

Machine Intelligence 1

2.1 Elements of Statistical Learning Theory

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

WS 2016/2017

Background

Central question

What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations? (Vapnik 1995)

- Regularization principles for solving ill-posed problems (e.g. Tikhonov)
- Nonparametric Statistics (e.g. Parzen)
- Law of large numbers in functional space (Vapnik & Chervonenkis)
- Algorithmic complexity (Kolmogorov, Solomonoff, Chaitin)

Empirical Risk Minimization (ERM)

$$E_{[\underline{\mathbf{w}}]}^G = \int P_{(y_T, \underline{\mathbf{x}})} e_{(y_T, \underline{\mathbf{x}}; \underline{\mathbf{w}})} d\underline{\mathbf{x}} dy_T \stackrel{!}{=} \min \quad (\text{generalization error})$$

mathematical expectation

\Downarrow ERM \Downarrow \leftarrow When does this work?
 empirical average

$$E_{[\underline{\mathbf{w}}]}^T = \frac{1}{p} \sum_{\alpha=1}^p e_{(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})} \stackrel{!}{=} \min \quad (\text{training error})$$

The law of large numbers

“The sequence of means converges to the expectation of a random variable (if it exists) as the number p of samples increases.”

- can only be applied if set of predictors has a finite number of elements (use Hoeffding's inequality for bound on deviations)
- extension to infinite sets (e.g. MLPs) necessary

Statistical learning theory (SLT)

Goal: Estimate a desired function from

- a wide set of functions
- a limited number of examples

Statistical learning theory (SLT)

Goal: Estimate a desired function from

- a wide set of functions
- a limited number of examples

Achievements of statistical learning theory

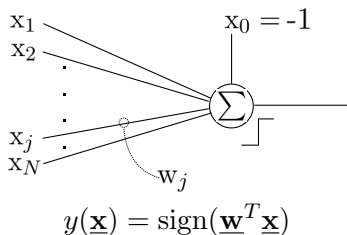
- ① formulation of conditions under which ERM works
- ② bounds describing **generalization ability** of ERM
- ③ inductive inference for **small sample sizes** based on these bounds
- ④ methods for implementing this new type of inference (\rightarrow **SVMs**)

2.1.1 Linear Classifiers: An Example

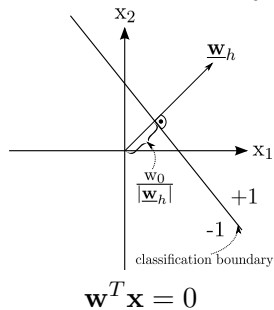
Linear classifiers

- **data representation:** samples $\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N$ and classes $y_T^{(\alpha)} \in \{-1, 1\}$
- **model class:** binary connectionist neuron

binary connectionist neuron



classification boundary



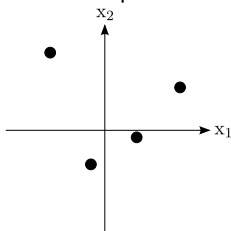
Optimization with ERM

- **performance measure:** classification error
- **optimization:** minimizing the training error (ERM)
 - which conditions yield small training errors?
- **validation:** evaluate test error on unseen data
 - which conditions yield small generalization errors?

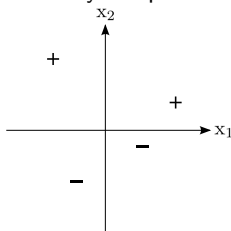
Linear separability

- \leadsto classes do not overlap
 \leadsto classes can be separated by a hyperplane
- $\left. \begin{array}{l} \leadsto \text{classes do not overlap} \\ \leadsto \text{classes can be separated by a hyperplane} \end{array} \right\} \begin{array}{l} \text{all data points can be correctly} \\ \text{classified by at least one classifier} \\ \text{from the set} \end{array}$

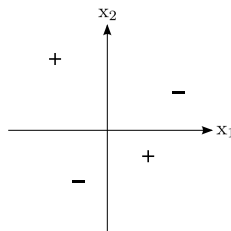
- observations $\underline{x}^{(\alpha)} \in \mathbb{R}^N, \alpha = 1, \dots, p$, in *general location* (\rightarrow each subset of N points should be linearly independent)



in general location



separable assignment



ass. not linearly separable

- 14 linearly separable configurations, 2 configurations not separable

Statistics of linear separability (Cover, 1965)

(1) **number** $C_{(p,N)}$ of linearly separable assignments $y_T : \underline{\mathbf{x}} \rightarrow y_T$:

$$C_{(p,N)} = 2 \sum_{k=0}^{N-1} \binom{p-1}{k}$$

Statistics of linear separability (Cover, 1965)

(1) **number** $C_{(p,N)}$ of linearly separable assignments $y_T : \underline{x} \rightarrow y_T$:

$$C_{(p,N)} = 2 \sum_{k=0}^{N-1} \binom{p-1}{k}$$

(2) **fraction** $\Pi_{(p,N)}$ of linearly separable assignments (total: 2^p):

$$\Pi_{(p,N)} = \frac{C_{(p,N)}}{2^p} = \frac{1}{2^{p-1}} \sum_{k=0}^{N-1} \binom{p-1}{k}$$

Asymptotic distribution

- change in the fraction of linearly separable assignments
(when one data point is added to the training set)

$$\Delta\Pi_{(p)} := \Pi_{(p,N)} - \Pi_{(p+1,N)} = \frac{C_{(p,N)}}{2^p} - \frac{1}{2} \frac{C_{(p+1,N)}}{2^p}$$

(see blackboard)

Asymptotic distribution

- change in the fraction of linearly separable assignments
(when one data point is added to the training set)

$$\Delta\Pi_{(p)} := \Pi_{(p,N)} - \Pi_{(p+1,N)} = \frac{C_{(p,N)}}{2^p} - \frac{1}{2} \frac{C_{(p+1,N)}}{2^p}$$

- $\Delta\Pi_{(p)}$ follows a **binomial distribution**

$$\Delta\Pi_{(p)} \sim \binom{p-1}{N-1} \left(\frac{1}{2}\right)^{N-1} \left(\frac{1}{2}\right)^{(p-1)-(N-1)}$$

(see blackboard)

Asymptotic distribution

- change in the fraction of linearly separable assignments
(when one data point is added to the training set)

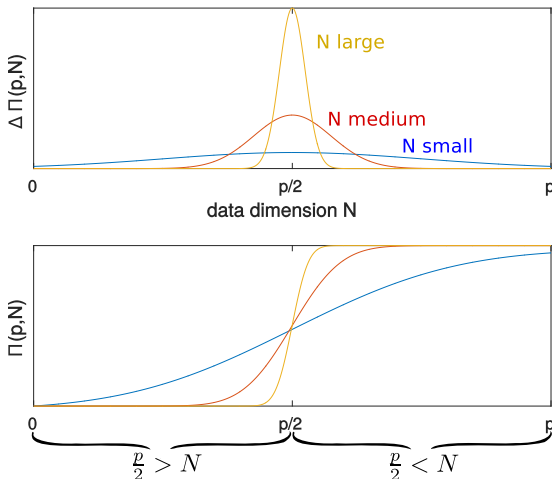
$$\Delta\Pi_{(p)} := \Pi_{(p,N)} - \Pi_{(p+1,N)} = \frac{C_{(p,N)}}{2^p} - \frac{1}{2} \frac{C_{(p+1,N)}}{2^p}$$

Binomial distribution converges to Gaussian distribution

In the limit $N, p \rightarrow \infty$, with $\frac{p}{N} = \text{const}$, holds:

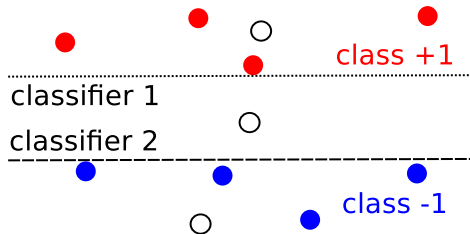
$$\Delta\Pi_{(p)} \approx \mathcal{N}_{(2N, \sqrt{N})} \quad (\text{asymptotic distribution})$$

Capacity of the set of linear classifiers

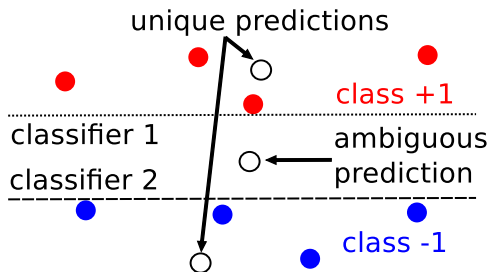


- all 2^p label configurations can be realized below $\beta := \frac{p}{N} = 2$
- β can be understood as a natural capacity measure

Generalization performance (1)



Generalization performance (1)



- number of ambiguous assignments: $C_{(p, N-1)}$ (Cover, 1965)

Generalization performance (2)

- number of ambiguous assignments: $C_{(p,N-1)}$
- fraction of ambiguous, linearly separable assignments

$$g_{(p,N)} := \frac{\text{ambiguous assignments}}{\text{all separable assignments}} = \frac{C_{(p,N-1)}}{C_{(p,N)}}$$

- if fraction of ambiguous assignments is large, prediction for new data points will not be informative

Generalization performance (2)

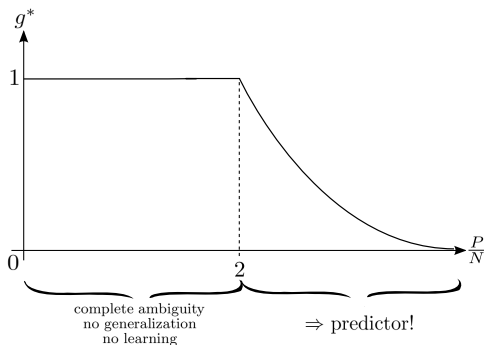
- number of ambiguous assignments: $C_{(p,N-1)}$
- fraction of ambiguous, linearly separable assignments

$$g_{(p,N)} := \frac{\text{ambiguous assignments}}{\text{all separable assignments}} = \frac{C_{(p,N-1)}}{C_{(p,N)}}$$

- if fraction of ambiguous assignments is large, prediction for new data points will not be informative
- asymptotic result (Cover, 1965):

$$g_{\left(\frac{p}{N}\right)}^* = \lim_{\substack{p, N \rightarrow \infty, \\ \frac{p}{N} = \text{const}}} g_{(p,N)} = \begin{cases} 1, & \text{for } 0 \leq \frac{p}{N} \leq 2 \\ \frac{1}{\frac{p}{N} - 1}, & \text{for } \frac{p}{N} > 2 \end{cases}$$

How much data do we need to predict?



- for $p/N \leq 2$ almost all predictions are ambiguous.
- for $p/N > 2$ classifier is able to generalize to new data points

“The probability of ambiguous generalization is large unless the number of training patterns exceeds the capacity of the set of separating surfaces.”

The Vapnik-Chervonenkis (VC) dimension

key question of inductive reasoning

When will conclusions inferred from a sample of data be meaningful?

→ concept of non-falsifiability (Popper)

- a theory is useful only, if it is **falsifiable** (cmp. astrology),
- i.e. if it cannot predict all potentially observable patterns.

The Vapnik-Chervonenkis (VC) dimension

key question of inductive reasoning

When will conclusions inferred from a sample of data be meaningful?

→ concept of non-falsifiability (Popper)

- a theory is useful only, if it is **falsifiable** (cmp. astrology),
- i.e. if it cannot predict all potentially observable patterns.

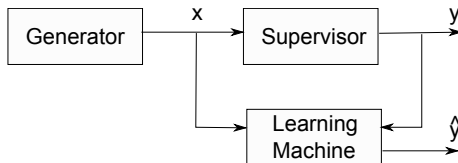
generalization of the capacity measure

VC-dimension d_{VC} describes the capacity of a family of functions \mathcal{M}

- d_{VC} is the **largest number** of samples in general locations...
 - that can be classified by some $f \in \mathcal{M}$ for any possible labeling
 - that can be classified by some $f \in \mathcal{M}$ without any empirical error

2.1.2 Formulation of the Inductive Learning Problem

Inductive learning for classification



- generator draws $\underline{x} \in R^n$ w.r.t. unknown $P(\underline{x})$
- supervisor returns output y_T to every \underline{x} w.r.t. unknown $P(y_T|\underline{x})$
- learning machine estimates function $f(\underline{x}, \underline{w}) \approx y_T$ by selecting $\underline{w} \in \Lambda$ from the set of possible parameters Λ

Empirical Risk Minimization (ERM)

$$E_{[\mathbf{w}]}^G = \int P_{(y_T, \mathbf{x})} e_{(y_T, \mathbf{x}; \mathbf{w})} d\mathbf{x} dy_T \stackrel{!}{=} \min \quad (\text{generalization error})$$

mathematical expectation

↓ ERM ↓

empirical average

← When does this work?

$$\min E_{[\mathbf{w}]}^T = \frac{1}{p} \sum_{\alpha=1}^p e_{(y_T^{(\alpha)}, \mathbf{x}^{(\alpha)}; \mathbf{w})} \stackrel{!}{=} \min \quad (\text{training error})$$

Change in nomenclature (\rightarrow Vapnik, 1998)

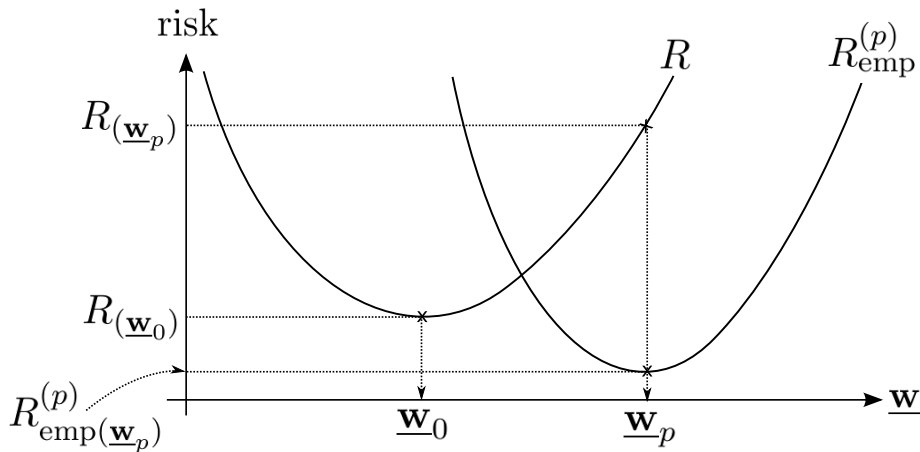
$$E_{[\mathbf{w}]}^G \rightarrow R_{[\mathbf{w}]} \quad (\text{risk})$$

$$E_{[\mathbf{w}]}^T \rightarrow R_{\text{emp}[\mathbf{w}]} \quad (\text{empirical risk})$$

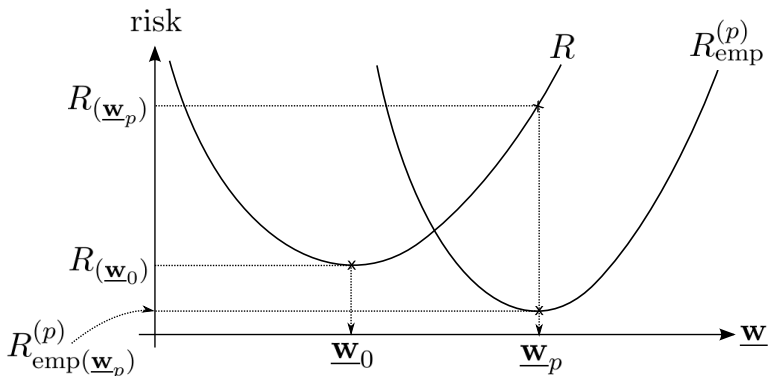
data representation

- p observations $\underline{\mathbf{z}}^{(\alpha)} := (\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)})$, $\alpha \in \{1, \dots, p\}$
- observations are drawn **iid** from unknown distribution
 $P(\underline{\mathbf{x}}, y_T) = P(y_T | \underline{\mathbf{x}}) P(\underline{\mathbf{x}})$.
- **stationarity**: training and test distributions are identical

The learning problem



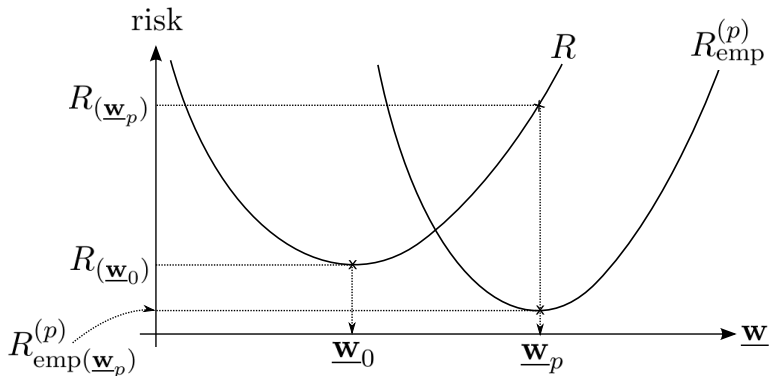
The key questions of Statistical Learning Theory



- ① When does inductive learning through ERM work?
 - $R(\underline{\mathbf{w}}_p)$ should reflect the true optimal risk $R(\underline{\mathbf{w}}_0)$ in the limit $p \rightarrow \infty$:

$$\lim_{p \rightarrow \infty} P\left\{|R_{(\underline{\mathbf{w}}_p)} - R_{(\underline{\mathbf{w}}_0)}| \geq \eta\right\} = 0, \quad \text{for all } \eta > 0.$$

The key questions of Statistical Learning Theory

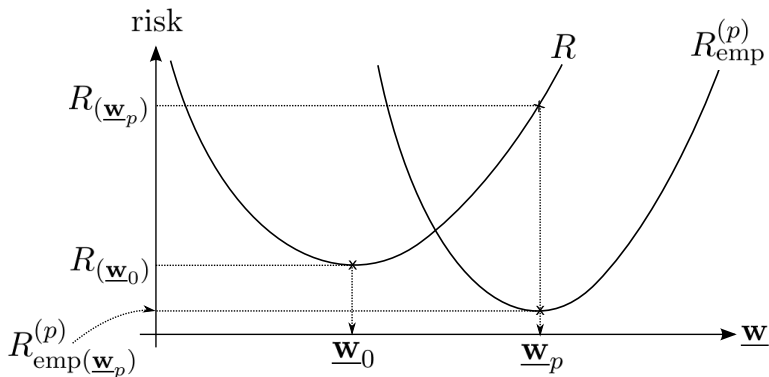


② How strongly does $R(\underline{w}_p)$ differ from $R(\underline{w}_0)$ for **finite** samples?

- For a given confidence ϵ , find η such that

$$P\left\{|R(\underline{w}_0) - R(\underline{w}_p)| \geq \eta\right\} < \epsilon.$$

The key questions of Statistical Learning Theory



- ③ Do training errors $R_{\text{emp}}(\underline{\mathbf{w}})$ reflect generalization performance $R(\underline{\mathbf{w}})$?
- Avoid overfitting for finite samples:

$$P\left\{|R(\underline{\mathbf{w}}) - R_{\text{emp}}(\underline{\mathbf{w}})| > \eta\right\} < \epsilon, \quad \forall \underline{\mathbf{w}} \in \Lambda.$$

2.1.3 General Classifiers: Conditions for Inductive Learning

Conditions for successful learning with ERM

The success of ERM depends on...

- the model class (\rightarrow set of classifiers)
- the availability of data (\rightarrow growth function)

Conditions for successful learning with ERM

The success of ERM depends on...

- the model class (\rightarrow set of classifiers)
- the availability of data (\rightarrow growth function)

Key concepts: growth function & capacity measures (VC-dimension)

results (milestones)

- ① asymptotic result: learning success for a potentially infinite number of data
- ② finite samples: deviation from the optimal model
- ③ finite samples: bound on the generalization error

for technical details, conditions for non-trivial consistency etc. see Vapnik (1998)

The growth function $G_{(p)}^{\Lambda}$

- **data representation:** $\underline{\mathbf{x}} \in \mathbb{R}^N$, $y_T \in \{-1, +1\}$
- **model class:** set Λ of functions $\hat{y}_{(\underline{\mathbf{x}}; \underline{\mathbf{w}})} \in \{-1, +1\}$
- **binary label vector:** $\underline{\mathbf{y}}_{(\underline{\mathbf{w}})} = \left(\hat{y}_{(\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{w}})}, \hat{y}_{(\underline{\mathbf{x}}^{(2)}, \underline{\mathbf{w}})}, \dots, \hat{y}_{(\underline{\mathbf{x}}^{(p)}, \underline{\mathbf{w}})} \right)$
 - different classifiers can induce the same label vector on the training set
- **number** of different vectors $\underline{\mathbf{y}}_{(\underline{\mathbf{w}})}$ induced by all $\hat{y} \in \Lambda$:

$$N_{(\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)})}^{\Lambda} \leq 2^p \quad (\text{depends on } \Lambda \text{ and the samples})$$

The growth function $G_{(p)}^{\Lambda}$

$$G_{(p)}^{\Lambda} = \ln \left(\underbrace{\sup_{\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)}} N_{(\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)})}^{\Lambda}}_{\text{worst case}} \right)$$

The growth function $G_{(p)}^{\Lambda}$

$$G_{(p)}^{\Lambda} = \ln \underbrace{\left(\sup_{\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)}} N_{(\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)})}^{\Lambda} \right)}_{\text{worst case}}$$

- bound on growth function (Vapnik, 1998)

$$G_{(p)}^{\Lambda} \begin{cases} = p \ln 2 & \text{for } p \leq d_{\text{VC}} \\ \leq d_{\text{VC}} \left(1 + \ln \frac{p}{d_{\text{VC}}} \right) & \text{for } p > d_{\text{VC}} \end{cases}$$

The growth function $G_{(p)}^{\Lambda}$

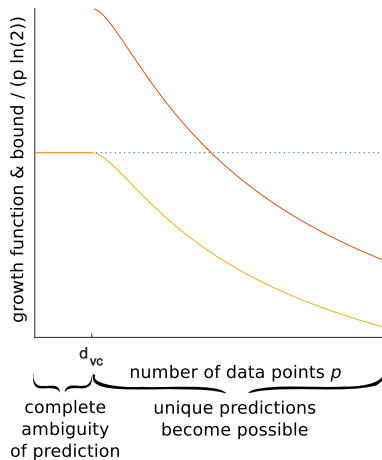
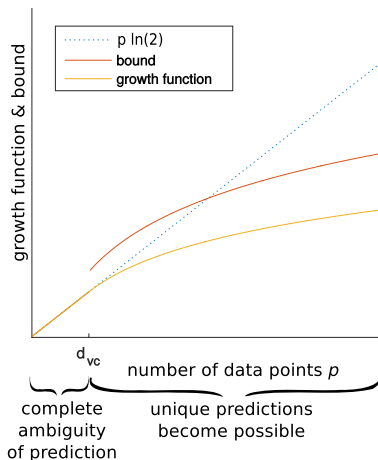
$$G_{(p)}^{\Lambda} = \ln \underbrace{\left(\sup_{\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)}} N_{(\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(p)})}^{\Lambda} \right)}_{\text{worst case}}$$

- bound on growth function (Vapnik, 1998)

$$G_{(p)}^{\Lambda} \begin{cases} = p \ln 2 & \text{for } p \leq d_{VC} \\ \leq d_{VC} \left(1 + \ln \frac{p}{d_{VC}} \right) & \text{for } p > d_{VC} \end{cases}$$

- Vapnik-Chervonkis dimension d_{VC} :
capacity measure of the model class

The growth function $G_{(p)}^{\Lambda}$



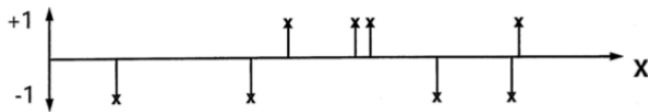
- d_{VC} : capacity measure of the model class
 - small d_{VC} : small sample size sufficient for learning
 - large d_{VC} : large sample size required

The Vapnik-Chervonenkis dimension

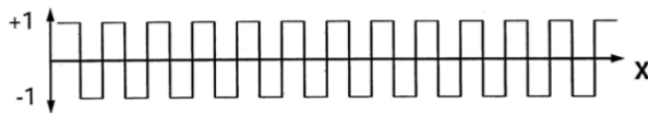
- d_{VC} is ...
 - largest number of data points on which (at least for one set) 2^p different loss vectors (i.e. label assignments) can be induced
 - capacity measure for a given set of models
 - purely combinatorial concept (worst case scenario)
 - provides a bound on the growth function ...
 - ... \leadsto determines the number of samples needed to reliably learn
- d_{VC} is typically hard to compute
 - bounds have been derived for some architectures
 - e.g. MLPs with sigmoidal units: $d_{VC} = \mathcal{O}(W^2)$,
where W is number of parameters, Koiran and Sontag (1996)

An unlearnable problem

Training examples: $\{(\underline{x}^{(\alpha)}, y_T^{(\alpha)})\}$, $\alpha = 1, \dots, p$ with $\underline{x}^{(\alpha)} \in \mathcal{R}$, $y_T^{(\alpha)} \in \{0, 1\}$

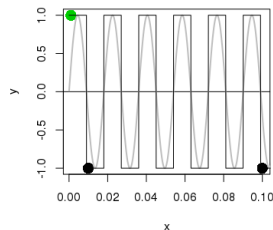
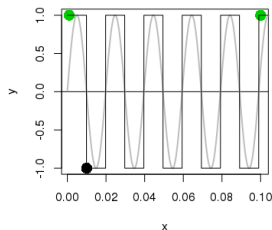
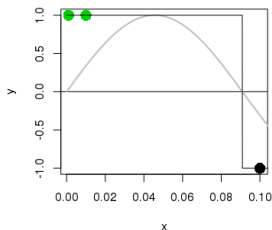


Set of classifiers: $Y(\underline{x}, a) = \text{sign}(\sin(ax))$



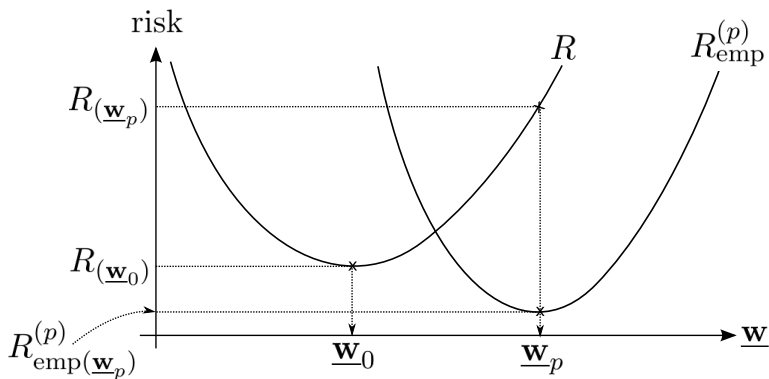
An unlearnable problem

- set of classifiers: $\hat{y}(x; a) = \text{sign}[\sin(ax)]$, single model parameter a
- selection of classifiers (cf. Vapnik, 1998)
 - take $x_i = 10^{-i}$, $i = 1, \dots, N$ and assign any $y_i \in \{-1, +1\}$
 - set $a = \pi(1 + \sum_{i=1}^N \frac{(1-y_i)10^i}{2})$



- a single model parameter, but an infinite d_{VC}

The learning problem and its solution



The learning problem and its solution

① learnability: asymptotic result

$$d_{VC} \text{ finite} \Rightarrow \lim_{p \rightarrow \infty} P \left\{ \left| R(\underline{\mathbf{w}}_p) - R(\underline{\mathbf{w}}_0) \right| \geq \eta \right\} = 0, \quad \forall \eta > 0$$

- convergence of ERM is only guaranteed if d_{VC} is finite

② finite samples: deviation from the optimal model

$$P \left\{ R(\underline{\mathbf{w}}_p) - R(\underline{\mathbf{w}}_0) > \left(\frac{G_{(2p)}^\Lambda - \ln \frac{\epsilon}{8}}{p} \right)^{\frac{1}{2}} + \left(-\frac{\ln \frac{\epsilon}{2}}{2p} \right)^{\frac{1}{2}} + \frac{1}{p} \right\} < \epsilon$$

- finite d_{VC} : increased sample size $p \Rightarrow$ reduction of the bound

③ finite samples: bound on the generalization error

$$P \left\{ \sup_{\underline{\mathbf{w}} \in \Lambda} \left| R(\underline{\mathbf{w}}) - R_{\text{emp}}^{(p)}(\underline{\mathbf{w}}) \right| > \eta \right\} < 4 \exp \left(G_{(2p)}^\Lambda - p \left(\eta - \frac{1}{p} \right)^2 \right)$$

- bound non-trivial only if $G_{(2p)}^\Lambda$ is sub-linear in p

Regression problems

- regression problems can be treated in a similar manner

End of Section 2.1

the following slides contain

OPTIONAL MATERIAL

Mathematical preliminaries

binomial Coefficient

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!}, & n \geq k \\ 0, & \text{else} \end{cases}$$

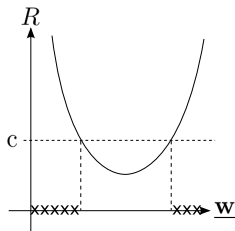
binomial distribution

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

cumulative distribution

$$F(k; n, p) = \Pr(X \leq k) = \sum_{x=0}^k f(x; n, p)$$

Strict Consistency

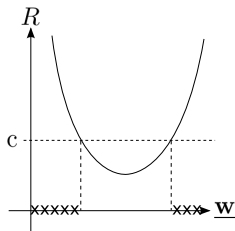


$\Lambda = \{\underline{\mathbf{w}}\}$: set of all predictors

$\Lambda_{(c)} \subset \Lambda$: set of all "bad" predictors

xxx : $\Lambda_{(c)} = \{\underline{\mathbf{w}} : R_{(\underline{\mathbf{w}})} \geq c\}$

Strict Consistency



$\Lambda = \{\underline{\mathbf{w}}\}$: set of all predictors
 $\Lambda_{(c)} \subset \Lambda$: set of all "bad" predictors

xxx : $\Lambda_{(c)} = \{\underline{\mathbf{w}} : R_{(\underline{\mathbf{w}})} \geq c\}$

definition and implications of strict consistency

$$\lim_{p \rightarrow \infty} P \left\{ \left| \inf_{\underline{\mathbf{w}} \in \Lambda_{(c)}} R_{\text{emp}}^{(p)}(\underline{\mathbf{w}}) - \inf_{\underline{\mathbf{w}} \in \Lambda_{(c)}} R_{(\underline{\mathbf{w}})} \right| \geq \eta \right\} = 0$$

strict consistency $\xrightarrow{\quad}$ inductive learning via ERM

proof: supplementary material

Strict Consistency \rightarrow convergence in the limit

key theorem of learning theory

let $a \leq R_{(\underline{\mathbf{w}})} \leq A$ for all $\underline{\mathbf{w}} \in \Lambda$ and $P_{(\underline{\mathbf{z}})} \in \Pi$, then:

the ERM procedure is strictly consistent for Λ and Π

$$\lim_{p \rightarrow \infty} P \left\{ \sup_{\underline{\mathbf{w}} \in \Lambda} \left(R_{(\underline{\mathbf{w}})} - R_{\text{emp}(\underline{\mathbf{w}})}^{(p)} \right) > \eta \right\} = 0, \quad \text{for all } P_{(\underline{\mathbf{z}})} \in \Pi.$$

\Updownarrow

proof: supplementary material

- uniform (one-sided) convergence of $R_{\text{emp}(\underline{\mathbf{w}})}^{(p)}$ to $R_{(\underline{\mathbf{w}})}$ for $p \rightarrow \infty$