Assignment 11

Statistical Machine Learning

Moritz Haas / Prof. Ulrike von Luxburg

Summer term 2022 — due on **July 11th at 12:00**

Exercise 1 (VC dimension, 1+2+1+3 points)
In this exercise we derive the VC dimension of finite function classes, closed intervals and linear functions.

(a) Assume that $\mathcal{F}$ is a finite function class. Prove that

$$\mathrm{VC}(\mathcal{F}) \leq \log_2 |\mathcal{F}|.$$

(b) For $a, b \in \mathbb{R}$, let $\mathbb{1}_{[a,b]}(x)$ be the characteristic function such that

$$\mathbb{1}_{[a,b]}(x) = \begin{cases} 1 & \text{if } x \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$$

Consider the function class

$$\mathcal{F} = \{f : \mathbb{R} \to \{0,1\} \mid f(x) = \mathbb{1}_{[a,b]}(x), a, b \in \mathbb{R}\}.$$

Prove that $\mathcal{N}(\mathcal{F}, n) = \frac{n^2+n}{2} + 1$ and $\mathrm{VC}(\mathcal{F}) = 2$.

(c) Consider the function class

$$\mathcal{F} = \{f : \mathbb{R} \to \{0,1\} \mid f(x) = \sum_{i=1}^{n} \mathbb{1}_{[a_i,b_i]}(x), n \in \mathbb{N}, a_i, b_i \in \mathbb{R}\}.$$

Prove that $\mathrm{VC}(\mathcal{F}) = \infty$.

(d) Consider the function class of hyperplanes, that is

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \to \{0,1\} \mid f(x) = \frac{1 + \mathrm{sign}(\langle w, x \rangle + b)}{2}, w \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

Prove that $\mathrm{VC}(\mathcal{F}) = d + 1$.

Hint: You can use the following theorem (without proving it):

> Two sets of points in $\mathbb{R}^d$ can be separated by a hyperplane if and only if the intersection of their convex hulls is empty.

Exercise 2 (Generalization bound using VC dimension, 1 + 3 + 1 points)
In this exercise, you will prove the result (Theorem 42, lecture slide 863) providing a generalization bound based on the VC dimension of a (possibly infinite) function class.

(a) Given a function class $\mathcal{F}$, prove that for any two independently drawn samples of size n from a probability distribution $\mathbb{P}$, for any function $f \in \mathcal{F}$ following inequality holds.

$$\forall t > 0, \mathbb{P}(R_n(f) - R_n^{'}(f) > t) \leq 2e^{-nt^2/2} \tag{1}$$

where $R_n(f)$ and $R_n^{'}(f)$ denote the risks of the function f computed on the two independent samples.

Hint: Use Hoeffding's inequality as given in Proposition 33, lecture slide 819.

(b) Use the result from part (a) and the symmetrization lemma (Proposition 40, lecture slide 850) to prove that $\forall\ 0 < \delta < 1$, with probability at least $1 - \delta$, all functions $f \in \mathcal{F}$ satisfy

$$R(f) \leq R_n(f) + 2\sqrt{2\frac{log(\mathcal{N}(\mathcal{F}, 2n)) + log(\frac{4}{\delta})}{n}}, \tag{2}$$

where $R(f)$ is the true risk of the function f with respect to $\mathbb{P}$ and $\mathcal{N}(\mathcal{F}, 2n)$ is the shattering coefficient.

(c) Finally, use the results from parts (a) and (b) and the Sauer-Shelah Lemma (see Exercise 4) to prove that, given $n \geq d$, $\forall\ \delta \in (0, 1)$, with probability at least $1 - \delta$, all functions $f \in \mathcal{F}$ satisfy

$$R(f) \leq R_n(f) + 2\sqrt{2\frac{d\log(2en/d) + log(\frac{4}{\delta})}{n}}. \tag{3}$$

Exercise 3 (Empirical Chebyshev and Hoeffding's inequality, 3+2+1 points)
In this exercise we investigate Chebyshev and Hoeffding's inequality empirically. We will also see a case where something goes wrong. Chebyshev's inequality states the following. Let X be a random variable with finite mean μ and finite non-zero variance σ then for all $k > 0$ we have

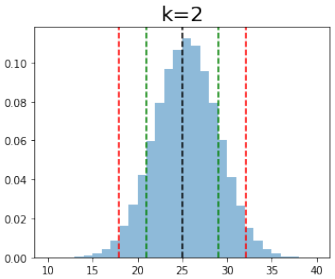
$$P\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}.$$

Meanwhile the Hoeffding's inequality states the following. Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables with support in $[0, 1]$. Then for any $\varepsilon \geq 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - E(Z)\right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}. \tag{4}$$

(a) Sample `nb_samples` points from a binomial distribution with $n = 50$ and $p = 0.5$.

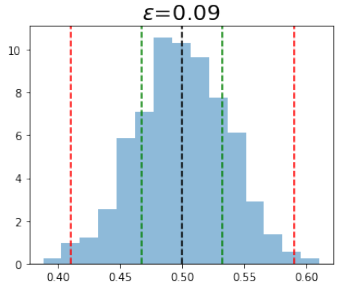
- For all `k` in `ks` compute and print the empirical probability and the bound given by Chebyshev's inequality. Does Chebyshev's inequality hold?
- Plot the empirical distribution of the sampled variable X. Indicate the empirical mean. For all `k` in `ks`, indicate the Chebyshev bound (red). To see how tight the bound is, indicate the minimal deviation from the mean such that the bounding probability $\frac{1}{k^2}$ is empirically fulfilled (green). Your figure(s) can, for example, look like the one below (only the case $k = 2$ is depicted).



(b) Consider a Beta distribution with $\alpha = \beta = 0.5$. Notice that the Beta distribution has support on $[0, 1]$. Compute the true expected value for this distribution.

- For all ε in `epsilons` print the empirical probability and the Hoeffding bound. This can be done by repeating `m` times the following procedure: Sample $n =$ `nb_samples` realisations z_i from the distribution, then check if $\left|\frac{1}{n}\sum_{i=1}^{n} z_i - E(Z)\right| \geq \varepsilon$. By computing the number of such occurrences, you get an empirical estimate of the probability in (4). Does Hoeffding's inequality hold?

- Plot the empirical distribution of the `m` realisations of the sample mean, together with the true mean, the various confidence levels ε and the minimal deviation from the mean that attains the bound $2e^{-2n\varepsilon^2}$. Your figure(s) can, for example, look like the one below (only the case $\varepsilon = 0.09$ is depicted).



(c) Consider the Student-t distribution with 10 degrees of freedom. In `xs` you find $(\texttt{m}, \texttt{nb\_samples})$ points from

$$\text{StudentT}(10)$$

Compute the true expected value for this distribution. For all ε in `epsilons` print the empirical probability and the expected Hoeffding bound like in (b). Does Hoeffding's inequality hold?

Exercise 4 (Bonus: Proving the Sauer-Shelah Lemma, 3+2 bonus points)

(a) The Sauer-Shelah Lemma claims that for any function class $\mathcal{F}$ with $\text{VC}(\mathcal{F}) = d \geq 0$ we have

$$\mathcal{N}(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i}, \quad n \geq 1.$$

Assume that we already proved the lemma for $d = 0$ and arbitrary $n \geq 1$ and also for $n = 1$ and arbitrary $d \geq 0$. Prove the lemma for general $d \geq 0, n \geq 1$ by induction over $d + n$.

(b) Show that

$$\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{ne}{d}\right)^d, \quad n \geq d > 1.$$