Statistical Learning Theory

Marius Kloft and Klaus-Robert Müller

TU Berlin

October 23, 2012

Introduction to Statistical Learning Theory

Outline:

- Problem setting and terminology
- Concentration Inequalities
- Vapnik-Chervonenkis theory

Problem setting

The goal in **statistical learning theory** is to find a classifier $g : \mathbb{R}^d \to \{0, 1\}$, predicting the correct class y of an observation $x \in \mathbb{R}^d$, based on data $(x_1, y_1), \ldots, (x_n, y_n)$.

Because we cannot learn a reasonable classifier, if no assumption is imposed on the relationship between the data and the test observation (x, y), we require:

Assumption

Let the data $D_n := (x_i, y_i)_{i=1}^n$ and test observation (x, y) be independently drawn from one and the same probability distribution $\mathbb{P}$.

Notation: we denote the random variables associated to (x_i, y_i) and (x, y) by capital letters, i.e., (X_i, Y_i) and (X, Y), respectively.

Bayes classifier

A classifier errs if $g(X) \neq Y$ so that $L(g) := \mathbb{P}(g(X) \neq Y | D_n)$ is the probability of error of g.

The **Bayes classifier**, defined as

$$g^*(x) := \underset{g}{\operatorname{argmin}}\, L(g) = \begin{cases} 1, & \text{if } \mathbb{P}(Y=1|X=x) > \frac{1}{2} \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

is, by definition, the **most accurate** classifier in average. If $\mathbb{P}$ is known, the Bayes classifier may be computed.

However, most often $\mathbb{P}$ is unknown in practice and needs to be **approximated** on base of the data:

$$\hat{L}_n(g) \quad := \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{g(X_i) \neq Y_i\}}}_{\textbf{empirical error}} \quad \approx \quad \underbrace{L(g)}_{\textbf{error probability}} \quad .$$

Empirical Risk Minimization (ERM)

The Bayes classifier is thus roughly approximated by:

Empirical risk minimization (ERM)

$$g^* := \operatorname*{argmin}_{g \in \mathcal{C}} \hat{L}_n(g)$$

In comparison to the Bayes classifier, ERM has two limitations

1. the empirical error $\hat{L}(g)$ is minimized, rather than the error probability $L(g)$
2. the minimization is over a sub-class $\mathcal{C}$ of classifiers, to avoid **overfitting**.

What is "lost" by the ERM approximation?

The **sub-optimality of ERM** is measured by $L(g_n^*) - L(g^*)$, i.e., the differences of the error probabilities of ERM and the Bayes classifier. We thus need to analyze $L(g_n^*) - L(g^*)$.

To this end, denote the most accurate classifier in the class $\mathcal{C}$ by $g_{\mathcal{C}}^* := \operatorname{argmin}_{g \in \mathcal{C}} L(g)$. Clearly, we may write:

$$L(g_n^*) - L(g^*) = \underbrace{L(g_n^*) - L(g_{\mathcal{C}}^*)}_{\text{called "estimation error"}} + \underbrace{L(g_{\mathcal{C}}^*) - L(g^*)}_{\text{called "approximation error"}} .$$

Approximation error: not controllable; may converge arbitrarily slowly when $n \to \infty$.

Estimation error: controllable; we will prove: converges to zero at a rate of $O(\sqrt{1/n})$.

Bounding the estimation error

Lemma

$$\underbrace{L(g_n^*) - L(g_{\mathcal{C}}^*)}_{\textit{estimation error}} \;\leq\; 2 \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| .$$

Proof.

$$\begin{aligned} & L(g_n^*) - L(g_{\mathcal{C}}^*) \\ &= L(g_n^*) - \hat{L}_n(g_n^*) + \Big(\underbrace{\hat{L}_n(g_n^*)}_{\leq \hat{L}(g_{\mathcal{C}}^*)} - L(g_{\mathcal{C}}^*) \Big) \\ &\leq 2 \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| . \end{aligned}$$

$\square$

Consequences of the Lemma

The above lemma states that upper bounds on $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$ automatically provide us with upper bounds on the sub-optimality of the ERM classifier g_n^* within $\mathcal{C}$, that is, a bound for the estimation error $L(g_n^*) - L(g_{\mathcal{C}}^*)$. This explains why...

The classical task in statistical learning theory is

to derive upper bounds on $\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right|$, i.e.,

$$\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right| \leq \mathrm{bound}(n)$$

with $\mathrm{bound}(n) \to 0$ when $n \to \infty$ at a reasonable speed (usually $O(\sqrt{1/n})$).

Warning: **pointwise convergence**, i.e., $\forall g \in \mathcal{C} : |\hat{L}_n(g) - L(g)| \to 0$ when $n \to \infty$ is not enough! We need that $\left|\hat{L}_n(g) - L(g)\right|$ convergences **uniformly** in $\mathcal{C}$.

What is coming up?

We bound $P(\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \geq t)$ in **two steps**:

1. showing that $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$ is **"concentrated"**, i.e., it is, with high probability over the draw of the data, very close to its mean $\mathbb{E} \sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$ (by "McDiarmid's inequality")
2. showing that $\mathbb{E} \sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \to 0$ when $n \to \infty$ at rate $O(\sqrt{1/n})$ (by "Vapnik-Chervonenkis theory")

This is justified by the following **decomposition**:

$$\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \leq$$

$$\underbrace{\left| \sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| - \mathbb{E} \sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \right|}_{\leq \text{ bound } (\text{Step 1})} + \underbrace{\mathbb{E} \sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|}_{\leq \text{ bound } (\text{Step 2})}$$

Outline

To reach step 1, we will introduce the theory of **concentration inequalities**, i.e., inequalities of the form: for a random variable Z and any real number $t > 0$,

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq \mathrm{bound}(t, n)\,.$$

To this end, we will step by step prove:

- Markov's inequality
- Chernoff's inequality

at the very end, reaching the very powerful **concentration inequality** of **McDiarmid (1989)**, which gives the required result of step 1.

Markov's inequality

The starting point of all concentration inequalities is the following simple, yet very useful result:

Proposition (Markov's inequality)

For any positive random variable Z and any real number $t > 0$,

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}Z}{t} .$$

Proof.

The core idea of the proof is to consider the random variable $Z_t := t\mathbb{I}_{\{Z \geq t\}}$. Note that Z_t is positive and it holds $Z_t \leq Z$ with probability one as well as, per construction, $EZ_t = t\mathbb{E}\mathbb{I}_{\{Z \geq t\}} = t\mathbb{P}(Z \geq t)$. Thus it follows

$$\mathbb{P}(Z \geq t) = \frac{EZ_t}{t} \leq \frac{\mathbb{E}Z}{t} ,$$

which was to show. □

From Markov's inequality, we easily prove:

Proposition (Chernoff's inequality)

For any random variable Z and any $t > 0$,

$$\mathbb{P}(Z \geq t) \leq \min_{s \in \mathbb{R}} \frac{M_Z(s)}{e^{st}},$$

where $M_Z(s) = \mathbb{E}e^{sZ}$ is the moment-generating function of Z.

Proof.

Note that by Markov's inequality $\mathbb{P}(Z \geq t) = \mathbb{P}(e^{sZ} \geq e^{st}) \leq \frac{\mathbb{E}e^{sZ}}{e^{st}}$, which was to show. □

Discussion (Chernoff's inequality)

The **moment-generating function (MGF)** occurring in Chernoff's inequality is, for many distributions, well known from the literature; e.g.:

Example (MGF of Gaussian random variables)

The MGF of a **Gaussian** random variable Z with expected value $E(Z) = 0$ and variance σ^2 is given by: for any $s \in \mathbb{R}$,

$$M_Z(s) = e^{\frac{1}{2}\sigma^2 s}\,.$$

Most relevant for us (because $0 \leq \hat{L}_n(g), L(g) \leq 1$) are **bounded** variables:

Lemma (Höffding's lemma. *For the proof, see lecture notes*)

A random variable Z is **bounded**, *if there exist constants $a, b > 0$ such that $\mathbb{P}(a \leq Z \leq b) = 1$. The MGF of a bounded random variable Z with expected value $\mathbb{E}(Z) = 0$ is upper bounded by: for any $s \in \mathbb{R}$,*

$$M_Z(s) \;\leq\; e^{s^2(b-a)^2/8}\,.$$

McDiarmid's inquality

We are now ready to prove the main concentration inequality of this lecture.

Assumption (Bounded difference assumption)

Let A be some set; a function $f : A^n \to \mathbb{R}$ satisfies the bounded difference assumption, if there exist real numbers $c_1, \ldots, c_n > 0$ so that for all $i = 1, \ldots, n$,

$$\sup_{z_1,\ldots,z_n,z_i' \in A} |f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i \, .$$

In words, if we change the ith variable while keeping all the others fixed, the value of the functionv g does not change by more than c_i.

Theorem (McDiarmid's inequality)

Under the bounded difference assumption, it holds, for all $t > 0$,

$$\mathbb{P}(|f(Z_1, \ldots, Z_n) - \mathbb{E}f(Z_1, \ldots, Z_n)| \geq t) \leq 2e^{-2t^2 / \sum_{i=1}^n c_i^2} \, .$$

Proof (McDiarmid's inequality)

Proof.

Write $f \equiv f(Z_1, \ldots, Z_n)$, $V := f - \mathbb{E}f$, and $V = \sum_{i=1}^{n} V_i$ with $V_i := \mathbb{E}[f|Z_1, \ldots, Z_i] - \mathbb{E}[f|Z_1, \ldots, Z_{i-1}]$, where $\mathbb{E}[f|Z_1, \ldots, Z_i]$ denotes the expected value conditioned on $Z_1, ..., Z_i$.

Changing the value of Z_i can, by the bounded difference assumption, change the value of V_i by at most c_i. Moreover $\mathbb{E}[V_i|Z_1, \ldots, Z_{i-1}] = 0$. Thus, by Höffding's lemma,

$$\mathbb{E}[e^{sV_i}|Z_1, \ldots, Z_{i-1}] \leq e^{s^2 c_i^2/8} . \quad (2)$$

Proof continued.

Hence, by Chernoff's inequality,

$$
\begin{aligned}
&\mathbb{P}(f - \mathbb{E}f \geq t) \\
&\quad \leq \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}e^{s(f-\mathbb{E}f)} = \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}e^{s\sum_{i=1}^{n} V_i} \\
&\quad = \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}\,\mathbb{E}[e^{s\sum_{i=1}^{n} V_i}|Z_1,\ldots,Z_{n-1}] \\
&\quad = \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}\,\mathbb{E}[e^{s\sum_{i=1}^{n-1} V_i}\mathbb{E}[e^{sV_n}|Z_1,\ldots,Z_{n-1}]|Z_1,\ldots,Z_{n-1}] \\
&\quad \overset{(2)}{\leq} \min_{s\in\mathbb{R}} e^{s^2c_i^2/8-st}\mathbb{E}\,\mathbb{E}[e^{s\sum_{i=1}^{n-1} V_i}|Z_1,\ldots,Z_{n-1}] \\
&\quad \leq \ \ldots \qquad (\text{REPEATING THE ARGUMENT } (n-1) \text{ TIMES}) \\
&\quad \leq \min_{s\in\mathbb{R}} e^{ns^2\sum_{i=1}^{n} c_i^2/8-st}.
\end{aligned}
$$

Proof continued.

$e^{ns^2 \sum_{i=1}^n c_i^2/8 - st}$ is minimized for $s := 4t/\sum_{i=1}^n c_i^2$, thus giving

$$\mathbb{P}(f - \mathbb{E}f \geq t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2} .$$

Analogously, repeating the argument for the function $-f$, we obtain the corresponding left-sided inequality

$$\mathbb{P}(f - \mathbb{E}f \leq -t) = \mathbb{P}(-f - \mathbb{E}(-f) \geq t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2} .$$

Combining both results gives the claimed result. $\square$

Consequences for Learning Theory

Corollary

Let $\mathcal{C}$ be a class of functions. Then, for any $t > 0$,

$$P\left(\left|\sup_{g\in\mathcal{C}}|\hat{L}_n(g) - L(g)| - \mathbb{E}\sup_{g\in\mathcal{C}}|\hat{L}_n(g) - L(g)|\right| \geq t\right) \leq 2e^{-2nt^2}.$$

Proof.

Put $Z_i := (X_i, Y_i)$, $i \in \mathbb{N}$, and $f(Z_1, \ldots, Z_n) := \sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$. Then f satisfies the bounded difference assumption with $c_i = 1/n$ for all $n \in \mathbb{N}$. The claimed inequality thus follows from McDiarmid's inequality. □

The big picture

Recall from the beginning of this lecture that our overall goal is to bound the **estimation error of ERM** and that it holds

$$\underbrace{L(g_n^*) - L(g_{\mathcal{C}}^*)}_{\text{estimation error}} \;\leq\; 2 \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| .$$

By the corollary from the previous slide, with probability $2e^{-2n\epsilon^2}$,

$$\sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right|$$

$$\underbrace{\left| \sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)| - \mathbb{E} \sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)| \right|}_{\leq\ \epsilon \quad (\text{BY MCDIARMID})} + \underbrace{\mathbb{E} \sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)|}_{\text{still left to bound!}}$$

We will bound the expected value $\mathbb{E} \sup_{g \in \mathcal{C}} |\hat{L}_n(g) - L(g)|$ using **Vapnik-Chervonenkis theory**.

Vapnik-Chervonenkis Theory

To bound the expected value $\mathbb{E} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right|$ we proceed in three steps:

1. relating $\mathbb{E} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right|$ with $\mathfrak{R}_n(\mathcal{C})$, the so-called *Rademacher complexity* of the class $\mathcal{C}$
2. relating $\mathfrak{R}_n(\mathcal{C})$ with the so-called *VC shattering coefficient* $\mathbb{S}_n(\mathcal{C})$
3. relating $\mathbb{S}_n(\mathcal{C})$ with the *VC dimension* V
4. computing V for specific classes $\mathcal{C}$.

Step 1: relating $E \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right|$ with $\mathfrak{R}_n(\mathcal{C})$

Definition (Rademacher complexity)

The *(empirical) Rademacher complexity* of a function class $\mathcal{C}$ is defined as

$$\mathfrak{R}_n(\mathcal{C}) := \mathbb{E}_\varsigma \sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varsigma_i \mathbb{I}_{\{g(X_i) \neq Y_i\}} \right| ,$$

where $\varsigma = (\varsigma_i)_{i=1,\dots,n}$ is an i.i.d. family of *Rademacher variables*, i.e., $\mathbb{P}(\varsigma_i = +1) = \mathbb{P}(\varsigma_i = -1)$.

The Rademacher complexity, intuitively, measures how well the empirical error can, when optimized over $g \in \mathcal{C}$, match with random signs.

Lemma (Rademacher lemma)

Let $\mathcal{C}$ be a class of functions. Then

$$\mathbb{E} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| \leq 2 \mathbb{E}_S \mathfrak{R}_n(\mathcal{C}) .$$

Proof of Rademacher lemma

Proof.

The core idea of the proof is to introduce $X'_1, \ldots, X'_n$ and $Y'_1, \ldots, Y'_n$, an independent copy of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively (called *ghost sample*), as well as $\varsigma = (\varsigma_i)_{i=1}^n$, an i.i.d. family of *Rademacher variables* that are independent of the sample and the ghost sample. Then, denoting $\hat{L}'_n(g) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{g(X'_i) \neq Y'_i\}}$, we have ...

Proof continued.

$$
\begin{aligned}
&\mathbb{E} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| \\
&\quad = \quad \mathbb{E}_S \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - \mathbb{E}_{S'} \hat{L}'_n(g) \right| \\
&\quad \leq \quad \mathbb{E}_S \mathbb{E}_{S'} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - \hat{L}'_n(g) \right| \\
&\qquad\qquad \text{(because } \sup_{i \in I} |\mathbb{E} Z_i| \leq \mathbb{E} \sup_{i \in I} |Z_i| \text{)} \\
&\quad = \quad \mathbb{E}_S \mathbb{E}_{S'} \mathbb{E}_\varsigma \sup_{g \in \mathcal{C}} \left| \sum_{i=1}^n \varsigma_i \left(\mathbb{I}_{\{g(X'_i) \neq Y'_i\}} - \mathbb{I}_{\{g(X'_i) \neq Y'_i\}} \right) \right| \\
&\qquad\qquad \text{(by the symmetry of the Rademacher variables)} \\
&\quad \leq \quad \underbrace{2 \mathbb{E}_S \, \mathbb{E}_\varsigma \sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \varsigma_i \mathbb{I}_{\{g(X_i) \neq Y_i\}} \right|}_{= \mathfrak{R}_n(\mathcal{C})} .
\end{aligned}
$$

□

Step 2: relating $\mathfrak{R}_n(\mathcal{C})$ with $\mathbb{S}_n(\mathcal{C})$

Definition (VC shatter coefficient)

The *VC shatter coefficient* of a function class $\mathcal{C}$ is defined as

$$\mathbb{S}_n(\mathcal{C}) = \max_{x_i \in \mathbb{R}^d,\, y_i \in \mathbb{R},\, i=1,\ldots,n} \left| \left\{ \left(\mathbb{I}_{\{g(X_1) \neq Y_1\}}, \ldots, \mathbb{I}_{\{g(X_n) \neq Y_n\}} \right) : g \in \mathcal{C} \right\} \right| .$$

The shatter coefficient, how many different functions "effectively" are in $\mathcal{C}$, after being processed by the loss function.

Theorem (Vapnik-Chervonenkis inequality)

Let $\mathcal{C}$ be a class of functions. Then

$$\mathfrak{R}_n(\mathcal{C}) \leq \sqrt{\frac{2 \log(2\, \mathbb{S}_n(\mathcal{C}))}{n}} .$$

Proof of Vapnik-Chervonenkis inequality.

Think of the variables $X_i, Y_i, i = 1, \ldots, n$ as being fixed, i.e., $\mathfrak{R}_n(\mathcal{C})$ only randomly depending on $\varsigma_1, \ldots, \varsigma_n$. Note that $\varsigma_i \mathbb{I}_{\{g(X_i) \neq Y_i\}}$, $i = 1, \ldots, n$ has zero mean and ranges in $[-1, 1]$. Thus, by Höffding's Lemma, $\mathbb{E} e^{s\varsigma_i \mathbb{I}_{\{g(X_i) \neq Y_i\}}} \leq e^{s^2/2}$. Thus it follows

$$\mathbb{E} e^{\frac{s}{n} \sum_i \varsigma_i \mathbb{I}_{\{g(X_i) \neq Y_i\}}} = \prod_{i=1}^{n} \mathbb{E} e^{\frac{s}{n} \varsigma_i \mathbb{I}_{\{g(X_i) \neq Y_i\}}} \leq \prod_{i=1}^{n} e^{\frac{s^2}{2n^2}} = e^{\frac{s^2}{2n}},$$

Hence, by the subsequent lemma,

$$\mathfrak{R}_n(\mathcal{C}) \leq \sqrt{\frac{2 \log(2 \mathbb{S}_n(\mathcal{C}))}{n}}$$

because, for fixed X_i, Y_i, $i = 1, \ldots, n$, the sup in the definition of $\mathfrak{R}_n(\mathcal{C})$ is effectively only over $\mathbb{S}_n(\mathcal{C})$ many values. □

Lemma

If $\mathbb{E} e^{sZ_i} \leq e^{\frac{\sigma^2 s^2}{2}}$, *then* $\mathbb{E} \max_{i=1,\ldots,k} |Z_i| \leq \sigma \sqrt{2 \log(2k)}$.

Proof.

By Jensen's inequality,

$$e^{s\mathbb{E}\max_{i=1,\ldots,n} Z_i} \overset{\text{JENSEN}}{\leq} \mathbb{E}e^{s\max_{i=1,\ldots,n} Z_i} = \mathbb{E}\max_{i=1,\ldots,n} e^{sZ_i}$$

$$\leq \sum_{i=1}^{n} \mathbb{E}e^{sZ_i} \leq ne^{s^2\sigma^2/2}.$$

Thus, $\mathbb{E}\max_{i=1,\ldots,n} Z_i \leq \log(n)/s + s\sigma^2/2$, which is minimized for $s := \sqrt{2\log(n)/\sigma^2}$. Resubstitution gives

$$\mathbb{E}\max_{i=1,\ldots,n} Z_i \leq \sigma\sqrt{2\log(n)}.$$

The result follows because

$$\max_{i=1,\ldots,n} |Z_i| = \max(Z_1, -Z_1, \ldots, Z_n, -Z_n).$$

$\square$

Step 3: Relating $\mathbb{S}_n(\mathcal{C})$ with the VC dimension V

For the Vapnik-Chervonenkis inequality to converge when $n \to \infty$, the quantity $\log(\mathbb{S}_n(\mathcal{C}))$ needs to decrease sublinearly in n. Thus we define:

Definition

The *V-C dimension* V is the smallest integer n such that $\mathbb{S}_n(\mathcal{C}) = 2^n$.

Example

For any non-colinear set of points $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$ and any choice of labels $y_1, \dots, y_n \in \{0, 1\}$, there is an affine-linear function, separating the two classes without any error, if and only if $n = d + 1$. Thus $V = d + 1$.

An interesting phase transition occurs for the VC shattering coefficient $\mathbb{S}_n(\mathcal{C})$ when $n > V$.

Lemma (Sauer's lemma)

For any $n > V$, $\mathbb{S}_n(\mathcal{C}) \leq (n+1)^V$.

Proof.

Fix the variables $x_i, y_i, i = 1, \ldots, n$ and consider the resulting table of values $\left\{ \left(\mathbb{I}_{\{g(x_1) \neq y_1\}}, \ldots, \mathbb{I}_{\{g(x_n) \neq y_n\}} \right) : g \in \mathcal{C} \right\}$. E.g., for $n = 5$, this could look as follows:

$T :=$

	x_1	x_2	x_3	x_4	x_5
g_1	0	1	0	1	1
g_2	1	0	0	1	1
g_3	1	1	1	0	1
g_4	0	1	1	0	0
g_5	0	0	0	1	0

Each row corresponds to one possible evaluation of a function in $\mathcal{C}$ on the sample, and the cardinality

$$\left| \left\{ \left(\mathbb{I}_{\{g(x_1) \neq y_1\}}, \ldots, \mathbb{I}_{\{g(x_n) \neq y_n\}} \right) : g \in \mathcal{C} \right\} \right|$$

equals the number of rows. □

Proof continued.

We translate the table by *shifting*, for each $i = 1, \ldots, n$, column i, that is, for each row, we replace a 1 in column i by a 0, unless this would produce a row that is already contained in the table.
After applying the shifting operation in order from x_1 to x_n, we get the following table, which contains mostly 0s.

$T^* :=$

	x_1	x_2	x_3	x_4	x_5
g_1	0	1	0	0	0
g_2	0	0	0	1	1
g_3	0	0	0	0	1
g_4	0	0	0	0	0
g_5	0	0	0	0	0

From the example, we can make the following observations:

1. The size of the table is unchanged because the rows are still distinct.
2. The shifted table T^* exhibits the is *closed below*, i.e., replacing any of the 1s in the table would produce a duplicate row in the table.

□

Proof continued.

Furthermore, the VC dimension of the original table T is at least as high as the one of the shifted table T^*, i.e., $\mathrm{VC}(T) \geq \mathrm{VC}(T^*)$. To see this, consider a subset of columns that is shattered in T^*; the same subset must also be shattered in T.

We conclude that T^* cannot have more than V 1s in a row and thus has $\leq \sum_{i=0}^{n} \binom{n}{i}$ rows (imagine assigning, for each $i = 0, \ldots, V$, i many 1s to the positions $1, \ldots, n$) and the same holds for T.

Moreover, by the binomial theorem,

$$\sum_{i=0}^{V} \binom{n}{i} = \sum_{i=0}^{V} \frac{n!}{((n-i)!i!} \leq \sum_{i=0}^{n} \frac{n^i}{i!}$$

$$\leq \sum_{i=0}^{V} \frac{n^i V!}{i!(V-i)!} = \sum_{i=0}^{V} n \binom{V}{i} \overset{\text{Bin.}}{=} (n+1)^V$$

$\square$

Conclusion

Putting things together, we obtain the following bound:

Corollary

With probability $1-\delta$*,*

$$\sup_{g \in \mathcal{C}} \left| \hat{L}(g) - L(g) \right| \;\leq\; \sqrt{\frac{\log(1/\delta)}{2n}} + 2\sqrt{\frac{2(V \log(n+1) + \log 2)}{n}}$$

Proof.

The result is obtained by setting $\epsilon := \sqrt{\frac{\log(2/\delta)}{2n}}$. □

Corollary

The estimation error of ERM with linear functions in $\mathbb{R}^d$*, is, with probability* $1-\delta$*, bounded by*

$$L(g_n^*) - L(g^*) \;\leq\; 2\sqrt{\frac{\log(1/\delta)}{2n}} + 4\sqrt{\frac{2(d+1)\log(n+1) + 2\log 2)}{n}}.$$

Interpretation

Going back to the slide from the beginning,

$$L(g_n^*) - L(g^*) = \underbrace{L(g_n^*) - L(g_{\mathcal{C}}^*)}_{\text{"estimation error"}} + \underbrace{L(g_{\mathcal{C}}^*) - L(g^*)}_{\text{"approximation error"}} \quad .$$

Estimation error: controllable; we just have shown we will prove: converges to zero at a rate of $O(\sqrt{V/n})$, where V is the VC dimension.

Approximation error: not controllable; may converge arbitrarily slowly when $n \to \infty$.

However, when increasing the size of the class, the approximation error may shrink. On the other hand, VC dimension may increase in this case, thus the estimation error decreases.

Bottom line: regarding the choice of the class $\mathcal{C}$, there is tradeoff between estimation and approximation error.

Bibliography

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sums of observations. *Annals of Mathematical Statistics*, 23: 409–507, 1952.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.

N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13 (1):145–147, 1972.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.