Statistical Data Analysis

Jana de Wiljes

wiljes@uni-potsdam.de

www.dewiljes-lab.com

19. Oktober 2022

Universität Potsdam

Chebyshev Inequality

Proposition: Let X be a random variable on R with mean μ and variance σ^2. Then for all $k \geq 0$

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{1}$$

Proof: Apply Markov's inequality to the random variable $(X - \mu)^2$, this yields

$$\mathbb{P}(|X - \mu| \geq \sigma k) = \mathbb{P}((X - \mu)^2 \geq \sigma^2 k^2) \tag{2}$$

$$\leq \frac{1}{\sigma^2 k^2} \mathbb{E}[(X - \mu)^2] \tag{3}$$

$$= \frac{\sigma^2}{\sigma^2 k^2} = \frac{1}{k^2} \tag{4}$$

On the way to Hoeffding's Inequality

Proposition: Let X be a centered ($\mathbb{E}[X] = 0$) random variable bounded in $[a, b]$. Then for any $s \in \mathbb{R}$

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

Reminder: convexity $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$

Proof: From convexity of the exponential function, for any $a \leq x \leq b$,

$$e^{sx} \leq \underbrace{\frac{x-a}{b-a}}_{t} e^{sb} + \underbrace{\frac{b-x}{b-a}}_{1-t} e^{as}$$

Note

$$sx = s\left(\frac{(x-a)}{b-a}b + \frac{(b-x)}{b-a}a\right) \tag{5}$$

On the way to Hoeffding's Inequality

Define $p = \Big(- a/(b-a) \Big)$ then (recall that $\mathbb{E}[X] = 0$)

$$\mathbb{E}[e^{sX}] \leq \underbrace{\frac{b}{b-a}}_{p} e^{sa} \underbrace{-\frac{a}{b-a}}_{1-p} e^{sb} \text{ (apply expectation on both sides)} \tag{6}$$

$$= \exp(\phi(u)) \tag{7}$$

where

$$\phi(u) := \log(pe^{sa}(1-p)e^{sb}) \tag{8}$$

$$= sa + \log(p + (1-p)\exp(s(b-a))) \tag{9}$$

$$\Big(\cdot \frac{\exp(sa)}{\exp(-sa)} \text{ and } \log(xy) = \log(x) + \log(y) \Big) \tag{10}$$

$$= \underbrace{(p-1)u}_{sa} + \log(p + (1-p)\exp(u)) \tag{11}$$

$$\tag{12}$$

with $u = s(b-a)$

On the way to Hoeffding's Inequality

Thus from Taylor theorem, the exists a $\theta \in [0, u]$ such that

$$\phi(\theta) = \phi(0) + \theta\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8} \tag{13}$$

where

$$\phi'(u) = (p-1) + \frac{(1-p)\exp(u)}{p + (1-p)\exp(u)} \tag{14}$$

and

$$\phi''(u) = -\frac{p(1-p)e^u}{(p + (1-p)e^u)^2} \tag{15}$$

Note that $\phi(0) = \phi'(0) = 0$ and $\phi''(u) \leq 1/4$ and thus

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8} \tag{16}$$

□

Chernoff-Hoeffding Inequality

Theorem: Let $X_i \in [a_i, b_i]$ be n independent r.v. with mean $\mu_i = \mathbb{E}[X_i]$. Then

$$\mathbb{P}[|\sum_{i=1}^{n}(X_i - \mu_i)| \geq \epsilon] \leq 2\exp(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}) \qquad (17)$$

Hoeffding Inequality

Proof:

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i-\mu_i)\geq\epsilon\Big) = \mathbb{P}(e^{s\sum_{i=1}^{n}X_i-\mu_i}\geq e^{s\epsilon}) \tag{18}$$

$$\leq e^{-s\epsilon}\mathbb{E}[e^{s\sum_{i=1}^{n}X_i-\mu_i}] \quad \text{(Markov inequality)} \tag{19}$$

$$= e^{-s\epsilon}\prod_{i=1}^{n}\mathbb{E}[e^{s(X_i-\mu_i)}] \quad \text{(independent random variables)} \tag{20}$$

$$= e^{-s\epsilon}\prod_{i=1}^{n}e^{s^2(b_i-a_i)^2/8} \quad \text{(Hoeffding inequality)} \tag{21}$$

$$= e^{-s\epsilon+s^2(b_i-a_i)^2/8} \tag{22}$$

If we choose $s=4\epsilon/(\sum_{i=1}^{n}(b_i-a_i)^2)$, the result follows. Similar arguments hold for $\mathbb{P}(\sum_{i=1}^{n}X_i-\mu_i\leq-\epsilon)$

Jensen's inequality

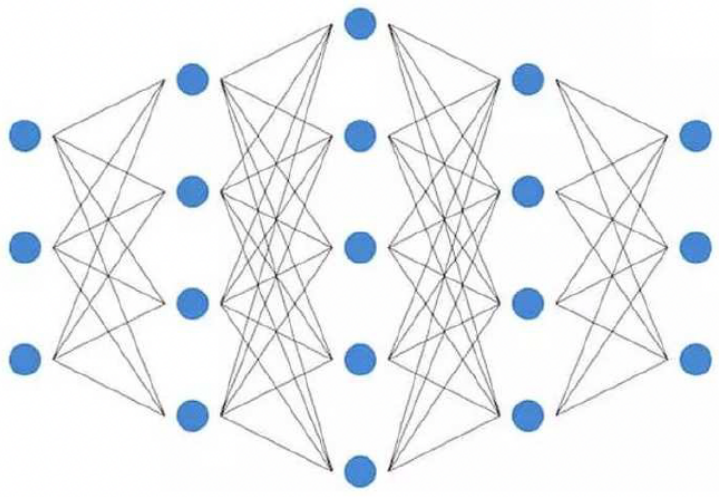
Proposition: Let g be a convex function and X random variable

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) \tag{23}$$

Motivation

Data

Different parametrisations possible



How can we efficiently process the data, to learn functions with a high prediction ability?

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

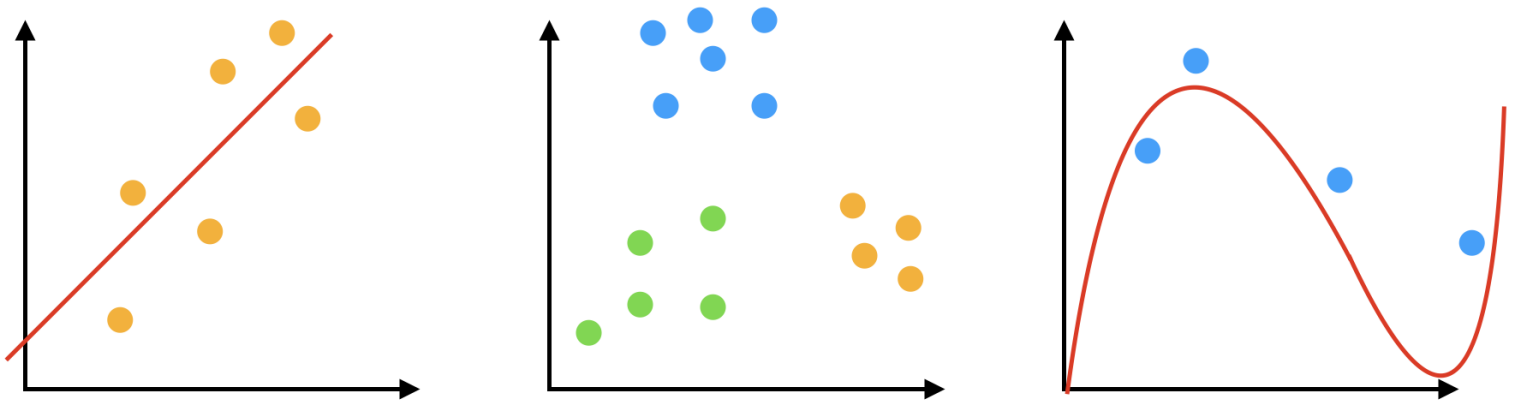
$$R(h) = \int_{\mathbf{Z}} l(y, h(x))\pi(z)dz \tag{24}$$

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

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \tag{25}$$

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 $i \in \{1, \ldots, N\}$



Empirisches Risiko

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

Empirischer Risiko-Minimierer

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.

Examples

Linear regression with regularisation

Data: $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}^h$

loss function:

$$l(x_i, f_w(x_i)) = \frac{1}{2}\|y_i - wx_i\|_2^2 + \|w_k\|_2^2$$

K-Means

Data: $x_i \in \mathbb{R}^d$, no labels y_i

loss function:

$$l(x_i, f_w(x_i)) = \min_{w_k} \frac{1}{2}\|x_i - w_k\|_2^2$$

where $w_1, \ldots w_K$ are unknown cluster centres

Perceptron

Data:$x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$

loss function:

$$l(x_i, f_w(x_i)) = \max(0, -yw^\top x)$$

Support Vector Machines

Data: $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$

loss function:

$$l(x_i, f_w(x_i)) = \lambda\|w\|_2^2 + \max(0, 1 - yw^\top x_i)$$

with hyper parameter $\lambda > 0$