Algorithmic Foundations of Learning

Problem Sheet 2

Lecturer: Patrick Rebeschini *Due date: 48 hours before class*

2.1 Maximal Inequalities for Sub-Gaussian Random Variables (Question type: A)

Let $X_1, \ldots, X_n$ be a collection of sub-Gaussian random variables (not necessarily independent) with mean μ and variance proxy σ^2. Prove that

$$\mathbf{E} \max_{i \in [n]} (X_i - \mu) \leq \sigma \sqrt{2 \log n},$$

$$\mathbf{P}\Big(\max_{i \in [n]} (X_i - \mu) > \varepsilon \Big) \leq n e^{-\varepsilon^2/(2\sigma^2)}.$$

2.2 Maximal Inequalities for Linear Predictors (Question type: B)

Let $X_1, \ldots, X_d$ be a collection of independent sub-Gaussian random variables with mean μ and variance proxy σ^2, and denote $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$. Let $\mathcal{B} := \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\}$. Using covering numbers, prove that

$$\mathbf{E} \sup_{w \in \mathcal{B}} w^\top (X - \mu) \leq 4\sigma\sqrt{d},$$

$$\mathbf{P}\Big(\sup_{w \in \mathcal{B}} w^\top (X - \mu) > \varepsilon \Big) \leq 6^d e^{-\varepsilon^2/(8\sigma^2)}.$$

Hint: Let $\mathcal{C} \subseteq \mathcal{B}$ be a 1/2-cover of the unit ball $\mathcal{B}$ with respect to the ℓ_2 norm. Relate $\sup_{w \in \mathcal{B}} w^\top (X - \mu)$ to $\max_{c \in \mathcal{C}} c^\top (X - \mu)$ and use that for any $w \in \mathcal{B}$, $w^\top (X - \mu)$ is sub-Gaussian with variance proxy σ^2.

2.3 Linear Predictors ℓ_1/ℓ_∞ Constraints (Question type: A)

Give a proof of Proposition 3.3 in the Lecture Notes that does not use Hölder's inequality.

2.4 Proof of Sauer-Shelah's Lemma 4.11 (Question type: C)

In the following, let $d = \texttt{VC}(\mathcal{A})$.

1. For any $n \geq 1$, prove that

$$\tau_{\mathcal{A}}(n) \leq \sum_{k=0}^{d} \binom{n}{k}$$

using that $|\mathcal{A} \circ \{x_1, \dots, x_n\}| \le |y \subseteq \{x_1, \dots, x_n\} : \mathcal{A}$ shatters $y|$. The last statement is known as Pajor's theorem, whose proof is a non-trivial exercise in combinatorics.

2. For any $d \le n$, write $\sum_{k=0}^{d} \binom{n}{k}$ in terms of the probability mass function of the binomial distribution and show that
$$\sum_{k=0}^{d} \binom{n}{k} \le \left(\frac{en}{d}\right)^d.$$

2.5 VC Dimension (Question type: C)

Compute the VC dimension of the following families of classifiers.

1. Signed intervals over the real line: $\mathcal{A} = \{x \in \mathbb{R} \to u(2\mathbf{1}_{w^- \le x \le w^+} - 1) : w^- \le w^+, u \in \{-1, 1\}\}$.
2. Axis-aligned rectangles in $\mathbb{R}^2$: $\mathcal{A} = \{(x_1, x_2) \in \mathbb{R}^2 \to 2\mathbf{1}_{w_1^- \le x_1 \le w_1^+, w_2^- \le x_2 \le w_2^+} - 1 : w_1^- \le w_1^+, w_2^- \le w_2^+\}$.
3. Axis-aligned rectangles in $\mathbb{R}^d$: $\mathcal{A} = \{x \in \mathbb{R}^d \to 2\mathbf{1}_{w_i^- \le x_i \le w_i^+ \ \forall i \in [d]} - 1 : w_i^- \le w_i^+ \ \forall i \in [d]\}$.

Let $\mathcal{A}_1, \dots, \mathcal{A}_\ell$ be given hypothesis classes over the same domain $\mathcal{X}$. Let $d = \max_{i \in [\ell]} \mathtt{VC}(\mathcal{A}_i)$. Prove that
$$\mathtt{VC}\Big(\bigcup_{i \in [\ell]} \mathcal{A}_i\Big) \le 6d \log(3d) + 3d + 3 \log \ell.$$

Hint: Take a set of k points in $\mathcal{X}$ and assume that they are shattered by the union class $\mathcal{A}$ so that $\tau_{\mathcal{A}}(k) = 2^k$. At the same time, use Sauer-Shelah's Lemma, Lemma 4.11, to show that $\tau_{\mathcal{A}}(k) \le \ell(ek)^d$. Use that if $x < a \log x + b$ for given $a \ge 1, b > 0$, then $x < 4a \log(2a) + 2b$ (see Lemma A.2 in the book *Understanding Machine Learning: From Theory to Algorithms Textbook* by Shai Ben-David and Shai Shalev-Shwartz)

2.6 Monotonicity of Packing and Covering Numbers (Question type: B)

Prove Proposition 5.1 in the Lecture Notes.

2.7 Concentration Inequalities: Confidence Intervals and Sensitivity to Variance (Question type: B)

A biased coin with heads probability p is tossed n times. Compute a lower bound for the probability that the number of heads obtained is between $pn - \sqrt{n}$ and $pn + \sqrt{n}$ using, respectively,

1. Markov's inequality;
2. Chebyshev's inequality;
3. Hoeffding's inequality;
4. Bernstein's inequality.

Evaluate the lower bounds in the case $p = 1/2$ (fair coin) and $p = 0.99$. Comment your findings.

2.8 Bounds in probability with empirical Rademacher complexity (Question type: B)

Assume that the loss function ℓ is bounded in the interval $[0, c]$. Prove that with probability at least $1 - \delta$ we have

$$r(A^\star) - r(a^\star) < 4\,\mathtt{Rad}(\mathcal{L} \circ \{Z_1, \ldots, Z_n\}) + 5c\sqrt{2\frac{\log(2/\delta)}{n}}.$$

This statement is analogous (modulo constants) to the statement in Theorem 6.13, when the (deterministic) quantity $\mathbf{E}\,\mathtt{Rad}(\mathcal{L} \circ \{Z_1, \ldots, Z_n\})$ is replaced with the (random) quantity $\mathtt{Rad}(\mathcal{L} \circ \{Z_1, \ldots, Z_n\})$, a.k.a. the *empirical* Rademacher complexity. See also Remark 2.12. This bound is data-dependent as it depends on the training set $Z_1, \ldots, Z_n$. The bound in Theorem 6.13 only depends on the *distribution* of the data (via the expected value), not on the data itself.

2.9 From Bounds in Probability to Bounds in Expectation (Question type: B)

Let X be a random variable such that for any $\varepsilon \in \mathbb{R}$ we have

$$\mathbf{P}(|X - \mathbf{E}X| > \varepsilon) \le 2e^{-\varepsilon^2/(2\sigma^2)}.$$

Prove the following.

1. For any $k \ge 1$ we have
$$\mathbf{E}[|X - \mathbf{E}X|^k] \le (2\sigma^2)^{k/2} k\Gamma(k/2),$$
where the Gamma function is defined as $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x$ for any $t > 0$.
Hint: Use that for any non-negative random variable X we have $\mathbf{E}X = \int_0^\infty \mathbf{P}(X > \varepsilon)\mathrm{d}\varepsilon$.

2. For any $\lambda \in \mathbb{R}$ we have
$$\mathbf{E}\,e^{\lambda(X - \mathbf{E}X)} \le e^{c\sigma^2\lambda^2},$$
for a constant $c > 0$.
Hint: Use the series expansion for the exponential: $e^x = \sum_{k=0}^\infty \frac{x^k}{k!}$, that the Gamma function is non-decreasing with $\Gamma(k) = (k-1)!$ and $2(k!)^2 \le (2k)!$ for any integer $k \ge 1$.