Probably Approximately Correct Learning

Concepts,
Instance Space, Hypothesis Space,
Risks,
the PAC Learning Model,
Rectangle Learning

Joachim M. Buhmann

December 26, 2019

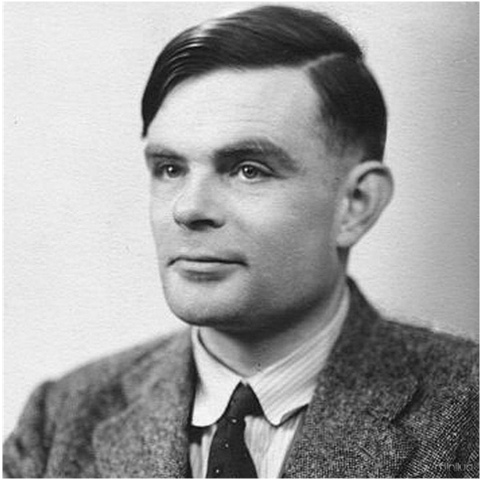
Statistical Learning Theory: the Setting

Mathematics and computer science has made remarkable breakthroughs in the last century.

- Better understanding of the world.
- Higher computing power.
- Outperforming humans in different tasks.

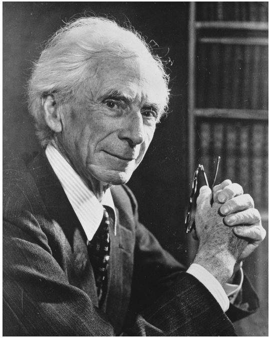
Capabilities of computers

- Can machines compute anything?
- No! Turing's halting problem, post correspondence problem.

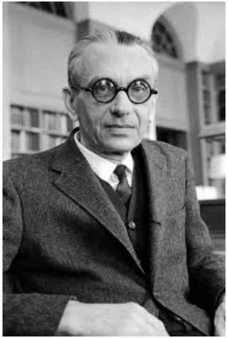


The breakthroughs of mathematics

- Bertrand Russell starts writing his Principia Mathematica
- Can formal logic prove anything?
- No! Gödel’s incompleteness theorem: there are infinitely many truths about arithmetic that cannot be proven formally.



Bertrand Russel

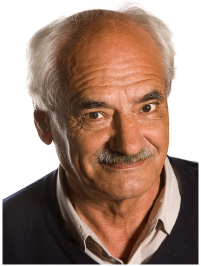


Kurt Gödel

The breakthroughs of machine learning

- Can machines learn anything?
- Can we learn a function to arbitrary precision with high probability?

Vladimir Vapnik



Alexey Chervonenkis

Statistical learning theory

Statistical learning theory is a framework for machine learning aiming at learning functions from data.

PAC learning is a subfield that concerns with the following questions.

- What is “learnable”? Can we learn anything?
- If something is “learnable”, how much can we learn it by empirically minimizing a “cost function”?

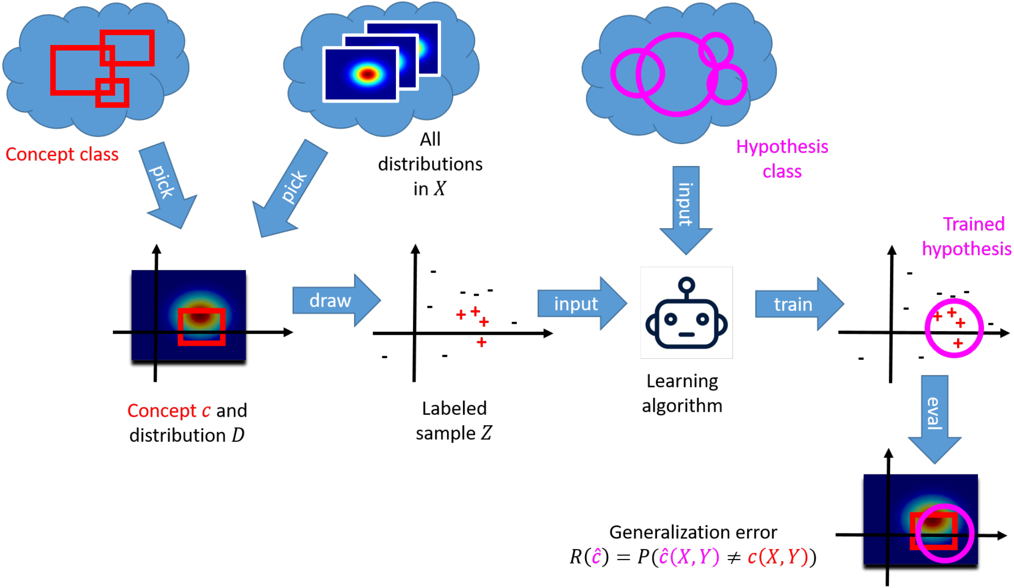
The following material is from the books

- Mohri, Rostamizadeh, and Talwalkar. Foundations of machine learning.
- Devroye, Györfi, and Lugosi. A probabilistic theory of pattern recognition.

Agenda

1. Motivation for statistical learning theory
2. Basic concepts
3. What is learnable.
4. Example of “learnable” concepts.
5. Useful inequalities from statistical learning theory.

The learning problem



Concept class

pick

pick

All distributions in X

Hypothesis class

input

Trained hypothesis $\hat{c}$

draw

input

train

Concept c and distribution D

Labeled sample Z

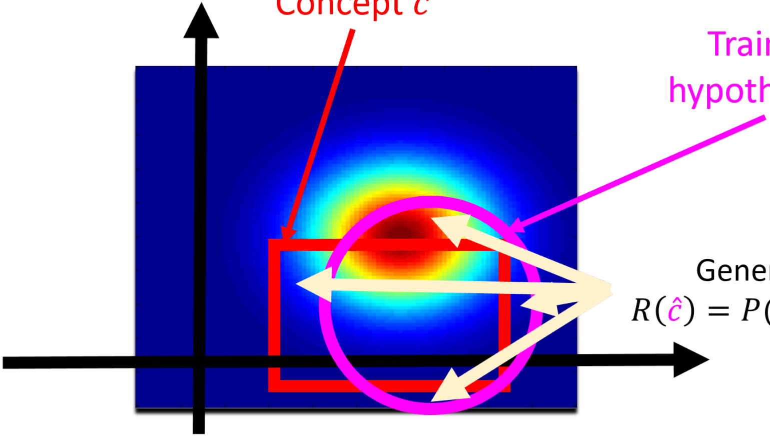
Learning algorithm

eval

Generalization error

$$R(\hat{c}) = P(\hat{c}(X,Y) \neq c(X,Y))$$

The learning problem



Concept c

Trained hypothesis $\hat{c}$

Generalization error

$$R(\hat{c}) = P(\hat{c}(X,Y) \neq c(X,Y))$$

Generalization error and empirical error

Generalization error Not computable by the learner:

$$\mathcal{R}(\hat{c}) := \mathbf{P}\left(\hat{c}(X) \neq c(X)\right).$$

Empirical error Computable by the learner:

$$\hat{\mathcal{R}}_n(\hat{c}) := \frac{1}{n} \sum_{i \leq n} \mathbf{1}_{\hat{c}(x_i) \neq c(x_i)}.$$

One can show that $\mathbb{E}\left[\hat{\mathcal{R}}_n(\hat{c})\right] = \mathcal{R}(\hat{c})$.

Notions from statistical learning theory

Instance space $\mathcal{X}$: think of $\mathcal{X}$ as being a set of instances or objects in the learner's world.

Concept: A concept is a subset c of $\mathcal{X}$ (we sometimes think of c as a function $c : \mathcal{X} \to \{0, 1\}$).

Concept class: A set of concepts we wish to learn.

Hypothesis class: Another set of concepts that we use to learn a target concept from the concept class.

No additional prior knowledge on the distribution on $\mathcal{X}$ is available.

Observe that this differs from Bayesian approaches, which require a prior on $\mathcal{X}$.

The PAC Learning Model

Definition

Let $\mathcal{H}$ and c be a hypothesis class and a concept. A learning algorithm is an algorithm that receives as input a labeled sample $\mathcal{Z} = \{(x_1, c(x_1)), \ldots, (x_n, c(x_n))\}$ and outputs a hypothesis $\hat{c} \in \mathcal{H}$.

The PAC Learning Model

A learning algorithm $\mathcal{A}$ can learn a concept c from $\mathcal{H}$ if, given as input a sufficiently large sample, it outputs a hypothesis that generalizes well with high probability.

Definition

A learning algorithm $\mathcal{A}$ can learn a concept c if there is a polynomial function *poly* $(\cdot,\cdot,\cdot)$ such that

1. for any distribution $\mathcal{D}$ on $\mathcal{X}$ and
2. for any $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$,

if $\mathcal{A}$ receives as input a sample $\mathcal{Z}$ of size $n \geq \textit{poly}\left(1/\epsilon, 1/\delta, \textit{dim}\left(\mathcal{X}\right)\right)$, then $\mathcal{A}$ outputs $\hat{c}$ such that

$$\mathbf{P}_{\mathcal{Z}\sim\mathcal{D}^n}\left(\mathcal{R}\left(\hat{c}\right) \leq \epsilon\right) \geq 1 - \delta.$$

This probability is taken over $\mathcal{Z}$ and any internal randomization of $\mathcal{A}$. The value $\textit{dim}\left(\mathcal{X}\right)$ indicates the instance space $\mathcal{X}$'s number of dimensions.

The PAC Learning Model

Definition

A concept class $\mathcal{C}$ is PAC learnable from a hypothesis class $\mathcal{H}$ if there is an algorithm that can learn any concept in $\mathcal{C}$.

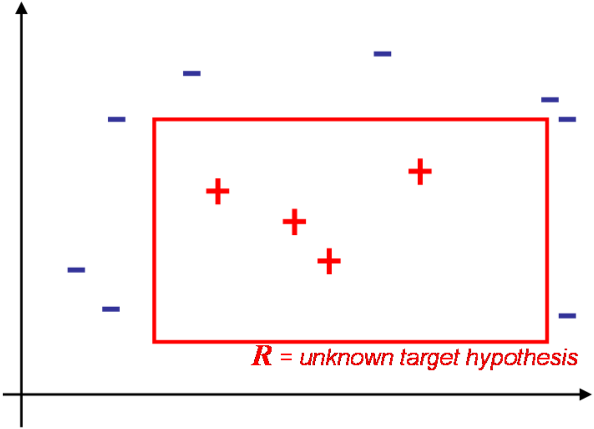
Efficient PAC learning: If $\mathcal{A}$ runs in time polynomial in $1/\epsilon$ and $1/\delta$, we say that $\mathcal{C}$ is efficiently PAC learnable.

- The input ϵ is called the **error parameter**, the parameter δ denotes the **confidence** value.

Remark: No assumptions on the distribution of instances are made!

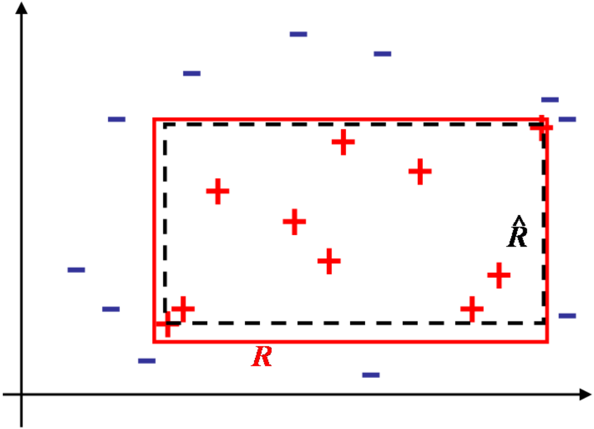
Example: Axis-aligned rectangles are PAC learnable

Let $\mathcal{C}$ be the concept of all axis-aligned rectangles. We show that $\mathcal{C}$ can be learned from $\mathcal{H} = \mathcal{C}$.



Example: Axis-aligned rectangles are PAC learnable

Consider the algorithm $\mathcal{A}$ that outputs the smallest rectangle $\hat{R}$ containing all positively labeled points. We show that $\mathcal{A}$ can learn any concept $R \in \mathcal{C}$.



How do we prove that $\mathcal{A}$ learns rectangles?

We show that there exists $poly\,(\cdot,\cdot)$ such that

- for any rectangle $R \in \mathcal{C}$,
- for any distribution $\mathcal{D}$ on $\mathbb{R}^2$,
- for any $\epsilon > 0$ and $\delta > 0$,

if $\mathcal{A}$ receives a sample $\mathcal{Z}$ of size $n \geq poly\,(1/\epsilon, 1/\delta, dim\,(\mathcal{X}) = 2)$, then

$$\mathbf{P}\left(\mathcal{R}(\hat{R}) \leq \epsilon\right) \geq 1 - \delta.$$

How do we prove that $\mathcal{A}$ learns rectangles?

We define next an event called $\hat{\mathcal{R}}IG$ (which stands for “$\hat{\mathcal{R}}$ is good enough”) such that

$$\mathbf{P}\left(\mathcal{R}(\hat{R}) \leq \epsilon\right) \geq \mathbf{P}\left(\hat{\mathcal{R}}IG\right) \geq 1 - 4e^{-\frac{n\epsilon}{4}}.$$

Observe that we just need to ensure that $1 - 4e^{-\frac{n\epsilon}{4}} \geq 1 - \delta$ or equivalently that

$$n \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}.$$

We can ensure this by letting

$$n \geq \underbrace{\frac{4}{\epsilon} \times \frac{4}{\delta}}_{poly(1/\epsilon, 1/\delta, 2)} \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}.$$

Tasks to do

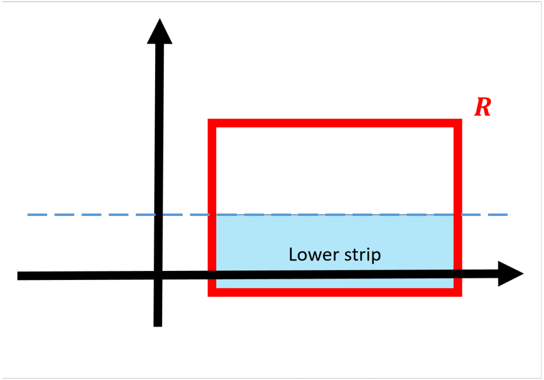
1. Define $\hat{\mathcal{R}}IG$
2. Prove that

$$\mathbf{P}\left(\mathcal{R}(\hat{R}) \leq \epsilon\right) \geq \mathbf{P}\left(\hat{\mathcal{R}}IG\right) \geq 1 - 4\exp\left(-\frac{n\epsilon}{4}\right).$$

The event $\hat{R}IG$

Consider the two rectangles that result from drawing a line through R that is parallel to the x-axis. We call the upper rectangle an upper strip.

In an analogous way, we define lower, left, and right strips.

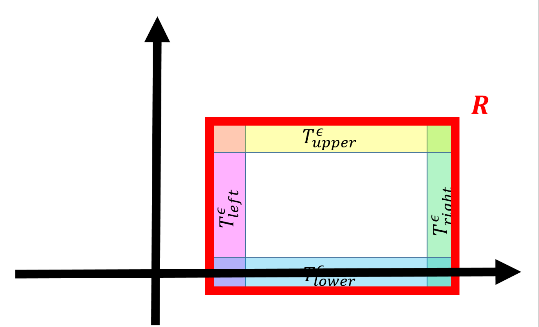


The event $\hat{R}IG$

Let T^{ϵ}_{upper} the upper strip such that $\mathbf{P}\left(T^{\epsilon}_{upper}\right) = \epsilon/4$.

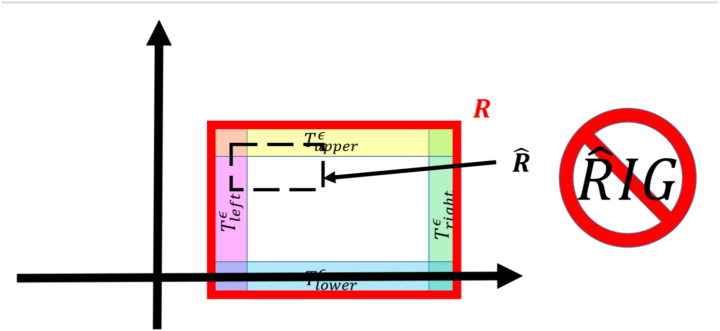
In an analogous way, we define T^{ϵ}_{lower}, T^{ϵ}_{left}, T^{ϵ}_{right}. We let

$$T^{\epsilon} := \bigcup_i T^{\epsilon}_i = T^{\epsilon}_{upper} \cup \ldots \cup T^{\epsilon}_{right}.$$



The event $\hat{\mathcal{R}}IG$

The event $\hat{\mathcal{R}}IG$ is the event that $\hat{R}$ intersects all four strips.



Tasks to do

1. Define $\hat{\mathcal{R}}IG$
2. Prove that

$$\mathbf{P}\left(\mathcal{R}(\hat{R}) \leq \epsilon\right) \geq \mathbf{P}\left(\hat{\mathcal{R}}IG\right) \geq 1 - 4\exp\left(-\frac{n\epsilon}{4}\right).$$

Proving the bound

See black board.

The universal concept class is not PAC-learnable

Let $\mathcal{X} = \{0, 1\}^*$ be the set of all finite binary sequences. The concept class $\mathcal{C}$ formed by all subsets of $\mathcal{X}$ is not PAC-learnable from $\mathcal{C}$.

The proof is hard though.

Consistent hypothesis and finite hypothesis classes

Let $\mathcal{C}$ be a finite concept class and assume that $\mathcal{H} = \mathcal{C}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in \mathcal{C}$ and any i.i.d. sample $\mathcal{Z}$ returns a consistent hypothesis $\hat{c}$ (i.e., $\hat{\mathcal{R}}_n(\hat{c}) = 0$). For any ϵ, $\delta > 0$, if

$$n \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right),$$

then

$$\mathbf{P}\left(\mathcal{R}(\hat{c}) \leq \epsilon\right) \geq 1 - \delta.$$

Proof

See black board.

The general stochastic setting

In general, an instance's label is not determined by the underlying concept. This is modeled with a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$. It reflects the fact that two instances with identical features may have different labels, like when two patients with very similar features show different reactions to the same drug.

The training dataset is therefore a sample $\mathcal{Z} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ from $\mathcal{D}$.

The goal then is to find a hypothesis $\hat{c} \in \mathcal{H}$ with small generalization error

$$\mathcal{R}(\hat{c}) = \mathbf{P}_{x,y\sim\mathcal{D}}(\hat{c}(x) \neq y) = \mathbb{E}_{x,y\sim\mathcal{D}}\left(\mathbf{1}_{\hat{c}(x)\neq y}\right).$$

It is now much harder to attain $\mathcal{R}(\hat{c}) \leq \epsilon$. Instead, we aim to attain $\mathcal{R}(\hat{c}) - \inf_{c\in\mathcal{C}} \mathcal{R}(c) \leq \epsilon$.

The general PAC learning model

A learning algorithm $\mathcal{A}$ can learn a concept class $\mathcal{C}$ from $\mathcal{H}$ if, given as input a sufficiently large sample, it outputs a hypothesis that generalizes well with high probability.

Definition

A learning algorithm $\mathcal{A}$ can learn a concept class $\mathcal{C}$ from $\mathcal{H}$ if there is a polynomial function *poly* $(\cdot, \cdot, \cdot)$ such that

1. for any distribution $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$ and
2. for any $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$,

if $\mathcal{A}$ receives as input a sample $\mathcal{Z}$ of size $n \geq \textit{poly}\,(1/\epsilon, 1/\delta, dim\,(\mathcal{X}))$, then $\mathcal{A}$ outputs $\hat{c} \in \mathcal{H}$ such that

$$\mathbf{P}_{\mathcal{Z} \sim \mathcal{D}^n} \left(\mathcal{R}\,(\hat{c}) - \inf_{c \in \mathcal{C}} \mathcal{R}\,(c) \leq \epsilon \right) \geq 1 - \delta.$$

Efficient PAC learning: If $\mathcal{A}$ runs in time polynomial in $1/\epsilon$ and $1/\delta$, we say that $\mathcal{A}$ is an efficient PAC learning algorithm.

Useful inequalities

Let $\epsilon > 0$. For a sample $\mathcal{Z} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, let $\hat{c}_n^*$ be the hypothesis obtained by empirical risk minimization:

$$\hat{c}_n^* = \arg\min_{c \in \mathcal{C}} \frac{1}{n} \left| \{(x_i, y_i) : c(x_i) \neq y_i, i \leq n\} \right| .$$

- If $\mathcal{C}$ is finite,

$$\mathbf{P}\left(\mathcal{R}\left(\hat{c}_n^*\right) - \inf_{c \in \mathcal{C}} \mathcal{R}\left(c\right) > \epsilon \right) \leq 2 \left|\mathcal{C}\right| \exp\left(-2n\epsilon^2\right) .$$

- The VC-dimension $VC_{\mathcal{C}}$ of a concept class $\mathcal{C}$ is a complexity measure for $\mathcal{C}$. If $VC_{\mathcal{C}} > 2$, then

$$\mathbf{P}\left(\mathcal{R}\left(\hat{c}_n^*\right) - \inf_{c \in \mathcal{C}} \mathcal{R}\left(c\right) > \epsilon \right) \leq 9n^{VC_{\mathcal{C}}} \exp\left(-\frac{n\epsilon^2}{32}\right) .$$

Concept classes with finite VC dimension are effective for learning

If $\mathcal{C}$ has a finite VC-dimension, then

$$\mathbf{P}\left(\mathcal{R}\left(\hat{c}_n^*\right) - \inf_{c \in \mathcal{C}} \mathcal{R}\left(c\right) > \epsilon\right) \leq 9 n^{VC_{\mathcal{C}}} \exp\left(-\frac{n\epsilon^2}{32}\right) \to 0, \text{as } n \to \infty.$$

VC-dimension

Definition

A set A of instances can be shattered by a concept class $\mathcal{C}$ if for every subset $S \subseteq A$, there is a concept $c_S \in \mathcal{C}$, such that $S = c_S \cap A$.

Examples:

- Any set of two points in $\mathbb{R}^2$ can be shattered by the class of axis-aligned rectangles.
- No set of three numbers in $\mathbb{R}$ can be shattered by the class of intervals.
- There is a set of three points in $\mathbb{R}^3$ that can be shattered by the class of axis-alligned rectangles. However, some sets of three points cannot be shattered by this class.

VC-dimension

Definition

The VC-dimension $VC_{\mathcal{C}}$ of a concept class $\mathcal{C}$ is computed as follows:

1. $n \leftarrow 1$
2. Is there a set of $n+1$ instances in $\mathcal{X}$ that can be shattered by $\mathcal{C}$?
3. If yes, then $n \leftarrow n+1$ and go to step 2.
4. Otherwise, $VC_{\mathcal{C}} = n$.

Contest

Open the Kahoot! website or download the app.

Link to the contest: `https://play.kahoot.it/#/k/908011b5-e067-4798-b16c-3644472aa64b`.

Conclusion

1. What is learnable?
2. PAC-learnability.
3. Useful inequalities to bound the probability $\mathbf{P}\left(\mathcal{R}\left(\hat{c}_n^*\right) - \inf_{c \in \mathcal{C}} \mathcal{R}\left(c\right) > \epsilon\right)$.

Bayes error and Bayes classifier

The Bayes error is

$$\mathcal{R}^* := \inf_f \mathcal{R}\left(f\right),$$

where f ranges over all measurable functions from $\mathcal{X}$ to $\{0,1\}$.

A Bayes classifier c^{Bayes} for $\mathcal{R}$ is a measurable function (not necessarily in the concept nor the hypothesis class) such that

$$\mathcal{R}\left(c^{Bayes}\right) = \mathcal{R}^*.$$

For $0/1$ loss, it holds that

$$c^{Bayes}(x) := \arg \max_{y \in \{0,1\}} \mathbf{P}\left(y \mid x\right).$$

Strong and Weak Learning

Assumption: Restricted classifier $c \in \mathcal{C}$ (hypothesis class).
Classification Error for empirically determined classifier $\hat{c}$

$$\mathbf{P}\left\{\mathcal{R}(\hat{c}) - \inf_{c\in\mathcal{C}} \mathcal{R}(c) > \epsilon\right\} < \delta \qquad (*)$$

Strong PAC Learning - Demand arbitrarily small error ϵ with high probability $1-\delta$.
Weak PAC Learning - Demand that $(*)$ holds for 'large' (but not trivial) error ϵ

- Example: Binary classification, require that

$$\epsilon \leq \frac{1}{2} - \gamma \qquad (\gamma > 0)$$

Weak learners are necessary to build ensemble classifiers, e.g., bagging classifiers as random forests or boosting methods like Adaboost.

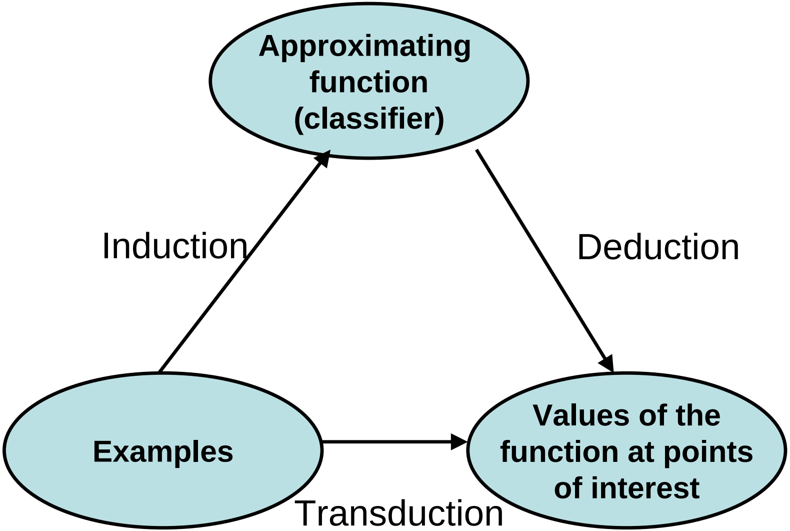
It is often assumed that $\mathcal{H} = \mathcal{C}$ in this setting. We make this assumption for the rest of the slides.

Statistical Learning Theory: Motivation

Vapnik (1982): Estimation of dependencies based on empirical data, Springer Verlag

“The method of [density estimation for] minimizing the risk generally is not reasonable, because the problem of density estimation is a more difficult problem than the minimization of the expected risk. Only when a substantial prior information is available about the desired density $\mathbf{P}(x, y)$, so that the function $\mathbf{P}(x, y)$ can be defined up to its parameters, is this approach plausible.” (Chap. 6 p. 139)

Relation of Inference Principles


Approximating function (classifier)
Induction
Deduction
Examples
Values of the function at points of interest
Transduction

Mathematics and its Effectiveness

I. M. Gelfand

... a famous mathematician who worked in biomathematics and molecular biology, as well as many other fields in applied mathematics is quoted as stating:

Eugene Wigner wrote a famous essay on the unreasonable effectiveness of mathematics in natural sciences. He meant physics, of course. There is only one thing which is more unreasonable than the unreasonable effectiveness of mathematics in physics, and this is the unreasonable ineffectiveness of mathematics in biology.

`http://en.wikipedia.org/wiki/Unreasonable_ineffectiveness_of_mathematics`

`http://www3.interscience.wiley.com/cgi-bin/fulltext/113397477/PDFSTART`

Classifier Selection

Induction Principle: **Empirical Risk Minimization**

Select the classifier $\hat{c}_n^\star \in \mathcal{C}$ with the smallest error on the training data $\mathcal{Z} = \{(x_1, y_1), \dots, (x_n, y_n)\}$:

$$\hat{c}_n^\star = \arg\min_{c \in \mathcal{C}} \underbrace{\frac{1}{n} \#\{(x_i, y_i) : c(x_j) \neq y_j,\ 1 \leq i \leq n\}}_{\text{training error } \hat{\mathcal{R}}_n(c)}$$

Empirical Classification Error $\hat{\mathcal{R}}_n(c) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{\{c(x_j) \neq y_j\}}$

can be measured without any a priori assumptions!

Expected Classification Error $\mathcal{R}(c) = \mathbf{P}\{c(X) \neq Y\}$

is the quality measure which we care about.

Goal: Derive a distribution independent bound for the probability of large deviations between the expected risk of the ERM classifier and the optimal classifier

$$\mathbf{P}\{\mathcal{R}(\hat{c}_n^\star) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon\}$$

Generalization Error: $\mathcal{R}(\hat{c}_n^\star) = \mathbf{P}\{\hat{c}_n^\star(X) \neq Y | \mathcal{Z}\}$

Remark: The generalization error $\mathcal{R}(\hat{c}_n^\star)$ of the trained classifier should not exceed the minimal generalization $\inf_{c \in \mathcal{C}} \mathcal{R}(c)$ achievable in class $\mathcal{C}$ by more than ϵ!

Problem: We cannot measure $\mathcal{R}(\hat{c}_n^\star)$!

Question: Can we give distribution independent bounds on this large deviation by selecting a specific classifier set $\mathcal{C}$?

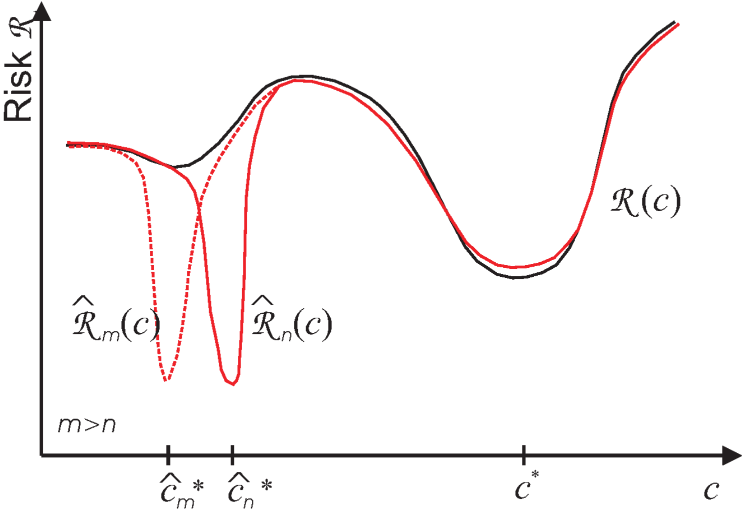
VC Inequality and Uniform Convergence

Theorem 1 *(Vapnik Chervonenkis 1974)* *($c^\star = \operatorname{argmin}_{c \in \mathcal{C}} \mathcal{R}(c)$)*

$$
\begin{aligned}
\mathcal{R}(\hat{c}_n^\star) \; & - \; \inf_{c \in \mathcal{C}} \mathcal{R}(c) \\
& = \; \mathcal{R}(\hat{c}_n^\star) - \hat{\mathcal{R}}_n(\hat{c}_n^\star) + \hat{\mathcal{R}}_n(\hat{c}_n^\star) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) \\
& \leq \; \underbrace{\mathcal{R}(\hat{c}_n^\star) - \hat{\mathcal{R}}_n(\hat{c}_n^\star)}_{} + \underbrace{\hat{\mathcal{R}}_n(c^\star) - \mathcal{R}(c^\star)}_{} \\
& \leq \; \sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| + \sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| \\
& \leq \; 2 \sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)|
\end{aligned}
$$

Bound on suboptimality of $\hat{c}_n^\star$: require uniform convergence!

$$
\mathbf{P}\{\mathcal{R}(\hat{c}_n^\star) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon\} \;\leq\; \mathbf{P}\{\sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \epsilon/2\}
$$



It often holds that $\hat{\mathcal{R}}_n(\hat{c}_n^\star) < \mathcal{R}(c^\star), \ \forall n < \infty$! (Any idea why?)

Reason for “Uniform Convergence”

We distinguish between two different situations:

1. A classifier c is selected **before** we see the training data $\mathcal{Z} = \{(x_i, y_i)\}$. Then the empirical classification risk of c converges against its expected classification risk due to the law of large numbers.
2. A classifier $\hat{c}_n$ is selected according to an induction principle like ERM **after** we have seen the data $\mathcal{Z}$.

   The law of large numbers does not apply in this situation since we select a different classifier $\hat{c}_n$ for each sample set $\mathcal{Z}_n$ which is used in the sequence $(n \to \infty)$ of classification problems.

Hoeffding's Inequality (1963)

Lemma 2 *(Markov Inequality) Let X be a non-negative random variable. Then*

$$\mathbf{P}\{X \geq \epsilon\} \leq \frac{\mathbb{E}[X]}{\epsilon}$$

Lemma 3 *Let X be a random variable with $\mathbb{E}[X] = 0$ and $a \leq X \leq b$. Then for $s > 0$ it holds*

$$\mathbb{E}[\exp(sX)] \leq \exp(s^2(b-a)^2/8) \ .$$

Proof: blackboard

Hoeffding's Theorem

Theorem 4 *(Hoeffding 1963)*
Let $X_1, \ldots, X_n$ be independent bounded random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one and let $S_n = \sum_{i=1}^n X_i$. Then for any $t > 0$ we have

$$\begin{aligned}
\mathbf{P}\{S_n - \mathbb{E}S_n \geq t\} \quad &\leq \quad \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
\mathbf{P}\{S_n - \mathbb{E}S_n \leq -t\} \quad &\leq \quad \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad .
\end{aligned}$$

Proof

Let s be an arbitrary positive number. Then it holds

$$\mathbf{P}\{X \geq t\} = \mathbf{P}\{\exp(sX) \geq \exp(st)\} \overset{\text{Markov}}{\leq} \frac{\mathbb{E}[\exp(sX)]}{\exp(st)}$$

Idea: find an appropriate s such that the bound is minimized. ($S_n = \sum_{i \leq n} X_i$)

$$\begin{aligned}
\mathbf{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-st}\mathbb{E}\Big[\exp(s\sum_{i\leq n}(X_i - \mathbb{E}X_i))\Big] \\
&= e^{-st}\prod_{i\leq n}\mathbb{E}[\exp(s(X_i - \mathbb{E}X_i))] && \text{independence} \\
&\leq e^{-st}\prod_{i\leq n}\exp(s^2(b_i - a_i)^2/8) && \text{Lemma} \\
&= \exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right) \\
\text{with} \quad s &= \frac{4t}{\sum_i (b_i - a_i)^2}
\end{aligned}$$

Analog proof for $\mathbf{P}\{S_n - \mathbb{E}S_n \leq -t\}$.

Special Case for Normalized Sums of iid R.V.'s

Normalized sums $\tilde{S} = S_n/n, \; t = n\epsilon$

$$\mathbf{P}\{\tilde{S}_n - \mathbb{E}\tilde{S}_n \geq \epsilon\} \leq \exp\left(-\frac{2n\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2/n}\right) \overset{n\to\infty}{\longrightarrow} 0$$

Absolute Deviation

$$\begin{aligned}
\mathbf{P}\{|\tilde{S}_n - \mathbb{E}\tilde{S}_n| \geq \epsilon\} &= \mathbf{P}\{\tilde{S}_n - \mathbb{E}\tilde{S}_n \geq \epsilon \vee \tilde{S}_n - \mathbb{E}\tilde{S}_n \leq -\epsilon\} \\
&= \mathbf{P}\{\tilde{S}_n - \mathbb{E}\tilde{S}_n \geq \epsilon\} + \mathbf{P}\{\tilde{S}_n - \mathbb{E}\tilde{S}_n \leq -\epsilon\} \\
&\leq 2\exp\left(-\frac{2n\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2/n}\right)
\end{aligned}$$

Remark: The Hoeffding bound for binomial random variables is also called Chernoff or Okamoto bound.

Error Bound for Finite Classifier Sets

Union Bound

$\mathbf{P}\{\bigcup_{i=1}^{\infty} A_i\} \le \sum_{i=1}^{\infty} \mathbf{P}\{A_i\}$ (shown graphically by drawing the Venn diagram of $\bigcup_i A_i$). Equality holds for partitions, i.e., if and only if $\forall i, j \; i \neq j \; A_i \cap A_j = \emptyset$.

Example: let $\mathcal{X} = \{X_1, \ldots, X_n\}, \; X_i \in \mathbb{R}$ be a set of random numbers. Then $\mathbf{P}\{\max_i X_i > \epsilon\} = \mathbf{P}\{X_1 > \epsilon \vee \cdots \vee X_n > \epsilon\} \le \sum_i \mathbf{P}\{X_i > \epsilon\}$.

Theorem 5 *Assume that the cardinality of $\mathcal{C}$ is bounded by N. Then we have for all $\epsilon > 0$*

$$\mathbf{P}\left\{\sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \epsilon\right\} \le 2N \exp(-2n\epsilon^2) \; .$$

Proof:

$$\mathbf{P}\left\{\sup_{c\in\mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \epsilon\right\} \overset{\text{(a)}}{\leq}$$

$$\sum_{c\in\mathcal{C}} \mathbf{P}\left\{|\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \epsilon\right\} \overset{\text{(b)}}{\leq} 2N\exp(-2n\epsilon^2)$$

(a) Union Bound
(b) Hoeffding's inequality and $\hat{\mathcal{R}}_n(c) \in [0,1]$ implies that for all i it holds $b_i = 1, a_i = 0$

Remark: Define the confidence $\delta \equiv 2N\exp(-2n\epsilon^2)$. Then the precision is

$$\epsilon = \sqrt{\frac{\ln N - \ln(\delta/2)}{2n}}$$

Confidence Interval: The inequality $\mathcal{R}(c) - \hat{\mathcal{R}}_n(c) \le \epsilon$ holds with high probability for all functions c!

$$\Rightarrow \quad \underbrace{\mathcal{R}(c)}_{\text{expected error}} \le \underbrace{\hat{\mathcal{R}}_n(c)}_{\text{empirical error}} + \underbrace{\sqrt{\frac{\ln N - \ln(\delta/2)}{2n}}}_{\text{variance}}$$

holds with probability $1 - \delta$. A large number of functions increases the variance logarithmically!

The expected error of the classifier c is bounded by its empirical error and a variance term which depends on sample size n as $1/\sqrt{n}$ but only logarithmically on the size of the hypothesis class.

Empirical Risk Minimization for Hyperplanes

The case of a finite hypothesis class can be generalized in two ways:

1. A hypothesis class with infinite cardinality is represented by finitely many hypotheses which yield different classifications on the data.
2. Measure the Vapnik-Chervonenkis ($\mathcal{VC}$) dimension of a set of functions and select a hypothesis class $\mathcal{H}$ with

$$\dim_{\mathcal{VC}}(\mathcal{H}) < \infty$$

Assumption: Samples are elements in a d-dimensional Euclidean space with $x \sim \mathbf{P}(x)$ (density exists).

Empirical Risk Minimization for Hyperplanes

Hypotheses: $\sum_{i=1}^{d} a_i x_i + a_0 = 0$ (set of all hyperplanes).

Idea: Select $\binom{n}{d}$ hyperplanes from the infinite set of all hyperplanes and estimate their error rate.
$\Rightarrow$ quantization of the hypothesis space (fingering argument).

Construction: Consider d arbitrary samples from $\mathcal{X}$. The points are in general position with probability 1 since we assume a density.
The hyperplane defined by these d points represents two classifiers

$$c_\alpha(x) = \begin{cases} 1 & a^T x + a_0 \overset{>}{\underset{\leq}{}} 0 \\ 0 & \text{otherwise} \end{cases} \qquad \alpha \in \{1, 2\}$$

$\#$ of possible classifiers equals $2 \cdot \#$ of hyperplanes: $2\binom{n}{d}$

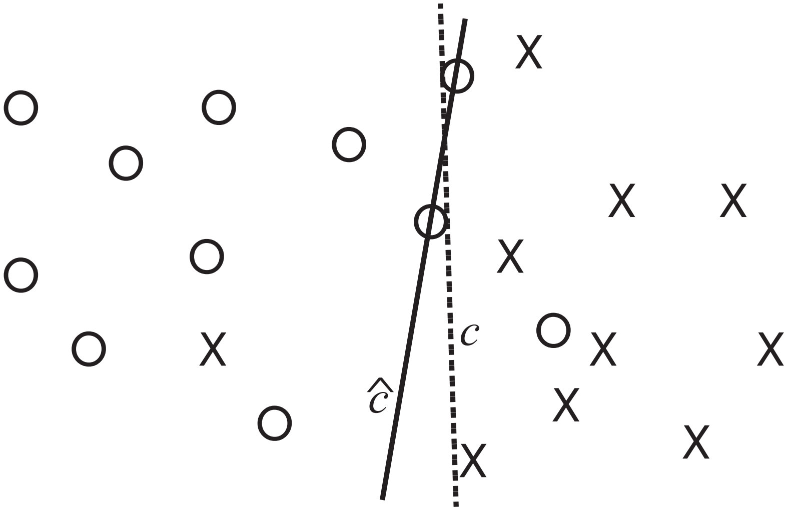
since every possible d tuple $(x_{i_1}, \ldots, x_{i_d}) \in \mathcal{X}^d$ has two classifiers associated.

Best Empirical Classifier, which is defined by d samples:

$$\hat{c} = \arg \min_{i=1,\ldots,2\binom{n}{d}} \hat{\mathcal{R}}_n(c_i)$$

Let c be any linear classifier. Then there exists a hyperplane containing d points such that the empirical error deviates by no more than $\frac{d}{n}$, i.e.,

$$\forall\, c \in \mathcal{C} \quad \hat{\mathcal{R}}_n(c) \geq \hat{\mathcal{R}}_n(\hat{c}) - \frac{d}{n}$$



"Empirical" classifications of c and $\hat{c}$ are the same except for the d data points on the plane!

Theorem 6 *Assume that X has a density. If $\hat{c}$ is found by empirical error minimization, then, for all possible distributions of (X, Y), if $n \geq d$ and $2\frac{d}{n} \leq \epsilon \leq 1$, we have*

$$\mathbf{P}\left\{\mathcal{R}(\hat{c}) > \inf_{c \in \mathcal{C}} \mathcal{R}(c) + \epsilon\right\} \leq e^{2d\epsilon}\left(2\binom{n}{d} + 1\right)\exp\left(-\frac{n\epsilon^2}{2}\right)$$

Moreover, if $n \geq d$, then

$$\mathbb{E}[\mathcal{R}(\hat{c}) - \mathcal{R}] \leq \sqrt{\frac{2}{n}(d+1)(\log n + 2)}$$

Proof: see blackboard

Remark: The factor $\binom{n}{d}$ grows polynomially for constant d! Learnability is shown since the probability of an ϵ-large deviation of the empirical from the expected risk vanishes for large sample size without assuming a specific distribution.

Classifiers with Vanishing Error

Theorem 7 *Assume that X has a density, and that the best linear classifier has zero error probability ($\mathcal{R}(c^\star) = 0$). Then for the ERM algorithm and for $n > d,\ \epsilon \leq 1$, it holds*

$$\mathbf{P}\left\{\mathcal{R}(\hat{c}_n) > \epsilon\right\} \leq 2\binom{n}{d}\exp(-\epsilon(n-d))$$

Remark

a) Improved convergence is achieved since $\exp(-n\epsilon)$ rather than $\exp(-n\epsilon^2)$ holds.
b) Given are n labeled points in $\mathbb{R}^d$. To find the best dichotomy is $\mathcal{NP}$-hard (Johnson & Preparata, 1978), i.e., to determine $\hat{c}$ is computationally difficult in the worst case.

Proof

We use the fingering argument of the previous theorem [6], the union bound (a) and the symmetry of classifiers (b).

$$
\begin{aligned}
\mathbf{P}\left\{\mathcal{R}(\hat{c}_n) > \epsilon\right\} &\le \mathbf{P}\left\{ \max_{i=1,\dots,2\binom{n}{d}\,:\,\hat{\mathcal{R}}(c_i)\le\frac{d}{n}} \mathcal{R}(c_i) > \epsilon \right\} \\
&\overset{(a)}{\le} \sum_{i=1}^{2\binom{n}{d}} \mathbf{P}\left\{ \hat{\mathcal{R}}_n(c_i) \le \frac{d}{n} \ \wedge\ \mathcal{R}(c_i) > \epsilon \right\} \\
&\overset{(b)}{=} 2\binom{n}{d}\mathbb{E}\Bigg[\underbrace{\mathbf{P}\left\{ \hat{\mathcal{R}}_n(c_1) \le \frac{d}{n} \ \wedge\ \mathcal{R}(c_1) > \epsilon \,\middle|\, X_1,\dots,X_d \right\}}_{\le \mathbf{P}\{c_1(X_j)=Y_j, d+1\le j\le n \ \wedge\ \mathcal{R}(c_1)>\epsilon | X_1,\dots,X_d\}}\Bigg] \\
&\le 2\binom{n}{d}(1-\epsilon)^{n-d} \le 2\binom{n}{d}\exp(-\epsilon(n-d)) \qquad \square
\end{aligned}
$$

This argument holds since errors are only made on the 1^{st} d samples and the probability of zero additional errors is less than $(1-\epsilon)^{n-d}$.