Statistical Data Analysis

Jana de Wiljes

December 5, 2022

VC Dimension

Problem setting

Goal: Approximate function f, that describes the link between two random variables X and Y which have the joint distribution $\pi(z) = \pi(x, y)$

Choice of parametrisation:

- choose model class $\mathcal{H}$
- and appropriate loss functional $l(y, h(x))$

Expected Risk

For $h \in \mathcal{H}$ we define the expected Risik as follows

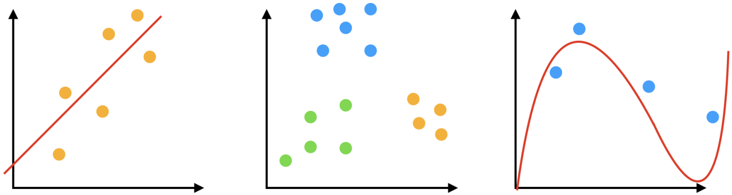
$$R(h) = \int_{\mathbf{Z}} l(y, h(x))\pi(z)dz \tag{1}$$

Approach: Want to find $h \in \mathcal{H}$ so that

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \tag{2}$$

Empirical Risk

Given in practice: independent and identical distributed Samples
$S = \{(x_i, y_i)\}_{i=1}^N$ with $(x_i, y_i) \sim \pi(x, y)$ for $i \in \{1, \dots, N\}$



Empirical Risk

For a given sample set S we define the corresponding empirical risk as follows:

$$R_S(h) = \frac{1}{N} \sum_{i=1}^{N} l(y_i, h(x_i))$$

Empirical Risk-Minimizer

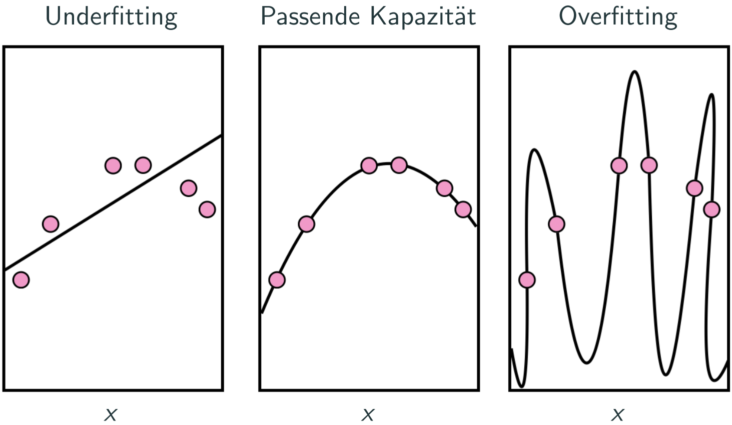
Empirical Risk-Minimizer

A learning algorithm $\hat{h}_N$ with $S = \{(x_i, y_i)\}_{i=1}^N$ where $(x_i, y_i) \sim \pi(x, y)$ of the form

$$\hat{h}_N \in \arg\min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(y_i, h(x_i))$$

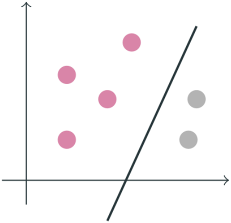
is called Empirical Risk-Minimizer.

Generalisability



Goal: want to find hypothesis class $\mathcal{H}$, so that $R_N(h) = 0$ with $h \in \mathcal{H}$ implies that $R(h)$ is small

Supervised classification



Consider:

- $f : \mathcal{X} \to \{0, 1\}$
- $f(x) = y$
- $\pi(x, y) = \begin{cases} \pi(x) & \text{für } f(x) = y \\ 0 & \text{für } f(x) \neq y \end{cases}$

Choose:

- Hypothesis class $\mathcal{H}$ with $h : \mathcal{X} \to \{0, 1\}$ for all $h \in \mathcal{H}$
- $0 - 1$ loss functional, i.e., $R_N(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{h(x_i) \neq y_i}$

Probably approximately correct learning

PAC-lernbar

We say that the hypothesis class $\mathcal{H}$ is PAC-learnable if there is a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm h_S such that for any $\epsilon, \delta \in (0,1)$, for every distribution π on $\mathcal{X}$ and for every labelling function $f : \mathcal{X} \to \{0,1\}$ if the realizable assumption holds with respect to $S = \{(X_i, f(X_i))\}_{i=1}^N$ and $(X_i)_{1\leq N} \sim \pi(x)$,

$$R(h_S) \geq \epsilon$$

with probability smaller δ for all $N \geq m_{\mathcal{H}}(\epsilon, \delta)$.

Remark: The smallest possible function $m_{\mathcal{H}}(\epsilon, \delta)$, to learn $\mathcal{H}$ is called sample complexity.

PAC Lernen

Theorem: Let $|S| = N$, $\epsilon, \delta \in (0,1)$ and $\mathcal{H}$ is finite and realizable . If $N \geq \frac{1}{\epsilon} ln(\frac{|\mathcal{H}|}{\delta})$, then

$$\mathbb{P}(R(h_S) < \epsilon) \geq 1 - \delta$$

holds for all $h \in \mathcal{H}$ with $R_S(h) = 0$ and iid samples S:

Sketch of proof: $R_S(h_S) = 0$ holds since $\mathcal{H}$ is realizable und ERM

$$\begin{aligned}
\mathbb{P}(R(h_S) \geq \epsilon) &= \mathbb{P}(\{S \in \mathcal{X}^N : \exists h \in \mathcal{H}, R_S(h) = 0 \text{ und } R(h) \geq \epsilon\}) \\
&= \mathbb{P}\Big(\bigcup_{h:R(h)\geq\epsilon} S_h\Big) \text{ where } S_h = \{S \in \mathcal{X}^N : R_S(h) = 0\} \\
&\leq \sum_{h:R(h)\geq\epsilon} \mathbb{P}(S_h) \quad \text{(Bonferroni-inequality)} \\
&\leq \sum_{h:R(h)\geq\epsilon} \prod_{i=1}^{N} \underbrace{\pi(\{x \in \mathcal{X} : h(x) = f(x)\})}_{1-R(h)} \quad \text{(iid)} \\
&\leq \sum_{h:R(h)\geq\epsilon} (1-\epsilon)^N \leq \underbrace{|\mathcal{H}|(1-\epsilon)^N \leq |\mathcal{H}| \exp(-N\epsilon)}_{1-x\leq exp(-x)} \leq \delta
\end{aligned}$$

Upper bound

Hoeffding-inequality: Let $\bar{X} = (X_1 + \cdots + X_N)/N$ with $X_i \in [0,1]$ iid, then:

$$\mathbb{P}(|\bar{X} - \mathbb{E}[X_i]| \geq \epsilon) \leq 2\exp\Big(-2N\epsilon^2\Big)$$

Theorem: Let $|S| \geq \frac{1}{2\epsilon^2}\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$ and $\mathcal{H}$ finite, then the following holds for all $h \in \mathcal{H}$:

$$\mathbb{P}(|R(h) - R_S(h)| < \epsilon) \geq 1 - \delta$$

Sketch of the proof:

- $\mathbb{1}_{h(X_i)\neq Y_i} \in [0,1]$ and $X_i \sim \pi(X)$ iid
- Bonferroni-inequality and Hoeffding-inequality

$$\mathbb{P}(|R(h) - R_S(h)| \geq \epsilon) \leq \underbrace{2|\mathcal{H}|\exp\Big(-2N\epsilon^2\Big)}_{\delta}$$

- reorder according to $N = |S|$

Shattering

Restriction

Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The restriction of $\mathcal{H}$ to C is the set $\mathcal{H}_C$ of functions from $C \to \{0,1\}$ that can be derived from $\mathcal{H}$

$$\mathcal{H}_C = \Big\{(c_1, \ldots, c_m) \to (h(c_1), \ldots, h(c_m))\Big\} \tag{3}$$

Shattering

A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if $\mathcal{H}_C = \{0,1\}^C$.

Vapnik-Chervonenkis Dimension

VC Dimension

The Vapnik Chervonenkis dimension **VCdim**$(\mathcal{H})$ of a hypothesis class $\mathcal{H}$ is the maximal size of a set $C \subset \mathcal{X}$, that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets $C \subset \mathcal{X}$ of arbitrarily large size we say that **VCdim**$(\mathcal{H}) = \infty$.

Example:

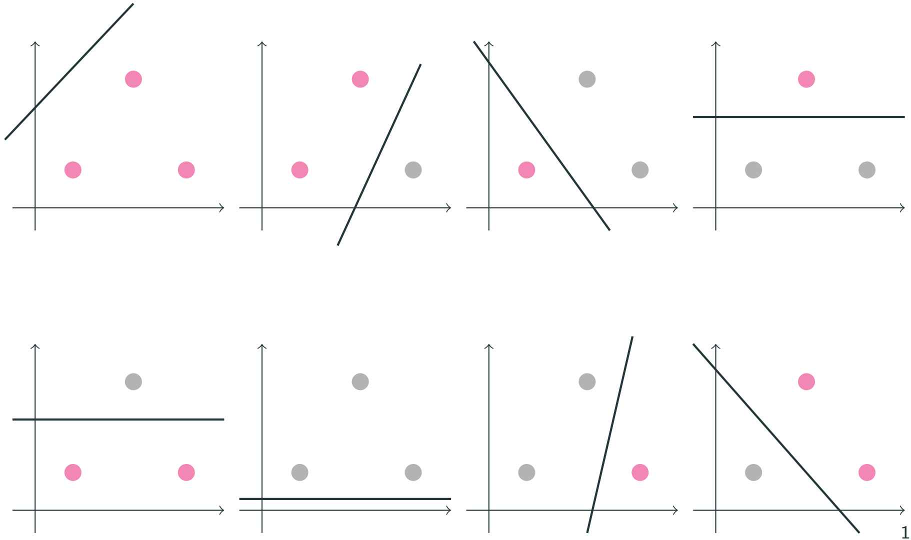
Consider $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{H} = \left\{ h_{w,\theta}(x) = \begin{cases} 1 & \text{if } w^\top x \geq \theta \\ 0 & \text{if } w^\top x < \theta \end{cases} \middle| w \in \mathbb{R}^2, \theta \in \mathbb{R} \right\}$

$C = \{(1,0), (2,0), (3,0)\}$

Not possible to shatter with $\mathcal{H}$

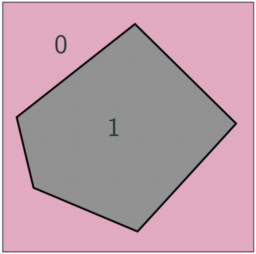
Example

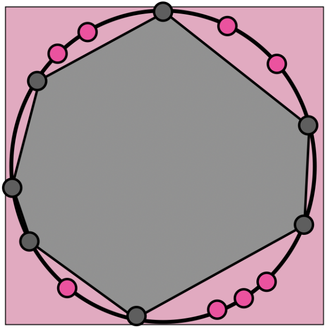
Three points in $\mathbb{R}^2$ shattered by $\mathcal{H}$



Example infinite dimensional

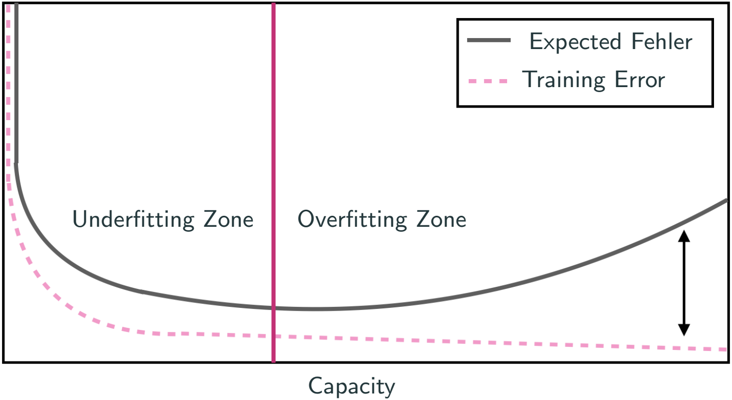
Hypothesis class $\mathcal{H}$:
$h : [0,1]^2 \to \{0,1\}$, are functions that have value 1 on a convex Polygon and have the value 0 outside of the polygon.





VC-Dimension of $\mathcal{H}$ is **infinite**: Choose m arbitrary points on the circle in $[0,1]^2$. For arbitrary y_i one connects the points on the circle that have label 1 to form a convex polygon.

Underfitting and Overfitting Dilemma



Theorem: Let $\mathcal{H}$ be a hypothesis class with $\textbf{VCdim}(\mathcal{H}) = \infty$. Then $\mathcal{H}$ is not PAC-learnable.

Sauer's Lemma

Growth function

Let $\mathcal{H}$ be a hypothesis class. The growth function $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ of $\mathcal{H}$ is defined via

$$\tau_{\mathcal{H}}(N) = \max_{C \subset \mathcal{X} : |C| = N} |\mathcal{H}_C| \tag{4}$$

Sauer's Lemma: Let $\mathcal{H}$ be a hypothesis class with $\mathbf{VCdim}(\mathcal{H}) \leq d < \infty$. Then

$$\tau_{\mathcal{H}}(N) \leq \sum_{i=0}^{d} \binom{N}{i} \tag{5}$$

for $N \geq d$. In particular $\tau_{\mathcal{H}}(N) \leq \left(\frac{eN}{d}\right)^d = \mathcal{O}(N^d)$ for $N > d + 1$.

Sauer's Lemma proof

Proof: In fact we prove a stronger claim

$$|\mathcal{H}_C| \leq |\{B \subset C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{N}{i}.$$

where the last inequality holds since no set of size larger than d is shattered by $\mathcal{H}$. The proof is done by induction.

$\mathbf{N = 1}$: The empty set is always considered to be shattered by $\mathcal{H}$. Hence, either $|\mathcal{H}(C)| = 1$ and $d = 0$, inequality then states $1 \leq 1$ or $d \leq 1$ and the inequality is $2 \leq 2$.

Induction: Let $C = \{c_1, \dots, c_N\}$ and let $C' = \{c_2, \dots, c_N\}$. We note functions like vectors, and we define

$$Y_0 = \{(y_2, \dots, y_N) : (0, y_2, \dots, y_N) \in \mathcal{H}_C \text{ or } (1, y_2, \dots, y_N) \in \mathcal{H}_C\}, \text{ and}$$
$$Y_1 = \{(y_2, \dots, y_N) : (0, y_2, \dots, y_N) \in \mathcal{H}_C \text{ and } (1, y_2, \dots, y_N) \in \mathcal{H}_C\}$$

Proof

Then $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Moreover $Y_0 = \mathcal{H}_{C'}$ an hence by the induction hypothesis:

$$\begin{aligned} |Y_0| \le |\mathcal{H}_{C'}| &\le |\{B \subset C' : \mathcal{H} \text{ shatters } B\}| \\ &= |\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| \end{aligned}$$

Next, define

$$\mathcal{H}' = \left\{ h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } h'(c) = \begin{cases} 1 - h(c) & \text{if } c = c_1 \\ h(c) & \text{otherwise} \end{cases} \right\}$$

Note that $\mathcal{H}'$ shatters $B \subset C'$ iff $\mathcal{H}'$ shatters $B \cup \{c_1\}$, and that $Y_1 = \mathcal{H}'_{C'}$. Hence, by the induction hypothesis,

$$\begin{aligned} |Y_1| = |\mathcal{H}'{}_{C'}| \le |\{B \subset C' : \mathcal{H}' \text{ shatters } B\}| = |\{B \subset C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}| \\ = |\{B \subset C : c_1 \text{ and } \mathcal{H}' \text{ shatters } B\}| \le |\{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| \end{aligned}$$

Proof

Overall

$$
\begin{aligned}
|\mathcal{H}_C| &= |Y_0| + |Y_1| \\
&\leq |\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}| + \{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}| \\
&= |\{B \subset C : \mathcal{H} \text{ shatters } B\}|
\end{aligned}
$$

For the last inequality, one may observe that if $N \geq 2d$, defining $X \sim \mathcal{B}(N, 1/2)$, Chernoff inequality and inequality $\log(u) \geq (u-1)/2$ yield

$$
\begin{aligned}
-\log \mathbb{P}(X \leq d) \geq Nkl(\frac{d}{N}, \frac{1}{2}) &\geq d \log\Big(\frac{2d}{N} + (N-d)\log\Big(\frac{2(N-d)}{N}\Big)\Big) \\
&\geq N \log 2 + d \log(\frac{d}{N}) + (N-d)\frac{-d/N}{(N-d)/N} \\
& N \log(2) + d \log(\frac{d}{eN})
\end{aligned}
$$

and hence

$$
\sum_{i=0}^{d} \binom{N}{i} = 2^d \mathbb{P}(X \leq d) \leq \exp\Big(-d \log(\frac{d}{eN})\Big) = \Big(\frac{eN}{d}\Big)
$$

Proof

Besides, for the case $d \leq N \leq 2d$, the inequality is obvious since $(eN/d)^d \geq 2^N$: indeed, function $f : x \mapsto -x \log(x/e)$ is increasing on $[0, 1]$, and hence for all $d \leq m \leq 2d$:

$$\frac{d}{N} \log \frac{eN}{d} = f(\frac{d}{N}) \geq f(1/2) = \frac{1}{2} \log(2e) \geq \log(2),$$

which implies

$$\left(\frac{eN}{d}\right)^d = \exp\left(d \log(\frac{eN}{d})\right) \geq \exp(N \log(2)) = 2^N$$

Alternately, you may simply observe that for all $N \geq d$,

$$\left(\frac{d}{N}\right)^d \sum_{i=0}^{d} \binom{N}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{N}\right)^i \binom{N}{i} \leq \sum_{i=0}^{N} \left(\frac{d}{N}\right)^i \binom{N}{i} = \left(1+\frac{d}{N}\right)^N \leq e^d$$

$\square$

Finite VC dimension

Theorem: Let $\mathcal{H}$ be a hypothesis class with $\textbf{VCdim}(\mathcal{H}) \leq d < \infty$. Then the following holds for $\delta \in (0, 1)$

$$R(h) \leq R_S(h) + \underbrace{\mathcal{O}\Big(\sqrt{\frac{d}{N}\log(N/d) - \frac{1}{N}\log(\delta)}\Big)}_{\mathcal{O}(\epsilon)} \tag{6}$$

with probability of at least $1 - \delta$.

Note: this result is sufficient to prove that finite VC-dim $\implies$ learnable, but the dependency in δ is not correct at all: roughly speaking, the factor $1/\delta$ can be replaced by $\log(1/\delta)$.

Finite VC dimension Proof

We consider the $0-1$ loss, or any $[0,1]$- valued loss. Observe that $R_\pi(h) = \mathbb{E}[R_{S'}(h)]$ where $S' = z_1', \ldots, z_m'$ is another iid sample of π. Hence,

$$\mathbb{E}_S\Big[\sup_{h\in\mathcal{H}} |R_\pi(h) - R_S(h)|\Big] = \mathbb{E}_S\Big[\sup_{h\in\mathcal{H}} |R_{S'}(h) - R_S(h)|\Big] \le \mathbb{E}_S\Big[\sup_{h\in\mathcal{H}} \Big|\mathbb{E}_{S'}[R_{S'}(h) - R_S(h)]\Big|\Big]$$

$$\le \mathbb{E}_S\Big[\sup_{h\in\mathcal{H}} \Big[\mathbb{E}_{S'}|R_{S'}(h) - R_S(h)|\Big]\Big] \le \mathbb{E}_S\Big[\mathbb{E}_{S'}\Big[\sup_{h\in\mathcal{H}} |R_{S'}(h) - R_S(h)|\Big]\Big]$$

$$= \mathbb{E}_{S,S'}\Big[\sup_{h\in\mathcal{H}} \frac{1}{m}\Big|\sum_{i=1}^m l(h,z_i') - l(h,z_i)\Big|\Big]$$

$$= \mathbb{E}_{S,S'}\Big[\sup_{h\in\mathcal{H}} \frac{1}{m}\Big|\sum_{i=1}^m \sigma_i(l(h,z_i') - l(h,z_i))\Big|\Big] \quad \text{for all } \sigma \in \{\pm 1\}^m$$

$$= \mathbb{E}_\Sigma\mathbb{E}_{S,S'}\Big[\sup_{h\in\mathcal{H}} \frac{1}{m}\Big|\sum_{i=1}^m \Sigma_i(l(h,z_i') - l(h,z_i))\Big|\Big] \quad \text{for all } \Sigma \sim \mathcal{U}\{\pm 1\}^m$$

$$= \mathbb{E}_{S,S'}\mathbb{E}_\Sigma\Big[\sup_{h\in\mathcal{H}} \frac{1}{m}\Big|\sum_{i=1}^m \Sigma_i(l(h,z_i') - l(h,z_i))\Big|\Big]$$

Now, for every S, S', let $C = C_{S,S'}$ be the instances appearing in S and S'. Then

$$\sup_{h\in\mathcal{H}} \frac{1}{m}\Big|\sum_{i=1}^m \Sigma_i(l(h,z_i') - l(h,z_i))\Big| = \max_{h\in\mathcal{H}_C} \frac{1}{m}\Big|\sum_{i=1}^m \Sigma_i(l(h,z_i') - l(h,z_i))\Big|.$$

Finite VC dimension Proof

Moreover, for every $h \in \mathcal{H}_C$ let $Z_h = \frac{1}{m}\sum_{i=1}^m \Sigma_i(l(h,z_i') - l(h,z_i))$. Then $\mathbb{E}_\Sigma[Z_h] = 0$, each summand belongs to $[-1,1]$ and by Hoeffding's inequality, for every $\epsilon > 0$:

$$\mathbb{P}_\Sigma\big[|Z_h| \geq \epsilon\big] \leq 2\exp\Big(-\frac{m\epsilon^2}{2}\Big)$$

Hence by the union bound

$$\mathbb{P}_\Sigma\big[\max_{h\in\mathcal{H}_C} |Z_h| \geq \epsilon\big] \leq 2|\mathcal{H}_C|\exp\Big(-\frac{m\epsilon^2}{2}\Big)$$

The following lemma permits to deduce that

$$\mathbb{E}_\Sigma\big[\max_{h\in\mathcal{H}_C} |Z_h|\big] \leq \frac{1+\sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{m/2}} \leq \frac{1+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{m/2}}$$

Hence,

$$\begin{aligned}\mathbb{E}_S\Big[\sup_{h\in\mathcal{H}} |R_\pi(h) - R_S(h)|\Big] &= \mathbb{E}_{S,S'}\mathbb{E}_\Sigma\Big[\sup_{h\in\mathcal{H}} \frac{1}{m}\Big|\sum_{i=1}^m \Sigma_i(l(h,z_i') - l(h,z_i))\Big|\Big] \\ &\leq \frac{1+\sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{m/2}}\end{aligned}$$

and we conclude by using Markov inequality (poor idea! Better: McDiarmid inequality). □

Auxiliary Lemma

Lemma: Let $a > 0$, $b > 0$, and let Z be a real-valued random variable such that for all $t \geq 0$,

$$\mathbb{P}(Z \geq 0) \leq 2b \exp\Big(-\frac{t^2}{a^2}\Big)$$

Then

$$\mathbb{E}[Z]\Big(\sqrt{\log(b)} + \frac{1}{\sqrt{\log(b)}}\Big)$$

Literature

- **Original Papers**
  1. L. G. Valiant (1984) *A theory of the learnable*, Communications of the ACM, 27(11):1134-1142.
  2. V. Vapnik und A. Chervonenkis (1971) *On the uniform convergence of relative frequencies of events to their probabilities* Theory of Probability and its Applications , 16 (2): 264-280.
- **Introduction:**
  1. C. J.C. Burges (1998) *A Tutorial on Support Vector Machines for Pattern Recognition* Data Mining and Knowledge Discovery 2, 121-167.
  2. S. Shalev-Shwartz and S. Ben-David.(2014) *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press.
  3. V. N. Vapnik (1998) *Statistical Learning Theory*, Wiley-Interscience
  4. L. Bottou, F. E. Curtis, J. Nocedal (2018) *Optimization Methods for Large-Scale Machine Learning*, SIAM Review.