Machine Learning Lecture Notes

Yevgeny Seldin

August 25, 2015

Contents

Chapter 1

Concentration of Measure Inequalities

1.1 Introduction

Machine Learning is (to a significant degree) about using past knowledge to make predictions about the future. Making successful predictions about the future based on past knowledge is possible only when the future is in some way similar to the past and up to the degree that it is similar to the past. One of the simplest and most commonly used ways to model the similarity between past and future are i.i.d. (independent identically distributed) processes. We assume that our past data are independent identically distributed samples from an unknown, but fixed distribution, and that the future observations will be sampled independently from the same distribution. Learning the parameters of this distribution (expectation, variance, etc.) based on past observations then allows to make better predictions about future observations. Thus, concentration of random variables around their expected values ("concentration of measure") plays the central role in theoretical foundations of learning theory. We start with presenting a number of basic results about the concentration of measure phenomenon and then demonstrate applications of these results to learning.

1.2 Markov's Inequality

Markov's Inequality is the simplest and relatively weak concentration inequality. Nevertheless, it forms the basis for many much stronger inequalities that we will see in the sequel.

Theorem 1 (Markov's Inequality)**.** *For any non-negative random variable X and $\varepsilon > 0$:*

$$\mathbb{P}\left\{X \geq \varepsilon\right\} \leq \frac{\mathbb{E}\left[X\right]}{\varepsilon}.$$

Proof. By definition of $\mathbb{E}\left[X\right]$ we have:

$$\mathbb{E}\left[X\right] = \int_0^\infty xp(x)dx = \int_0^\varepsilon xp(x)dx + \int_\varepsilon^\infty xp(x)dx \geq \int_\varepsilon^\infty xp(x)dx \geq \varepsilon \int_\varepsilon^\infty p(x)dx = \varepsilon\mathbb{P}\left\{X \geq \varepsilon\right\}.$$

By dividing both sides by ε we obtain the inequality. □

1.3 Hoeffding's Inequality

Hoeffding's inequality is a much more powerful concentration result.

Theorem 2 (Hoeffding's Inequality)**.** *Let $X_1, \dots, X_n$ be independent real-valued random variables, such that for each $i \in \{1, \dots, n\}$ there exist $a_i \leq b_i$, such that $\mathbb{P}\left\{a_i \leq X_i \leq b_i\right\} = 1$. Then for every $\varepsilon > 0$:*

$$\mathbb{P}\left\{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] > \varepsilon\right\} \leq e^{-2\varepsilon^2 / \sum_{i=1}^n (b_i - a_i)^2} \tag{1.1}$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] < -\varepsilon\right\} \leq e^{-2\varepsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}. \tag{1.2}$$

By taking a union bound of the events in (1.1) and (1.2) we obtain the following corollary.

Corollary 3. *Under the assumptions of Theorem 2:*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right| > \varepsilon\right\} \leq 2e^{-2\varepsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}. \tag{1.3}$$

Equations (1.1) and (1.2) are known as "one-sided Hoeffding's inequalities" and (1.3) is known as "two-sided Hoeffding's inequality".

If we assume that X_i-s are identically distributed and belong to the $[0,1]$ interval we obtain the following corollary.

Corollary 4. *Let $X_1, \dots, X_n$ be independent random variables, such that $\mathbb{P}\{X_i \in [0,1]\} = 1$ and $\mathbb{E}[X_i] = \mu$ for all i, then:*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu > \varepsilon\right\} \leq e^{-2n\varepsilon^2} \tag{1.4}$$

and

$$\mathbb{P}\left\{\mu - \frac{1}{n}\sum_{i=1}^{n} X_i > \varepsilon\right\} \leq e^{-2n\varepsilon^2}. \tag{1.5}$$

The proof of Hoeffding's inequality is based on Hoeffding's lemma.

Lemma 5 (Hoeffding's Lemma)**.** *Let X be a random variable, such that $\mathbb{P}\{X \in [a,b]\} = 1$. Then for any $\lambda \in \mathbb{R}$:*

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

See a separate handout for a proof of the lemma.

Proof of Theorem 2. We prove the first inequality in Theorem 2. The second inequality follows by applying the first inequality to $-X_1, \dots, -X_n$. The proof is based on Chernoff's bounding technique. For any $\lambda > 0$ the following holds:

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] > \varepsilon\right\} &= \mathbb{P}\left\{e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)} > e^{\lambda\varepsilon}\right\}\\ &\leq \frac{\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)}\right]}{e^{\lambda\varepsilon}},\end{aligned}$$

where the first line holds since $e^{\lambda x}$ is a monotonously increasing function for $\lambda > 0$ and the second line holds by Markov's inequality. We now take a closer look at the nominator:

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)}\right] &= \mathbb{E}\left[e^{\left(\sum_{i=1}^{n} \lambda(X_i - \mathbb{E}[X_i])\right)}\right]\\ &= \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda(X_i - \mathbb{E}[X_i])}\right]\\ &= \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \qquad (1.6)\\ &\leq \prod_{i=1}^{n} e^{\lambda^2(b_i-a_i)^2/8} \qquad (1.7)\\ &= e^{(\lambda^2/8)\sum_{i=1}^{n}(b_i-a_i)^2},\end{aligned}$$

where (1.6) holds since $X_1, \ldots, X_n$ are independent and (1.7) holds by Hoeffding's lemma applied to a random variable $Z_i = X_i - \mathbb{E}[X_i]$ (note that $\mathbb{E}[Z_i] = 0$ and that $Z_i \in [a_i - \mu_i, b_i - \mu_i]$ for $\mu_i = \mathbb{E}[X_i]$). *Put attention to the crucial role that independence of $X_1, \ldots, X_n$ plays in the proof! Without independence we would not be able to exchange the expectation with the product and the proof would break down!* To complete the proof we substitute the bound on the expectation into the previous calculation and obtain:

$$\mathbb{P}\left\{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] > \varepsilon\right\} \leq e^{(\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon}.$$

This expression is minimized by

$$\lambda^* = \arg\min_\lambda e^{(\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon} = \arg\min_\lambda \left((\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon\right) = \frac{4\varepsilon}{\sum_{i=1}^n (b_i - a_i)^2}.$$

It is important to note that the best choice of λ does not depend on the sample. In particular, it allows to fix λ before observing the sample. By substituting λ^* into the calculation we obtain the result of the theorem. □

Chapter 2

Supervised Learning

In this chapter we derive a number of generalization bounds for supervised learning. We start with a formal setting of supervised learning, formulate it in the language of probability theory, and then provide the analysis.

2.1 Supervised Learning Setting

We start with a bunch of notations.

- $\mathcal{X}$ - the sample space.
- $\mathcal{Y}$ - the label space.
- $X \in \mathcal{X}$ - unlabeled sample.
- $(X, Y) \in (\mathcal{X} \times \mathcal{Y})$ - labeled sample.
- $S = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ - a training set. We assume that (X_i, Y_i) pairs in S are sampled i.i.d. according to an unknown, but fixed distribution $p(X, Y)$.
- $h : \mathcal{X} \to \mathcal{Y}$ - a hypothesis, which is a function from $\mathcal{X}$ to $\mathcal{Y}$.
- $\mathcal{H}$ - a hypothesis set.

The classical supervised learning acts according to the following protocol:

1. The learner gets a training set S of size n sampled i.i.d. according to $p(X, Y)$.
2. The learner returns a prediction rule h.
3. New instances (X, Y) are sampled according to $p(X, Y)$, but only X is observed. h is used to predict the unobserved Y and the quality of h is judged by the quality of predictions, as defined below.

We define:

- $\ell(Y', Y)$ - the loss function for predicting Y' instead of Y.
- $\hat{L}(h, S) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$ - the empirical loss (or error) of h on S. Sometimes, when it is clear from the context, we will omit S and write $\hat{L}(h)$.
- $L(h) = \mathbb{E}\left[\ell(h(X), Y)\right]$ - the expected loss (or error) of h, where the expectation is taken with respect to $p(X, Y)$.

The goal of the learner is to return h that minimizes $L(h)$, which is the expected error on future samples.

2.2 Generalization Bound for a Single Hypothesis

We start with the simplest case, where we have a sample S and a single prediction rule h. We are interested in the quality of h, measured by $L(h)$, but all we can measure is $\hat{L}(h,S)$. What can we say about $L(h)$ based on $\hat{L}(h,S)$? Let $Z_i = \ell(h(X_i),Y_i)$ be the loss of h on the sample (X_i,Y_i). Note that Z_i is a random variable and that since (X_i,Y_i)-s are sampled i.i.d., $Z_1,\ldots,Z_n$ are i.i.d. random variables with $\mathbb{E}[Z_i] = L(h)$. And we also have $\hat{L}(h,S) = \frac{1}{n}\sum_{i=1}^n Z_i$. Thus we have the following result.

Theorem 6. *Assume that ℓ is bounded in the $[0,1]$ interval ($\ell(Y',Y) \in [0,1]$ for all Y', Y), then for a single h and any $\delta \in (0,1)$ we have:*

$$\mathbb{P}\left\{L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right\} \leq \delta \tag{2.1}$$

and

$$\mathbb{P}\left\{\left|L(h) - \hat{L}(h,S)\right| > \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right\} \leq \delta. \tag{2.2}$$

Proof. For (2.1) take $\varepsilon = \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}$ in (1.5) and rearrange the terms. Equation (2.2) follows in a similar way from the two-sided Hoeffding's inequality. Note that in (2.1) we have $\frac{1}{\delta}$ and in (2.2) we have $\frac{2}{\delta}$. □

There is an alternative way to read equation (2.1) - with probability greater than $1-\delta$ we have:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}.$$

In other words, with probability greater than $1-\delta$ the sample S is *representative* of the performance of h and $L(h)$ is close to $\hat{L}(h,S)$ up to $\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}$ and with probability at most δ the sample S is not representative of the performance of h and we can say nothing about $L(h)$ by observing $\hat{L}(h,S)$.

Theorem 6 is directly analogous to the problem of estimating a bias of a coin based on coin flip outcomes. Lets say that we have a fair coin, but we do not know that this is a fair coin. With high probability, if we flip the coin 1000 times the empirical bias will be close to $\frac{1}{2}$, but there is always a small probability that we flip the coin 1000 times and get all heads or other non-representative outcome for this coin. And if this happens we are doomed - there is nothing we can do when the sample does not represent the reality faithfully. Fortunately for us, this happens with a small probability that decreases exponentially with the growth of the sample size n.

Whether we use the one-sided bound (2.1) or the two-sided bound (2.2) depends on the situation. Generally we care about the upper bound on the expected performance of the prediction rule (2.1), however, if we want to show that we did the best we could we may need to use (2.2).

2.3 Generalization Bound for Finite Hypothesis Classes

Taking a single prediction rule and estimating its expected error is a validation procedure. In learning we operate with multiple prediction rules and pick the best one based on the sample (like the best separating hyperplane in SVMs). So what can we say about the expected error $L(h)$, also called generalization error, when h was selected from a hypothesis class $\mathcal{H}$? The tricky point is that S may be representative for h_1, but not representative for h_2 and vice versa, and without any additional knowledge the best we can do is to apply a union bound.

Theorem 7. *Assume that ℓ is bounded in the $[0,1]$ interval and that $|\mathcal{H}| = M$. Then for any $\delta \in (0,1)$ we have:*

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right\} \leq \delta. \tag{2.3}$$

Proof.

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}\right\} \leq \sum_{h\in\mathcal{H}} \mathbb{P}\left\{L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}\right\} \leq \sum_{h\in\mathcal{H}} \frac{\delta}{M} = \delta,$$

where the first inequality is by the union bound and the second is by Hoeffding's inequality. □

Another way of reading Theorem 7 is - with probability greater than $1-\delta$ for all $h \in \mathcal{H}$:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}, \tag{2.4}$$

which means that no matter which h from $\mathcal{H}$ is returned by the algorithm, with high probability we have guarantee (2.4) on its expected performance. The price that we paid for considering M hypotheses instead of a single one is $\ln M$. Note that it grows only logarithmically with M!

Similar to theorem 6 it is possible to derive a two-sided bound on the error.

2.4 Occam's Razor Bound

Now we take a deeper look at Hoeffding's inequality. It says that

$$\mathbb{P}\left\{L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right\} \leq \delta,$$

where δ is the probability that things go wrong and $L(h)$ happens to be far away from $\hat{L}(h,S)$ because S is not representative for the performance of h. But there is an interdependence between the probability that things go wrong and the requirement on the closeness between $L(h)$ and $\hat{L}(h,S)$. If we want them to be very close (meaning that $\ln\left(\frac{1}{\delta}\right)$ is small) then δ will be large, but if we can allow larger distance then δ can be smaller.

So, δ can be seen as our "confidence budget" (or, more precisely, "uncertainty budget") - the probability that we allow things go completely wrong. The idea behind Occam's Razor bound is to distribute this budget unevenly among the hypotheses in $\mathcal{H}$.

Theorem 8. *Let ℓ be bounded in $[0,1]$, let $\mathcal{H}$ be a countable hypothesis set and let $p(h)$ be such that $\sum_{h\in\mathcal{H}} p(h) \leq 1$ and $p(h)$ is independent of the sample. Then:*

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{p(h)\delta}\right)}{2n}}\right\} \leq \delta.$$

Proof.

$$\begin{aligned}\mathbb{P}\left\{\exists h \in \mathcal{H} : L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{p(h)\delta}\right)}{2n}}\right\} &\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left\{L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{p(h)\delta}\right)}{2n}}\right\} \\ &\leq \sum_{h\in\mathcal{H}} p(h)\delta \\ &\leq \delta,\end{aligned}$$

where the first inequality is the union bound, the second inequality is by Hoeffding's inequality, and the last inequality is by our assumption on $p(h)$. Note that $p(h)$ has to be selected before we observe the sample (or, in other words, independently of the sample), otherwise the second inequality does not hold. □

Another way of reading Theorem 8 is that with probability greater than $1-\delta$, for all $h \in \mathcal{H}$:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{p(h)\delta}\right)}{2n}}.$$

Again, note that the bound on $L(h)$ depends both on $\hat{L}(h,S)$ and on $p(h)$. Therefore, according to the bound, the best generalization is achieved by h that optimizes the trade-off between empirical performance $\hat{L}(h,S)$ and $p(h)$, where $p(h)$ can be interpreted as a complexity measure or a prior belief. Also, note that $p(h)$ can be designed arbitrarily, but it should be independent of the sample S. If $p(h)$ happens to put more mass on h-s with low $\hat{L}(h,S)$ the bound will be tighter, otherwise the bound will be loser, but it will still be a valid bound. But we cannot readjust $p(h)$ after observing S! Some considerations behind the choice of $p(h)$ are provided in Section 2.4.1.

2.4.1 Applications of Occam's Razor bound

We now consider two applications of Occam's Razor bound.

Generalization bound for finite hypotheses spaces

An immediate corollary from Occam's razor bound is the generalization bound for finite hypotheses classes that we have already seen earlier in the course.

Corollary 9. *Let $\mathcal{H}$ be a finite hypotheses class of size M, then*

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : L(h) > \hat{L}(h,S) + \sqrt{\frac{\ln(M/\delta)}{2n}}\right\} \leq \delta.$$

Proof. We set $p(h) = \frac{1}{M}$ (which means that we distribute the confidence budget δ uniformly among the hypotheses in $\mathcal{H}$) and apply Theorem 8. □

Generalization bound for binary decision trees

Theorem 10. *Let $\mathcal{H}_d$ be the set of binary decision trees of depth d and let $\mathcal{H} = \bigcup_{d=0}^{\infty} \mathcal{H}_d$ be the set of binary decision trees of unlimited depth. Let $d(h)$ be the depth of tree (hypothesis) h. Then*

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : L(h) > \hat{L}(h) + \sqrt{\frac{\ln\left(2^{2^{d(h)}} 2^{d(h)+1}/\delta\right)}{2n}}\right\} \leq \delta.$$

Proof. We first note that $|\mathcal{H}_d| = 2^{2^d}$. We define $p(h) = \frac{1}{2^{d(h)+1}} \frac{1}{2^{2^{d(h)}}}$. The first part of $p(h)$ distributes confidence budget δ among $\mathcal{H}_d$-s (we can see it as $p(\mathcal{H}_d) = \frac{1}{2^{d(h)+1}}$ - the share of confidence budget that goes to $\mathcal{H}_d$) and the second part of $p(h)$ distributes confidence budget uniformly within $\mathcal{H}_d$. Since $\sum_{d=0}^{\infty} \frac{1}{2^{d+1}} = 1$, the assumption $\sum_{h \in \mathcal{H}} p(h) \leq 1$ is satisfied. The result follows by application of Theorem 8. □

Note that the bound depends on $\ln\left(\frac{1}{p(h)\delta}\right)$ and the dominating term in $\frac{1}{p(h)}$ is $2^{2^{d(h)}}$. We could have selected a different distribution of confidence over $\mathcal{H}_d$-s, for example, $p(\mathcal{H}_d) = \frac{1}{(d+1)(d+2)}$ (for which we also have $\sum_{d=0}^{\infty} \frac{1}{(d+1)(d+2)} = 1$), which is perfectly fine, but does not make significant difference for the bound. The dominating complexity term $\ln\left(2^{2^{d(h)}}\right)$ comes from uniform distribution of confidence within $\mathcal{H}_d$, which makes sense unless we have some prior information about the problem. In absence of such information there is no reason to give preference to any of the trees within $\mathcal{H}_d$, because $\mathcal{H}_d$ is symmetric.