IST Austria: Statistical Machine Learning 2020/21

Christoph Lampert <chl@ist.ac.at>
TAs: Alexandra Peste <alexandra.peste@ist.ac.at>,
Bernd Prach <bernd.prach@ist.ac.at>

Exercise Sheet 5/5 (due date 09/11/2020, 10:15am)
Please send your solutions via email to the TAs

1 Complexity measures

Let $\mathcal{H} = \{h_\theta(x) = \operatorname{sign}(x-\theta), \text{for } \theta \in \mathbb{R}\} \subset \{\mathcal{X} \to \{\pm 1\}\}$ be the set of threshold functions on $\mathcal{X} = [0,1]$. Let $\|f\|_p$ denote the p-norm of a function $f : \mathcal{X} \to \mathbb{R}$, where $\|f\|_p = \big(\int_x |f(x)|^p \mathrm{d}x\big)^{1/p}$ for $1 \leq p < \infty$, and $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

1) What is the ϵ-covering number of $\mathcal{H}$ with respect to $\|\cdot\|_p$ for $p = 1$, $p = 2$ and $p = \infty$ (you can assume that ϵ is small)

2) Derive an explicit expression for the growth function of $\mathcal{H}$.

3) What is the VC-dimension of $\mathcal{H}$?

4) (optional puzzle) What is the empirical Rademacher complexity of $\mathcal{H}$ for $\mathcal{D}_m = \{\frac{1}{m+1}, \frac{2}{m+1}, \dots, \frac{m}{m+1}\}$

5) (optional puzzle) What is the Rademacher complexity of $\mathcal{H}$ when p is the uniform distribution?

6) (optional puzzle) Can you find a distribution for which the Rademacher complexity is *lower*? Can you find a distribution for which the Rademacher complexity is *higher*?

2 Missing Proofs.

Prove **Hoeffding's Lemma**:
Let Z be a random variable that takes values in $[a,b]$ and fulfills $\mathbb{E}[Z] = 0$. Then, for every $\lambda > 0$,

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}.$$

Hints:
1) First show that $\mathbb{E}[e^{\lambda X}] \leq \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}$ by using the definition of convexity on the (convex) function $\exp(x)$.
2) For $h = \lambda(b-a)$, find $L(h)$ such that $\frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{L(h)}$.
3) Show $L(h) \leq \frac{h^2}{8}$ (which completes the proof).

3 Fun with Bounds.

The generalization bounds discussed in the lecture are well thought through and probably don't allow trivial improvements. Let's try anyway:

a) According to the SVM bound, small $\|w\|$ are preferable over large $\|w\|$. Luckily, for SVMs without bias term we can make $\|w\|$ arbitrary small, because the classification rule $c(x) = \operatorname{sign}\langle w, x\rangle$ does not change if we replace w by αw for arbitrarily small $\alpha > 0$.

b) Here's another way to make $\|w\|$ smaller: define a feature function $\phi(x) = \alpha x$ for $\alpha > 1$ before learning. It's easy to check that the solution learned by the SVM will be $w' = w/\alpha$, where w is the solution for the SVM on the original data. Everything else, in particular the learned linear function is preserved ($\langle w', \phi(x)\rangle = \langle w, x\rangle$).

c) The bound for finite hypothesis sets holds for all hypotheses. Clearly, this is wasteful, because most hypotheses will not yield good results even on the training set and will therefore never be used. In practice, one should therefore always use the bound with a hypothesis set of the form $\mathcal{H}' := \{f \in \mathcal{H} : \hat{\mathcal{R}}(f) \leq \eta\}$ for a user-specified threshold η, which typically can make $\mathcal{H}'$ much smaller than the original $\mathcal{H}$.

d) In general, the VC-dimension for linear classifiers in $\mathbb{R}^d$ is $d+1$, because that's the number of points that can be labeled arbitrarily with a function of the form $\text{sign}(\langle w, x\rangle)$. Remember, though, that in the definition of VC-dimension, we are free to choose the data points any way we want. By arranging points such that at least 3 points lie on a line, not all labeling are possible anymore, so the VC-dimension is reduced, and we obtain improved bounds.

e) A hypothesis set, $\mathcal{H} : \{\mathcal{X} \to \{\pm 1\}\}$, has VC-dimension d, if we can assign all possible 2^d labelings to a set of d points using hypotheses in $\mathcal{H}$. The moment even one of the labelings is not possible, the VC-dimension is automaticaly smaller. We can therefore obtain a lower VC dimension (and therefore a tighter bound) by removing, e.g., the function that assigns label +1 to every datapoint ($f(x) = +1$) from $\mathcal{H}$. Note that we won't need this hypothesis later anyway: it corresponds to the case when all training examples have the same training label, and we won't be able to learn a reasonable classifier in that situation anyway.

For each case, give a short explanation why the argument is not valid (or argue convincingly that it is).

4 Do-it-yourself bounds.

Prove your own generalization bound for finite hypothesis sets. Follow the same steps as in the lecture, but use Chebyshev's inequality instead of Hoeffding's. What's the main differences? Can you imagine a situation where one would prefer your bound over the classical one?

5 How tight is my bound?

The generalization bound

$$\forall h \in \mathcal{H} \quad \mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$$

holds (with probability $1-\delta$) for arbitray data distributions and (finite) hypothesis sets $\mathcal{H}$.
Try to find out how *tight* it is: for each

a) $m = 2$, $k = 2$, $\delta = \frac{1}{2}$

b) $m = 10$, $k = 2$, $\delta = \frac{1}{10}$

c) $m = 10$, $k = 10$, $\delta = \frac{1}{10}$

d) (bonus) arbitrary $m \geq 2$, $k \geq 2$ and $\delta > 0$.

define a data distribution on $\mathcal{X} \times \mathcal{Y}$ and a hypothesis set $\mathcal{H} \subset \{\mathcal{X} \to \mathcal{Y}\}$ with $|\mathcal{H}| = k$ such that $\mathcal{R}(h) - \hat{\mathcal{R}}(h)$ is as big as possible for a proportion of at least $(1-\delta)$ of possible i.i.d.training sets. In each case compare the value to the corresponding bound. For d), can you make the inequality an *equality*?

Hint: ideally use finite $\mathcal{X} \subset \mathbb{R}$ and $\mathcal{Y} = \{\pm 1\}$, to simplify the discussion about solutions. It might also help to look at the proof of the bound to understand which situation makes each step of the proof as loose as possible.

e) (optional puzzle) In individual cases it can happen that for a learned classifier h we have $\mathcal{R}(h) < \hat{\mathcal{R}}(h)$. Can you find a situation in which this is true uniformly, i.e. with probability $1-\delta$, for all $h \in \mathcal{H}$ we have $\mathcal{R}(h) < \hat{\mathcal{R}}(h)$?