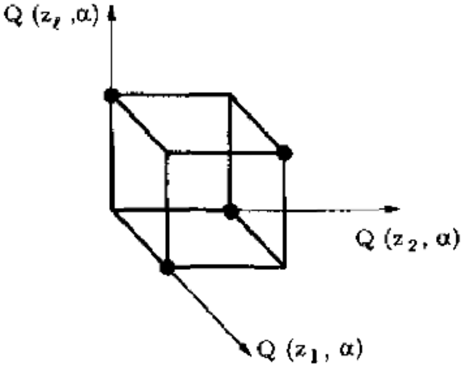


Figure: The set of l-dimensional binary vectors $q(\alpha)$, $\alpha \in \Lambda$, is a subset of the set of vertices of the l-dimensional unit cube.

VC Entropy of the set of indicator functions

Random entropy and VC entropy

- The random entropy

$$H^{\wedge}(\mathbf{z_1},\ldots,\mathbf{z_l}) = \ln N^{\wedge}(\mathbf{z_1},\ldots,\mathbf{z_l})$$

  describes the diversity of the set of functions on the given data.
- The expectation of the random entropy over the joint distribution function $F(\mathbf{z_1},\ldots,\mathbf{z_l})$:

$$H^{\wedge}(l) = E\left[\ln N^{\wedge}(\mathbf{z_1},\ldots,\mathbf{z_l})\right] \tag{7}$$

  is the *VC entropy* of the set or indicator functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, on samples of size l.

VC Entropy of the set of real functions

Diversity

- Let $A \leq Q(\mathbf{z}, \alpha) \leq B$, $\alpha \in \Lambda$, a set of bounded loss functions.
- Considering this set of functions and the training set $\mathbf{z_1}, \ldots, \mathbf{z_l}$ one can construct the following set of l-dimensional vectors:

$$q(\alpha) = (Q(\mathbf{z_1}, \alpha), \ldots, Q(\mathbf{z_l}, \alpha)), \quad \alpha \in \Lambda$$

- The diversity, $N = N^{\wedge}(\varepsilon, \mathbf{z_1}, \ldots, \mathbf{z_l})$, indicates the number of elements of the minimal ε-net of this set of vectors $q(\alpha)$, $\alpha \in \Lambda$.

VC Entropy of the set of real functions

Minimal ε-net

- The set of vectors $q(\alpha)$, $\alpha \in \Lambda$, has a minimal ε-net $q(\alpha_1), \ldots, q(\alpha_N)$ if:
  1. There exist $N = N^{\wedge}(\varepsilon, \mathbf{z_1}, \ldots, \mathbf{z_l})$ vectors $q(\alpha_1), \ldots, q(\alpha_N)$ such that for any vector $q(\alpha^*)$, $\alpha^* \in \Lambda$, one can find among these N vectors one $q(\alpha_r)$ that is ε-close to $q(\alpha^*)$ in a given metric.
  2. N is the minimum number of vectors that posseses this property.

VC Entropy of the set of real functions

Diversity (geometrics)

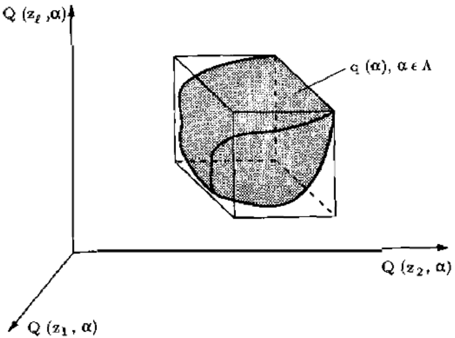


Figure: The set of l-dimensional vectors $q(\alpha)$, $\alpha \in \Lambda$, belongs to an l-dimensional cube.

VC Entropy of the set of real functions

Random entropy and VC entropy

- The random VC entropy of the set of functions $A \leq Q(\mathbf{z}, \alpha) \leq B$, $\alpha \in \Lambda$, on the sample $\mathbf{z_1}, \ldots, \mathbf{z_l}$ is given by:

$$H^{\wedge}(\varepsilon; \mathbf{z_1}, \ldots, \mathbf{z_l}) = \ln N^{\wedge}(\varepsilon; \mathbf{z_1}, \ldots, \mathbf{z_l})$$

- The expectation of the random VC entropy over the joint distribution function $F(\mathbf{z_1}, \ldots, \mathbf{z_l})$:

$$H^{\wedge}(\varepsilon; l) = E\left[\ln N^{\wedge}(\varepsilon; \mathbf{z_1}, \ldots, \mathbf{z_l})\right]$$

  is the *VC entropy* of the set of real functions $A \leq Q(\mathbf{z}, \alpha) \leq B$, $\alpha \in \Lambda$, on samples of size l.

Outline

Conditions for uniform two-sided convergence

Theorem

Under some conditions of measurability on the set of real bounded functions $A \leq Q(\mathbf{z}, \alpha) \leq B$, $\alpha \in \Lambda$, for uniform two-sided convergence it is necessary and sufficient that the equality

$$\lim_{l \to \infty} \frac{H^{\wedge}(\varepsilon; l)}{l} = 0, \quad \forall \varepsilon > 0 \tag{8}$$

be valid.

Conditions for uniform two-sided convergence

Corollary

Corollary

Under some conditions of measurability on the set of indicator functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, *for uniform two-sided convergence it is necessary and sufficient that*

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

which is a particular case of (8).

Uniform one-sided convergence

- Uniform two-sided convergence can be described as

$$\lim_{l\to\infty} P\left\{\left[\sup_{\alpha}\left(R(\alpha)-R_{emp}(\alpha)\right)\right]\vee\left[\sup_{\alpha}\left(R_{emp}(\alpha)-R(\alpha)\right)\right]\right\}=0 \quad (9)$$

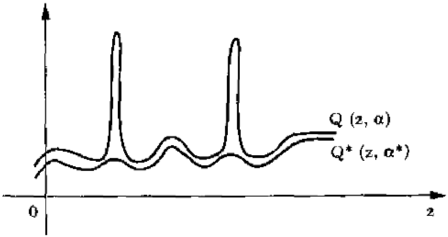
  which includes uniform one-sided convergence, and it's sufficient condition for ERM consistency.
- But for consistency of ERM principle, left-hand side of (9) can be violated.

Conditions for uniform one-sided convergence

- Consider the set of bounded real functions $A \le Q(\mathbf{z}, \alpha) \le B$, $\alpha \in \Lambda$, together with a new set of functions $Q^*(\mathbf{z}, \alpha^*)$, $\alpha^* \in \Lambda^*$, such that

$$Q(\mathbf{z}, \alpha) - Q^*(\mathbf{z}, \alpha^*) \ge 0, \quad \forall \mathbf{z}$$

$$\int (Q(\mathbf{z}, \alpha) - Q^*(\mathbf{z}, \alpha^*))\, dF(\mathbf{z}) \le \delta \qquad (10)$$



Conditions for uniform one-sided convergence

Theorem

Under some conditions of measurability on the set of real bounded functions $A \leq Q(\mathbf{z}, \alpha) \leq B$, $\alpha \in \Lambda$, *for uniform one-sided convergence it is necessary and sufficient that for any positive* δ, η *and* ε *there exist a set of functions* $Q^*(\mathbf{z}, \alpha^*)$, $\alpha^* \in \Lambda^*$, *satisfying* (10) *such that the following holds:*

$$\lim_{l \to \infty} \frac{H^{\wedge}(\varepsilon; l)}{l} < \eta \qquad (11)$$

Outline

Models of reasoning

- Deductive:
  - Moving from general to particular.
  - The ideal approach is to obtain corollaries (consequences) using a system of axioms and inference rules.
  - Guarantees that true consequences are obtained from true premises.
- Inductive:
  - Moving from particular to general.
  - Formation of general judgements from particular assertions.
  - Judgements obtained from particular assertions are not always true.

Demarcation problem

- Proposed by Kant, it is a central question of inductive theory.

Demarcation problem

What is the difference between the cases with a justified inductive step and those for which the inductive step is not justified?

- Is there a formal way to distinguish between true theories and false theories?

Example

- Assume that metereology is a true theory and astrology is a false one.
- What is the formal difference between them?
  - The complexity of the models?
  - The predictive ability of their models?
  - Their use of mathematics?
  - The level of formality of inference?
- None of the above gives a clear advantage to either of these theories.

Criterion for demarcation

- Suggested by Popper (1930), a necessary condition for justifiability of a theory is the feasibility of its falsification.
- By falsification, Popper means the existence of a collection of particular assertions that cannot be explained by the given theory although they fall into its domain.
- If the given theory can be falsified it satisfies the necessary conditions of a scientific theory.

Outline

Nature of the ERM principle

- What happens if the condition of one-side convergence (theorem 11) is not valid?
- Why is the ERM method not consistent is this case?
- Answer: if uniform two-sided convergence does not take place, then the method of minimizing the empirical risk is non-falsifiable.

Complete (Popper's) non-falsiability

- According to the definition of VC entropy the following expressions are valid for a set of indicator functions:

$$H^{\wedge}(l) = E[\ln N^{\wedge}(\mathbf{z_1},\ldots,\mathbf{z_l})] \quad and \quad N^{\wedge}(\mathbf{z_1},\ldots,\mathbf{z_l}) \leq 2^l$$

- Suppose that for a set of indicator fuctions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, the following equality is true:

$$\lim_{l\to\infty} \frac{H^{\wedge}(l)}{l} = \ln 2$$

- It can be shown that the ratio of the entropy to the number of observations decreases monotonically as the number of observations l increases. Therefore for any finite number l the following equality holds true:

$$\frac{H^{\wedge}(l)}{l} = \ln 2$$

Complete (Popper's) non-falsiability

- This means that for almost all samples $\mathbf{z_1}, \ldots, \mathbf{z_l}$ (all but a set of measure zero) the following equality is true:

$$N^{\wedge}(\mathbf{z_1}, \ldots, \mathbf{z_l}) = 2^l$$

- That is, the set of functions of this learning machine is such that almost any sample $\mathbf{z_1}, \ldots, \mathbf{z_l}$ of arbitrary size l can be separated in all possible ways.
- This implies that the minimum of the empirical risk for this machine equals zero independently of the value of the actual risk.
- This learning machine is non-falsiable because it can give a general explanation (function) for almost any data.

Partial non-falsiability

- When entropy of a set of indicator functions over the number of observations tends to a nonzero limit, there exits some subspace of the original space Z where the learning machine is non-falsifiable.

Partial non-falsiability

- Given a set of indicator functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, for which the following convergence is valid:

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

then, there exists a subset Z^* of Z such that

$$P(Z^*) = c$$

and for the subset $\mathbf{z_1^*},\ldots,\mathbf{z_k^*} = (\mathbf{z_1},\ldots,\mathbf{z_l}) \cap Z^*$ and for any given sequence of the binary values $\delta_1,\ldots,\delta_k$, $\delta_i \in \{0,1\}$, there exists a function $Q(\mathbf{z},\alpha^*)$ for which the equalities $\delta_i = Q(\mathbf{z_i^*},\alpha^*)$ holds true.

Potential non-falsiability

Definition

- Considering a set of uniformly bounded real functions $|Q(\mathbf{z}, \alpha)|$, $\alpha \in \Lambda$.
- A learning machine that has an admissible set of real functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, is potentially non-falsiable for a generator of inputs $F(\mathbf{z})$ if there exist two functions

$$\psi_1(\mathbf{z}) \geq \psi_0(\mathbf{z})$$

such that:

1. $\int (\psi_1(\mathbf{z}) - \psi_0(\mathbf{z}))\, dF(\mathbf{z}) = c > 0$
2. For almost any sample $\mathbf{z_1}, \ldots, \mathbf{z_l}$, any sequence of binary values $\delta(1), \ldots, \delta(l)$, $\delta(i) \in \{0, 1\}$, and any $\varepsilon > 0$, one can find a function $Q(\mathbf{z}, \alpha^*)$ in the set of functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, for which the following inequality holds true:

$$\left|\psi_{\delta(i)}(\mathbf{z_i}) - Q(\mathbf{z_i}, \alpha^*)\right| < \varepsilon$$

Potential non-falsiability

Graphical representation

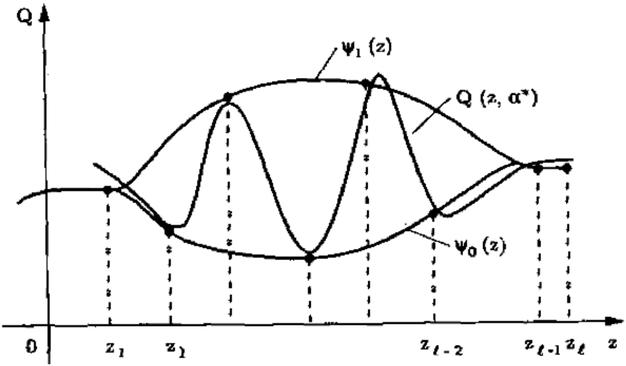


Figure: A potentially non-falsiable learning machine

Potential non-falsiability

Generalization

- This definition of non-falsiability generalizes Popper's concept:
  - Of complete non-falsiability for a set of indicator functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, where $\psi_1(\mathbf{z}) = 1$ and $\psi_0(\mathbf{z}) = 0$.
  - Of partial non-falsiability for a set of indicator functions $Q(\mathbf{z},\alpha)$, $\alpha \in \Lambda$, where

$$\psi_1(\mathbf{z}) = \begin{cases} 1 & if \quad \mathbf{z} \in Z^* \\ Q(\mathbf{z}) & if \quad \mathbf{z} \notin Z^* \end{cases}$$

$$\psi_0(\mathbf{z}) = \begin{cases} 0 & if \quad \mathbf{z} \in Z^* \\ Q(\mathbf{z}) & if \quad \mathbf{z} \notin Z^* \end{cases}$$

Potential non-falsiability

Theorem

Suppose that for the set of uniformly bounded real functions $Q(\mathbf{z}, \alpha)$, $\alpha \in \Lambda$, there exists an ε_0 such that the following convergence is valid:

$$\lim_{l \to \infty} \frac{H^{\wedge}(\varepsilon_0, l)}{l} = c^* > 0$$

Then, the learning machine with this set of functions is potentially non-falsiable.

For Further Reading

- The Nature of Statistical Learning Theory. Vladimir N. Vapnik. ISBN: 0-387-98780-0. 1995.
- Statistical Learning Theory. Vladimir N. Vapnik. ISBN: 0-471-03003-1. 1998.

Questions?

Thank you very much for your attention.

- Contact:
  - Miguel Angel Veganzones
  - Grupo Inteligencia Computacional
  - Universidad del País Vasco - UPV/EHU (Spain)
  - E-mail: miguelangel.veganzones@ehu.es
  - Web page: http://www.ehu.es/computationalintelligence