NIP, O-minimality and Neural Networks

Lothar Sebastian Krapp

Universität Konstanz, Fachbereich Mathematik und Statistik

14 May 2018

Oberseminar mathematische Logik, Mengenlehre und Modelltheorie

1 NIP

2 O-minimality

3 Neural Networks

VC Dimension

VC Dimension

Throughout this section we fix a language $\mathcal{L}$, a complete $\mathcal{L}$-theory T and a monster model $\mathbb{M}$ of T. We abbreviate $\mathbb{M} \models \varphi$ by $\models \varphi$.

VC Dimension

Throughout this section we fix a language $\mathcal{L}$, a complete $\mathcal{L}$-theory T and a monster model $\mathbb{M}$ of T. We abbreviate $\mathbb{M} \models \varphi$ by $\models \varphi$.

Definition

Let $\varphi(\underline{x}, \underline{y})$ be a formula. The **Vapnik–Chervonenkis dimension** (**VC dimension**) of φ is defined as

$$\begin{aligned}&\mathrm{vc}(\varphi(\underline{x};\underline{y})) :=\\ &\qquad \max\{n<\omega \mid \exists(\underline{a}_i)_{i<n}\exists(\underline{b}_J)_{J\subseteq\{0,\ldots,n-1\}}\ [\models \varphi(\underline{a}_i;\underline{b}_J) \text{ if and only if } i\in J]\}\end{aligned}$$

if the maximum exists, and ∞ otherwise.

VC Dimension

VC Dimension

Example

Let T_{dlo} be the theory of dense linear ordered without endpoints. Consider the formula $\varphi(x; y)$ given by $x < y$.

VC Dimension

Example

Let T_{dlo} be the theory of dense linear ordered without endpoints. Consider the formula $\varphi(x; y)$ given by $x < y$.

Let $a_0 \in \mathbb{M}$ be arbitrary and set $b_\emptyset = a_0$ and $b_{\{0\}} > a_0$. Then $\not\models a_0 < b_\emptyset$ and $\models a_0 < b_{\{0\}}$. We obtain $\mathrm{vc}(x < y) \geq 1$.

VC Dimension

Example

Let T_{dlo} be the theory of dense linear ordered without endpoints. Consider the formula $\varphi(x; y)$ given by $x < y$.
Let $a_0 \in \mathbb{M}$ be arbitrary and set $b_\emptyset = a_0$ and $b_{\{0\}} > a_0$. Then $\not\models\ a_0 < b_\emptyset$ and $\models a_0 < b_{\{0\}}$. We obtain $\mathrm{vc}(x < y) \geq 1$.
Now let $a_0, a_1 \in \mathbb{M}$ be arbitrary. Assume that there exist $b_{\{0\}}, b_{\{1\}} \in \mathbb{M}$ such that $a_i < b_J$ if and only if $i \in J$ for $(i, J) \in \{0, 1\} \times \{\{0\}, \{1\}\}$. Then

$$a_0 < b_{\{0\}} \leq a_1 < b_{\{1\}} \leq a_0,$$

a contradiction. Hence, $\mathrm{vc}(x < y) \leq 1$.

NIP

NIP

Definition

A formula $\varphi(\underline{x};\underline{y})$ has the **independence property** (or is **IP**) if $\mathrm{vc}(\varphi) = \infty$. If φ does not have the independence property, it is called **NIP**. A complete theory T in which every formula is NIP is also called NIP.

NIP

Definition

A formula $\varphi(\underline{x}; \underline{y})$ has the **independence property** (or is **IP**) if $\mathrm{vc}(\varphi) = \infty$. If φ does not have the independence property, it is called **NIP**. A complete theory T in which every formula is NIP is also called NIP.

Examples of NIP theories:

NIP

Definition

A formula $\varphi(\underline{x};\underline{y})$ has the **independence property** (or is **IP**) if $\mathrm{vc}(\varphi) = \infty$. If φ does not have the independence property, it is called **NIP**. A complete theory T in which every formula is NIP is also called NIP.

Examples of NIP theories:

- o-minimal theories
- weakly o-minimal theories
- C-minimal theories
- ...

Alternation Number

Alternation Number

A sequence $(\underline{a}_i)_{i<\omega}$ is indiscernible if for every $n < \omega$ and any $i_1 < \ldots < i_n < \omega$ and $j_1 < \ldots < j_n < \omega$ the tuples $(\underline{a}_{i_1}, \ldots, \underline{a}_{i_n})$ and $(\underline{a}_{j_1}, \ldots, \underline{a}_{j_n})$ have the same type, i.e. $\varphi(\underline{a}_{i_1}, \ldots, \underline{a}_{i_n})$ holds if and only if $\varphi(\underline{a}_{j_1}, \ldots, \underline{a}_{j_n})$.

Alternation Number

A sequence $(\underline{a}_i)_{i<\omega}$ is indiscernible if for every $n < \omega$ and any $i_1 < \ldots < i_n < \omega$ and $j_1 < \ldots < j_n < \omega$ the tuples $(\underline{a}_{i_1}, \ldots, \underline{a}_{i_n})$ and $(\underline{a}_{j_1}, \ldots, \underline{a}_{j_n})$ have the same type, i.e. $\varphi(\underline{a}_{i_1}, \ldots, \underline{a}_{i_n})$ holds if and only if $\varphi(\underline{a}_{j_1}, \ldots, \underline{a}_{j_n})$.

Definition

Let $\varphi(\underline{x}; \underline{y})$ be a formula. Let $X \subseteq \omega$ be the set of all $n < \omega$ for which there are an indiscernible sequence $(\underline{a}_i)_{i<\omega}$ and a tuple $\underline{b}$ such that for any $i < n-1$ we have $\models \varphi(\underline{a}_i; \underline{b}) \leftrightarrow \neg\varphi(\underline{a}_{i+1}; \underline{b})$. The **alternation number** $\mathrm{alt}(\varphi)$ of φ is defined as the maximum of X if it exists, or ∞ otherwise.

Alternation Number and VC Dimension

Alternation Number and VC Dimension

Proposition

Let $\varphi(\underline{x}; \underline{y})$ be a formula. Then $\mathrm{alt}(\varphi) \leq 2\,\mathrm{vc}(\varphi) + 1$.

Alternation Number and VC Dimension

Proposition

Let $\varphi(\underline{x}; \underline{y})$ be a formula. Then $\mathrm{alt}(\varphi) \leq 2\,\mathrm{vc}(\varphi) + 1$.

Corollary

1. φ is NIP.
2. $\mathrm{vc}(\varphi) < \infty$.
3. $\mathrm{alt}(\varphi) < \infty$.
4. For every indiscernible sequence $(\underline{a}_i)_{i<\omega}$ and every tuple $\underline{b}$ the set of indices $i < \omega$ such that $\models \varphi(\underline{a}_i; \underline{b})$ holds is finite or cofinite.

Further Tools

Further Tools

Proposition

The set of NIP formulas in T is closed under boolean combinations.

Further Tools

Proposition

The set of NIP formulas in T is closed under boolean combinations.

Proof.

Let $\varphi(\underline{x};\underline{y})$ and $\psi(\underline{x};\underline{y})$ be formulas with finite alternation number.

Further Tools

Proposition

The set of NIP formulas in T is closed under boolean combinations.

Proof.
Let $\varphi(\underline{x}; \underline{y})$ and $\psi(\underline{x}; \underline{y})$ be formulas with finite alternation number. By the previous corollary, it suffices to show that $\neg\varphi(\underline{x}; \underline{y})$ and $\varphi(\underline{x}; \underline{y}) \wedge \psi(\underline{x}; \underline{y})$ have finite alternation number.

Further Tools

Proposition

The set of NIP formulas in T is closed under boolean combinations.

Proof.
Let $\varphi(\underline{x};\underline{y})$ and $\psi(\underline{x};\underline{y})$ be formulas with finite alternation number. By the previous corollary, it suffices to show that $\neg\varphi(\underline{x};\underline{y})$ and $\varphi(\underline{x};\underline{y}) \wedge \psi(\underline{x};\underline{y})$ have finite alternation number. Indeed, $\mathrm{alt}(\neg\varphi) = \mathrm{alt}(\varphi)$ and $\mathrm{alt}(\varphi \wedge \psi) \leq \mathrm{alt}(\varphi) + \mathrm{alt}(\psi) - 1$.

Further Tools

Proposition

The set of NIP formulas in T is closed under boolean combinations.

Proof.
Let $\varphi(\underline{x}; \underline{y})$ and $\psi(\underline{x}; \underline{y})$ be formulas with finite alternation number. By the previous corollary, it suffices to show that $\neg\varphi(\underline{x}; \underline{y})$ and $\varphi(\underline{x}; \underline{y}) \wedge \psi(\underline{x}; \underline{y})$ have finite alternation number. Indeed, $\mathrm{alt}(\neg\varphi) = \mathrm{alt}(\varphi)$ and $\mathrm{alt}(\varphi \wedge \psi) \leq \mathrm{alt}(\varphi) + \mathrm{alt}(\psi) - 1$.

Theorem

Let T be a complete theory. Suppose that every formula of the form $\varphi(x; \underline{y})$ is NIP in T. Then T is NIP.

T_0-minimality

T_0-minimality

Definition

Let T_0 be a theory in a language $\mathcal{L}_0$. A complete theory $T \supseteq T_0$ in a language $\mathcal{L} \supseteq \mathcal{L}_0$ is said to be T_0-**minimal** if for every $\mathcal{L}$-formula $\varphi(x, y_1, \ldots, y_m)$, any model $\mathcal{M} \models T$ and any parameters $b_1, \ldots, b_m \in M$, there exist a quantifier-free $\mathcal{L}_0$-formula $\psi(x, z_1, \ldots, z_n)$ and parameters $c_1, \ldots, c_n \in M$ such that

$$\mathcal{M} \models \forall x(\varphi(x, b_1, \ldots, b_m) \leftrightarrow \psi(x, c_1, \ldots, c_n)).$$

If T_0 is the theory of linear orders, then a T_0-minimal theory T is called **o-minimal**.

T_0-minimality

Definition

Let T_0 be a theory in a language $\mathcal{L}_0$. A complete theory $T \supseteq T_0$ in a language $\mathcal{L} \supseteq \mathcal{L}_0$ is said to be T_0-**minimal** if for every $\mathcal{L}$-formula $\varphi(x, y_1, \ldots, y_m)$, any model $\mathcal{M} \models T$ and any parameters $b_1, \ldots, b_m \in M$, there exist a quantifier-free $\mathcal{L}_0$-formula $\psi(x, z_1, \ldots, z_n)$ and parameters $c_1, \ldots, c_n \in M$ such that

$$\mathcal{M} \models \forall x(\varphi(x, b_1, \ldots, b_m) \leftrightarrow \psi(x, c_1, \ldots, c_n)).$$

If T_0 is the theory of linear orders, then a T_0-minimal theory T is called **o-minimal**.

Example

T_0-minimality

Definition

Let T_0 be a theory in a language $\mathcal{L}_0$. A complete theory $T \supseteq T_0$ in a language $\mathcal{L} \supseteq \mathcal{L}_0$ is said to be T_0-**minimal** if for every $\mathcal{L}$-formula $\varphi(x, y_1, \ldots, y_m)$, any model $\mathcal{M} \models T$ and any parameters $b_1, \ldots, b_m \in M$, there exist a quantifier-free $\mathcal{L}_0$-formula $\psi(x, z_1, \ldots, z_n)$ and parameters $c_1, \ldots, c_n \in M$ such that

$$\mathcal{M} \models \forall x(\varphi(x, b_1, \ldots, b_m) \leftrightarrow \psi(x, c_1, \ldots, c_n)).$$

If T_0 is the theory of linear orders, then a T_0-minimal theory T is called **o-minimal**.

Example

T_{rcf}, the theory of real closed fields, is T_{lo}-minimal. E.g. $(\mathbb{R}, +, \cdot, 0, 1, <) \models \forall x \ (x^2 < 2 \leftrightarrow (-\sqrt{2} < x \wedge x < \sqrt{2}))$.

O-minimality Implies NIP

O-minimality Implies NIP

Proposition

Let T_0 be a theory. Suppose that for any complete theory $T' \supseteq T_0$ every quantifier-free T_0-formula of the form $\varphi(x; \underline{y})$ is NIP in T'. Then every T_0-minimal theory T is NIP.

O-minimality Implies NIP

Proposition

Let T_0 be a theory. Suppose that for any complete theory $T' \supseteq T_0$ every quantifier-free T_0-formula of the form $\varphi(x; \underline{y})$ is NIP in T'. Then every T_0-minimal theory T is NIP.

Theorem

Let T be an o-minimal theory. Then T is NIP.

O-minimality Implies NIP

Proposition

Let T_0 be a theory. Suppose that for any complete theory $T' \supseteq T_0$ every quantifier-free T_0-formula of the form $\varphi(x; \underline{y})$ is NIP in T'. Then every T_0-minimal theory T is NIP.

Theorem

Let T be an o-minimal theory. Then T is NIP.

Proof.

O-minimality Implies NIP

Proposition

Let T_0 be a theory. Suppose that for any complete theory $T' \supseteq T_0$ every quantifier-free T_0-formula of the form $\varphi(x; \underline{y})$ is NIP in T'. Then every T_0-minimal theory T is NIP.

Theorem

Let T be an o-minimal theory. Then T is NIP.

Proof.

Any quantifier-free formula is (equivalent to) a boolean combination of formulas of the form $x < y$.

O-minimality Implies NIP

Proposition

Let T_0 be a theory. Suppose that for any complete theory $T' \supseteq T_0$ every quantifier-free T_0-formula of the form $\varphi(x; \underline{y})$ is NIP in T'. Then every T_0-minimal theory T is NIP.

Theorem

Let T be an o-minimal theory. Then T is NIP.

Proof.

Any quantifier-free formula is (equivalent to) a boolean combination of formulas of the form $x < y$. Since $\mathrm{vc}(x < y) = 1$, these are NIP.

Brief Historical Overview

Brief Historical Overview

- **Vapnik, Chervonenkis, 1971:** paper on statistical learning theory introducing the notion of VC dimensions (for sets rather than formulas)

Brief Historical Overview

- **Vapnik, Chervonenkis, 1971:** paper on statistical learning theory introducing the notion of VC dimensions (for sets rather than formulas)
- **Shelah, 1971:** paper on model theory introducing the independence property

Brief Historical Overview

- **Vapnik, Chervonenkis, 1971:** paper on statistical learning theory introducing the notion of VC dimensions (for sets rather than formulas)
- **Shelah, 1971:** paper on model theory introducing the independence property
- **Pillay, Steinhorn, 1986:** proof that o-minimality implies NIP

Brief Historical Overview

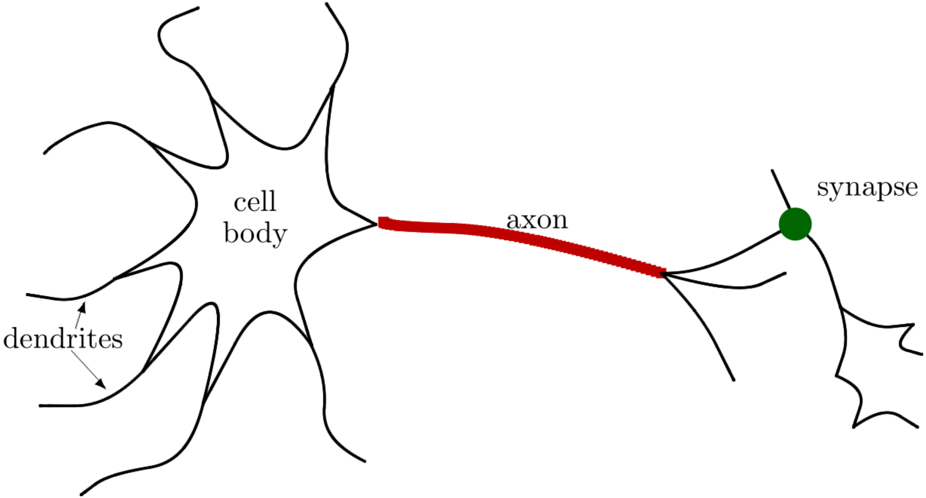
- **Vapnik, Chervonenkis, 1971:** paper on statistical learning theory introducing the notion of VC dimensions (for sets rather than formulas)
- **Shelah, 1971:** paper on model theory introducing the independence property
- **Pillay, Steinhorn, 1986:** proof that o-minimality implies NIP
- **Laskowski, 1992:** connecting the notion of VC dimension to the independence property

Brief Historical Overview

- **Vapnik, Chervonenkis, 1971:** paper on statistical learning theory introducing the notion of VC dimensions (for sets rather than formulas)
- **Shelah, 1971:** paper on model theory introducing the independence property
- **Pillay, Steinhorn, 1986:** proof that o-minimality implies NIP
- **Laskowski, 1992:** connecting the notion of VC dimension to the independence property
- **Wilkie, 1996:** proof that $\mathbb{R}_{\text{an,exp}}$ is o-minimal

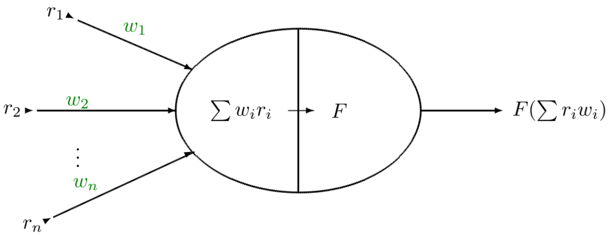
Neurons

Neurons

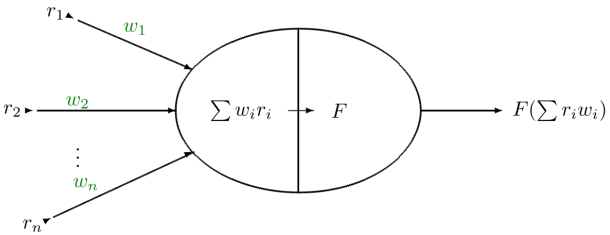


Artificial Neurons

Artificial Neurons

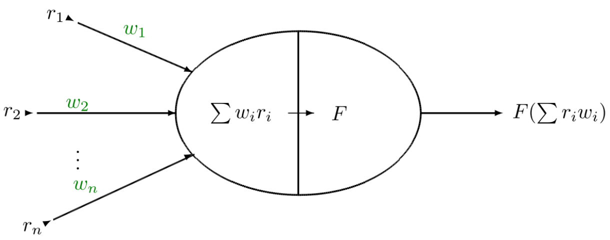


Artificial Neurons



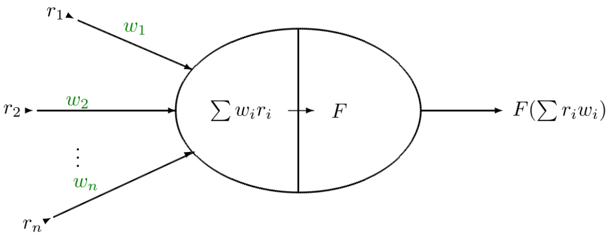
r_i: real numbers, **input**

Artificial Neurons



r_i: real numbers, **input**
w_i: real numbers, **weights**

Artificial Neurons

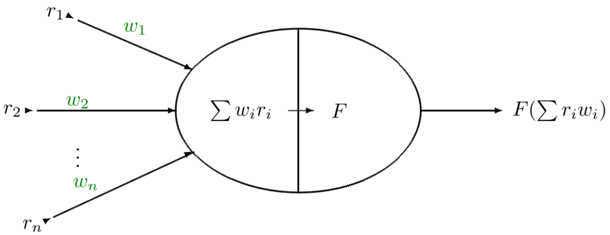


r_i: real numbers, **input**

w_i: real numbers, **weights**

$\sum w_i r_i$: weighted sum

Artificial Neurons



r_i: real numbers, **input**
w_i: real numbers, **weights**
$\sum w_i r_i$: weighted sum
F: real valued function,
activation function

Activation Functions

Typical activation functions F:

Activation Functions

Typical activation functions F:

- characteristic functions on intervals (a, ∞)

Activation Functions

Typical activation functions F:

- characteristic functions on intervals (a, ∞)
- piecewise linear functions

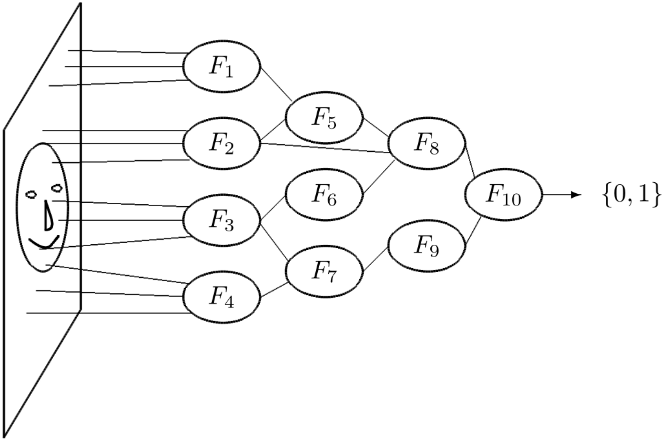
Activation Functions

Typical activation functions F:

- characteristic functions on intervals (a, ∞)
- piecewise linear functions
- sigmoid function $F(t) = \frac{1}{1+\exp(-t)}$

Artificial Neural Network

Artificial Neural Network



(10 neurons, 4 layers)

Artificial Neural Network

X: input space, e.g. $(\mathbb{R}^2 \times \{0, \ldots, 255\})^{12}$

Artificial Neural Network

X: input space, e.g. $(\mathbb{R}^2 \times \{0, \ldots, 255\})^{12}$
Y: output space, e.g. $\{0, 1\}$

Artificial Neural Network

X: input space, e.g. $(\mathbb{R}^2 \times \{0, \ldots, 255\})^{12}$
Y: output space, e.g. $\{0, 1\}$
$F_i(\underline{x}, \underline{w})$: activation functions

Artificial Neural Network

X: input space, e.g. $(\mathbb{R}^2 \times \{0, \ldots, 255\})^{12}$
Y: output space, e.g. $\{0, 1\}$
$F_i(\underline{x}, \underline{w})$: activation functions

The network computes a class of functions $X \to Y$.

Learning Cycle

Learning Cycle

1. network is in an initial state h coded by the weights

Learning Cycle

1. network is in an initial state h coded by the weights
2. training sample $(x, y) \in X \times Y$ is chosen

Learning Cycle

1. network is in an initial state h coded by the weights
2. training sample $(x, y) \in X \times Y$ is chosen
3. $h(x)$ is computed

Learning Cycle

1. network is in an initial state h coded by the weights
2. training sample $(x, y) \in X \times Y$ is chosen
3. $h(x)$ is computed
4. the weights are adjusted depending on $h(x) = y$ or $h(x) \neq y$ (also considering previous training samples)

Learning Cycle

1. network is in an initial state h coded by the weights
2. training sample $(x, y) \in X \times Y$ is chosen
3. $h(x)$ is computed
4. the weights are adjusted depending on $h(x) = y$ or $h(x) \neq y$ (also considering previous training samples)

Goal: After finitely many training samples the network is in a state h which gives a good approximation to recognising the pattern.

Formal Learning

Neural network H: set of all possible functions depending on the weights

Formal Learning

Neural network H: set of all possible functions depending on the weights

Sample space $Z = X \times Y$

Formal Learning

Neural network H: set of all possible functions depending on the weights

Sample space $Z = X \times Y$

Learning algorithm L:

$$L : \bigcup_{m=1}^{\infty} Z^m \to H.$$

Learning Algorithm

p – probability measure on Z measuring the probability that a sample is chosen as training sample

Learning Algorithm

p – probability measure on Z measuring the probability that a sample is chosen as training sample

$\mathrm{er}_p(h) = p\{(x, y) \in Z \mid h(x) \neq y\}$ – error of $h \in H$

Learning Algorithm

p – probability measure on Z measuring the probability that a sample is chosen as training sample

$\mathrm{er}_p(h) = p\{(x, y) \in Z \mid h(x) \neq y\}$ – error of $h \in H$

$\mathrm{opt}_p(H) = \inf_{h \in H} \mathrm{er}_p(h)$ – best approximation in H for given p

Learning Algorithm

Definition

Learning Algorithm

Definition

Let H be a collection of functions $X \to Y$ for a given sample space $Z = X \times Y$. A **learning algorithm** L is a map

$$L : \bigcup_{m=1}^{\infty} Z^m \to H$$

such that it has the following property:

Learning Algorithm

Definition

Let H be a collection of functions $X \to Y$ for a given sample space $Z = X \times Y$. A **learning algorithm** L is a map

$$L : \bigcup_{m=1}^{\infty} Z^m \to H$$

such that it has the following property:
$\forall \varepsilon, \delta \in (0,1)\ \exists m_0 \in \mathbb{N}\ \forall m \geq m_0 :$

Learning Algorithm

Definition

Let H be a collection of functions $X \to Y$ for a given sample space $Z = X \times Y$. A **learning algorithm** L is a map

$$L : \bigcup_{m=1}^{\infty} Z^m \to H$$

such that it has the following property:
$\forall \varepsilon, \delta \in (0,1)\ \exists m_0 \in \mathbb{N}\ \forall m \geq m_0 :$
for any probability measure p on Z we have

$$p^m\{z \in Z^m \mid \mathrm{er}_p(L(z)) < \mathrm{opt}_p(H) + \varepsilon\} \geq 1 - \delta,$$

where p^m is the product measure on Z^m.

Learning Algorithm

Definition

Let H be a collection of functions $X \to Y$ for a given sample space $Z = X \times Y$. A **learning algorithm** L is a map

$$L : \bigcup_{m=1}^{\infty} Z^m \to H$$

such that it has the following property:
$\forall \varepsilon, \delta \in (0,1)\ \exists m_0 \in \mathbb{N}\ \forall m \geq m_0 :$
for any probability measure p on Z we have

$$p^m\{z \in Z^m \mid \mathrm{er}_p(L(z)) < \mathrm{opt}_p(H) + \varepsilon\} \geq 1 - \delta,$$

where p^m is the product measure on Z^m.
H is called **learnable** if there exists a learning algorithm for H.

NIP Implies Learnability

NIP Implies Learnability

Theorem

Let $\mathcal{R} = (\mathbb{R}, +, \cdot, <, \ldots)$ be an expansion of $(\mathbb{R}, +, \cdot, <)$, $X \subseteq \mathbb{R}^d$ a (parametrically) definable set over $\mathcal{R}$ and let H be a collection of activation functions of a neural network $X \to \{0, 1\}$ (parametrically) definable over $\mathcal{R}$. Suppose that the complete theory of $\mathcal{R}$ is NIP. Then H is learnable.

NIP Implies Learnability

Theorem

Let $\mathcal{R} = (\mathbb{R}, +, \cdot, <, \ldots)$ be an expansion of $(\mathbb{R}, +, \cdot, <)$, $X \subseteq \mathbb{R}^d$ a (parametrically) definable set over $\mathcal{R}$ and let H be a collection of activation functions of a neural network $X \to \{0, 1\}$ (parametrically) definable over $\mathcal{R}$. Suppose that the complete theory of $\mathcal{R}$ is NIP. Then H is learnable.

Since $\mathbb{R}_{\text{an,exp}}$ is o-minimal and thus NIP, a set H of $\mathbb{R}_{\text{an,exp}}$-definable activation functions of a neural network is learnable.

References

[1] A. Ya. Chervonenkis and V. N. Vapnik, 'The uniform convergence of frequencies of the appearance of events to their probabilities', *Teor. Verojatnost. i Primenen.* 16 (1971) 264–279 (Russian), *Theor. Probability Appl.* 16 (1971) 264–280 (English).

[2] M. C. Laskowski, 'Vapnik–Chervonenkis classes of definable sets', *J. London Math. Soc. (2)* 45 (1992) 377–384.

[3] A. Pillay and C. Steinhorn, 'Definable sets in ordered structures', I, *Trans. Amer. Math. Soc.* 295 (1986) 565–592.

[4] S. Shelah, 'Stability, the f.c.p., and superstability; model theoretic properties of formulas in first order theory', *Ann. Math. Logic* 3 (1971) 271–362.

Graphics from:

[5] M. Tressl, 'Introduction to o-minimal structures and an application to neural network learning', Preprint, 2010.