IST Austria: Statistical Machine Learning 2018/19

Christoph Lampert <chl@ist.ac.at>
TAs: Nikola Konstantinov <nkonstan@ist.ac.at>,
Mary Phuong <bphuong@ist.ac.at>
Amelie Royer <aroyer@ist.ac.at>

Exercise Sheet 5/6 (due date 12/11/2018)

1 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).

2 Fun with Bounds.

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

a) According to the SVM bounds 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) Another way to make $\|w\|$ smaller: define a feature function $\phi(x) = \alpha x$ for $\alpha > 1$, the before learning. According to exercise sheet 4, the SVM solution w' will be w/α, 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 is 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 $\operatorname{sign}(\langle w, x\rangle)$. This changes, however, if we arrange such that at least 3 points lie on a line. Then, not all labeling are possible anymore, so the VC-dimension is reduced, and we obtain an improved bound.

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 better bound (i.e.with lower VC dimension) by removing, e.g., the function that assigns label $+1$ to every datapoint ($f(x) = +1$) from $\mathcal{H}$. Note that we probably won't need this hypothesis later anyway: it corresponds to the case when all training examples have the same training label, and we cannot learn a reasonable classifier in that situation anyway.

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

3 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?

4 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 $\mathcal{R}(h) - \hat{\mathcal{R}}$ 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.

e) (bonus bonus) 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 $\mathcal{R}(h) < \hat{\mathcal{R}}(h)$?

5 (bonus) Lower bounds?

All generalization bounds we saw are *upper bounds*. Imagine you want to show a *lower bound* on $|\mathcal{R} - \hat{\mathcal{R}}|$. What form of the statement would you be satisfied with?

Hint: really try to come up with something, don't take something from Google.

6 Practical Experiments VIII

In the lecture we saw **Hoeffding's inequality**: For $Z_1, \ldots, Z_m \overset{i.i.d.}{\sim} p$ and $\mu = \mathbb{E}[Z_i]$,

$$\mathbb{P}[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon] \leq 2e^{-m\frac{\epsilon^2}{(b-a)^2}},$$

For a Bernoulli variables Z_i write a program that for fixed μ, ϵ and m:

- estimates the left hand side from 100 random experiments,
- computes the right hand side value,
- plots the values in a graph with sample size on the x-axis and the bound or estimate on the y-axis.

Do this for a) $\mu = 0.5$, $\epsilon = 0.25$, $m = 1, 5, 10, 15, \ldots, 50$, b) $\mu = 0.1$, $\epsilon = 0.25$, $m = 1, 5, 10, 15, \ldots, 50$,
c) $\mu = 0.5$, $\epsilon = 0.05$, $m = 10, 50, 100, 150, \ldots, 500$, d) $\mu = 0.1$, $\epsilon = 0.05$, $m = 10, 50, 100, 150, \ldots, 500$.