**10-806 Foundations of Machine Learning and Data Science**

Maria-Florina Balcan Lecture 7: September 30th, 2015

---

# 1 Concentration Inequalities (Tail Inequalities)

Recall our concentration inequalities from last time.

Consider a coin of bias $p$ flipped $m$ times. Let $S$ be the number of observed number of heads. So $\mathbf{E}[S/m] = p$.

*Hoeffding bounds* state that for any $\epsilon \in [0, 1]$,

1. $\Pr[\frac{S}{m} > p + \epsilon] \leq e^{-2m\epsilon^2}$, and
2. $\Pr[\frac{S}{m} < p - \epsilon] \leq e^{-2m\epsilon^2}$.

*Chernoff bounds* state that under the same conditions,

1. $\Pr[\frac{S}{m} > p(1 + \epsilon)] \leq e^{-mp\epsilon^2/3}$, and
2. $\Pr[\frac{S}{m} < p(1 - \epsilon)] \leq e^{-mp\epsilon^2/2}$.

# 2 The Non-realizable Case

So far, we have been assuming the target function belongs to the class $C$. In general, the target function might not be in the class of functions we consider. Formally, in the non-realizable or agnostic passive supervised learning setting, we assume that the input to a learning algorithm is a set $S$ of labeled examples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. We assume that these examples are drawn i.i.d. from some fixed but unknown distribution $D$ over the the instance space $X$ and that they are labeled by some target concept $c^*$. So $y_i = c^*(x_i)$. *However, $c^*$ might not belong to $C$.* The goal is just as in the realizable case to do optimization over the given sample $S$ in order to find a hypothesis $h : X \to \{0, 1\}$ of low error over whole distribution $D$. The error of $h$ is defined as

$$err(h) = \Pr_{x \sim D}(h(x) \neq c^*(x)).$$

We denote by

$$err_S(h) = \Pr_{x \sim S}(h(x) \neq c^*(x))$$

the empirical error over the sample. Our goal is to compete with the best function (the function of smallest true error rate) in some concept class $C$.

A natural hope is that picking a concept $c$ with a small observed error rate gives us small true error rate. It is therefore useful to find a relationship between *observed* error rate for a sample and the *true* error rate.

We now present sample complexity guarantees for the non-realizable case, where we will use Hoeffding and Chernoff bounds to extend the results we previously showed for the realizable case.

## 2.1 Simple sample complexity results for finite hypotheses spaces

We can use the Hoeffding bounds to show the following:

**Theorem 1** *Let $C$ be a finite hypothesis space. Let $D$ be an arbitrary, fixed unknown probability distribution over $X$ and let $c^*$ be an arbitrary unknown target function. For any $\epsilon$, $\delta > 0$, if we draw a sample $S$ from $D$ of size*

$$m \geq \frac{1}{2\epsilon^2}\left(\ln(2|C|) + \ln\left(\frac{1}{\delta}\right)\right),$$

*then probability at least $(1-\delta)$, all hypotheses $h$ in $C$ have*

$$|err(h) - err_S(h)| \leq \epsilon. \tag{1}$$

*Proof:* Let us fix a hypothesis $h$. By Hoeffding, we get that the probability that its observed error within $\epsilon$ of its true error is at most $2e^{-2m\epsilon^2} \leq \delta/|C|$. By union bound over all all $h$ in $C$, we then get the desired result. ■

**Note:** A statement of type one is called a *uniform convergence* result. It implies that the hypothesis that minimizes the empirical error rate will be very close in generalization error to the best hypothesis in the class. In particular if $\widehat{h} = argmin_{h \in C} err_S(h)$ we have $err(\widehat{h}) \leq err(h^*) + 2\epsilon$, where $h^*$ is a hypothesis of smallest true error rate.

**Note:** The sample size grows quadratically with $1/\epsilon$. Recall that the learning sample size in the realizable (PAC) case grew only linearly with $1/\epsilon$.

**Note:** Another way to write the bound in Theorem 1 is as follows:

For any $\epsilon$, $\delta > 0$, if we draw a sample from $D$ of size $m$ then with probability at least $1-\delta$, all hypotheses $h$ in $C$ have

$$err(h) \leq err_S(h) + \sqrt{\frac{\ln(2|C|) + \ln\left(\frac{1}{\delta}\right)}{2m}}$$

This is the more "statistical learning theory style" way of writing the same bound.

We can get rid of that pesky $\epsilon^2$ by relaxing our goal, to say that for hypotheses whose true error is greater than $\epsilon$, we are satisfied if their observed error comes just within a *factor of 2*. If you think about it, this is often all we need since we don't care so much about the high-error hypotheses. Now we can use Chernoff bounds.

**Theorem 2** *Let $C$ be a finite hypothesis space. Let $D$ be an arbitrary, fixed unknown probability distribution over $X$ and let $c^*$ be an arbitrary unknown target function. For any $\epsilon$, $\delta > 0$, if we draw a sample $S$ from $D$ of size*

$$m \geq \frac{6}{\epsilon}[\ln(|C|) + \ln(1/\delta)]$$

*then with probability at least $1-\delta$, all $h \in C$ with $err(h) > 2\epsilon$ have $err_S(h) > \epsilon$, and all $h \in C$ with $err(h) \leq \epsilon/2$ have $err_S(h) \leq \epsilon$. Thus, if the hypothesis $h^*$ of minimum true error has $err(h^*) \leq \epsilon/2$ then the hypothesis $\hat{h}$ of minimum* empirical *error has $err(\hat{h}) \leq 2\epsilon$.*

Or, to be analogous to Note 3, given $m$ examples, with probability at least $1-\delta$, all $h \in C$ with $err(h) > \frac{12}{m}\ln(2|C|/\delta)$ satisfy $err_S(h) \geq err(h)/2$ and all $h \in C$ with $err(h) < \frac{3}{m}\ln(2|C|/\delta)$ satisfy $err_S(h) < \frac{6}{m}\ln(2|C|/\delta)$.

*Proof of Theorem:* If $err(h) = p \geq 2\epsilon$ and we want empirical error at least $p/2 \geq \epsilon$ with confidence $\delta'$, we can solve $e^{-mp/8} \leq \delta'$ to get that it suffices to have $m \geq \frac{8}{p}\ln(1/\delta')$ (and we can replace $p$ with $2\epsilon$). On the other hand, if $err(h) = p \leq \epsilon/2$, we can rewrite our additive error goal by saying that we want the observed error to be no more than $\frac{\epsilon}{2}(1+1)$, which by Chernoff bounds implies that $m \geq \frac{6}{\epsilon}\ln(1/\delta')$ examples suffice. So, setting $\delta' = \delta/|C|$ we get that:

$$m \geq \frac{6}{\epsilon}[\ln(|C|/\delta)]$$

examples are sufficient for both conditions of the theorem.

## 2.2 Sample complexity results for infinite hypothesis spaces

We now consider the case of VC-dimension-based sample complexity for the non-realizable case, which will apply when $|C|$ may be infinite. Recall that for a class $C$ and set of examples $S = \{x_1, \cdots, x_m\}$, we define

$$C(S) = \{(c(x_1), \cdots, c(x_m)) ; c \in \mathcal{C}\}.$$

Also, for any natural number $m$, we consider $C[m]$ to be the maximum number of ways to split $m$ points using concepts in $C$, that is

$$C[m] = \max\{|C(S)| ; |S| = m, S \subseteq X\}.$$

Earlier we proved the following result:

**Theorem 3** *Let $C$ be an arbitrary hypothesis space. Let $D$ be an arbitrary, fixed unknown probability distribution over $X$ and let $c^*$ be an arbitrary unknown target function. For any $\epsilon$, $\delta > 0$, if we draw a sample $S$ from $D$ of size*

$$m > \frac{2}{\epsilon} \cdot \left[\log_2(2 \cdot \mathcal{C}[2m]) + \log_2\left(\frac{1}{\delta}\right)\right] \tag{2}$$

*then with probability $(1-\delta)$, all bad hypothesis in $\mathcal{C}$ (with error $> \epsilon$ with respect to $c$ and $D$) are inconsistent with the data.*

For the non-realizable case, the analogous result is as follows:

**Theorem 4** *Let $C$ be an arbitrary hypothesis space. Let $D$ be an arbitrary, fixed unknown probability distribution over $X$ and let $c^*$ be an arbitrary unknown target function. For any $\epsilon$, $\delta > 0$, if we draw a sample $S$ from $D$ of size $m > (8/\epsilon^2)[\ln(2C[2m]) + \ln(1/\delta)]$ then with probability $1-\delta$, all $h$ in $C$ have $|err_D(h) - err_S(h)| < \epsilon$.*

*Proof Sketch:* We just need to redo the proof we had for the realizable case using Hoeffding bounds.

Draw $2m$ examples. Let $B$ be the event that on first $m$, there exists hypothesis in $C$ with empirical and true error that differ by at least $\epsilon$. (This is what we want to bound for the theorem). Let $B'$ be

the event that there exists a concept in $C$ whose empirical error on 1st half differs from empirical error on 2nd half by at least $\epsilon/2$.

For large enough $m$, we have $\Pr[B'|B] \geq 1/2$ so $\Pr[B] \leq 2 \cdot \Pr[B']$. Now we have to show that $\Pr[B']$ is low.

As before, let's first pick $S$, $S'$, then we do the symmetrization (or swapping). Once $S \cup S'$ is determined, there are only $C[2m]$ hypotheses we need to worry about. Using Hoeffding bounds, we can show that for any fixed $h$, $S$, and $S'$,

$$\Pr_{Swap}[|err_T(h) - err_{T'}(h)| > \epsilon/2] \leq e^{-\epsilon^2 m/8}.$$

Performing a union bound over all $C[2m]$ hypotheses as in Theorem 3, we get the desired result. ■

### 2.3 Combining with VC-dimension

As in the realizable case, we can apply Sauer's lemma to get a bound in terms of VC-dimension that only has $m$ on one side.

**Theorem 5** *Let $C$ be an arbitrary hypothesis space of VC-dimension $d$ and let $D$ be an arbitrary unknown probability distribution over the instance space. There exists constant $c > 0$ such that for any $\epsilon$, $\delta > 0$, if we draw a sample $S$ from $D$ of size $m$ satisfying*

$$m \geq \frac{c}{\epsilon^2}\left[d \ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right],$$

*then with probability at least $1 - \delta$, all the hypotheses in $C$ satisfy $|err_D(h) - err_S(h)| \leq \epsilon$.*

Interestingly, here there are certain tricks that can be used to get rid of the $\log(1/\epsilon)$ term. See Anthony and Bartlett [1].

### 2.4 Sample complexity lower bounds

As in the realizable case, one can also get lower bounds for the sample complexity of learning in the agnostic case. Let $OPT(C) = \min_{h \in C} err_D(h)$.

**Theorem 6** *Any algorithm for agnostically learning a concept class $C$ of VC dimension $d$, that achieves error at most $OPT(C) + \epsilon$ with probability at least $1 - \delta$, must for some constant $c > 0$ use at least*

$$\frac{c}{\epsilon^2}\left[d + \ln\left(\frac{1}{\delta}\right)\right]$$

*examples in the worst case.*

## References

[1] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 1999.