CS 7545: Machine Learning Theory **Fall 2018**

Lecture 4: Concentration Inequalities

Lecturer: Jacob Abernethy *Scribes: Zihao Hu, Nathan Hatch*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

4.1 Concentration Inequalities

4.1.1 Review from last lecture

Theorem 4.1 *(Markov's Inequality) For a random variable $X \geq 0$*

$$Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \tag{4.1}$$

This is "the most basic deviation bound".

Theorem 4.2 *(Chebyshev's Inequality) For any random variable with mean μ and variance σ^2*

$$Pr(|X - \mu| > t\sigma) \leq \frac{1}{t^2} \tag{4.2}$$

This deviation bound is also very general. It works for any random variable with finite mean and variance. It's slightly better than Markov's inequality, but still "not good enough".

4.1.2 Hoeffding's Inequality

Hoeffding's Inequality will give us a deviation bound that decays exponentially. This is much better than $1/t$ or $1/t^2$. It is also non-asymptotic (unlike the central limit theorem), which is nice for engineering purposes when you don't have an infinite amount of data.

Before stating the theorem, we state a lemma which will be used in the proof.

Lemma 4.3 *(Hoeffding's Lemma) Let X be a random variable such that $a \leq X \leq b$, $\mathbb{E}[X] = 0$. Then*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \tag{4.3}$$

Proof: See Foundations of Machine Learning book, p. 369. ■

Theorem 4.4 *(Hoeffding's Inequality) Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ and $\mathbb{E}[X_i] = 0$. Then*

$$Pr\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right) \tag{4.4}$$

Remark Note that there is no absolute value in the theorem statement. However, "using symmetry", it is possible to argue that $Pr(|\sum X_i| > t) \leq 2Pr(\sum X_i > t)$. Also, if your random variables are bounded but not zero-mean, you can still apply the theorem to the zero-mean variables $X_i - \mathbb{E}[X_i]$.

Proof: (Chernoff Bounding Technique) For all $\lambda > 0$, the following holds:

$$\begin{aligned}
Pr\left(\sum_{i=1}^{n} X_i > t\right) &= Pr\left(\exp\left(\lambda\sum_{i=1}^{n} X_i\right) > \exp(\lambda t)\right) && \text{monotonicity of } e^{\lambda x} \\
&\le \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n} X_i\right)\right] / \exp(\lambda t) && \text{Markov's Inequality} \\
&= e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}[\exp(\lambda X_i)] && \text{independence of } X_i \\
&\le \exp(-\lambda t)\prod_{i=1}^{n}\exp\left(\frac{\lambda^2(b_i-a_i)^2}{8}\right) && \text{Hoeffding's Lemma} \\
&= \exp\left(\lambda^2\frac{\sum(b_i-a_i)^2}{8} - \lambda t\right)
\end{aligned}$$

The exponent is convex quadratic in λ. Since this is true for all $\lambda > 0$, we can choose λ to minimize the quadratic and achieve the best bound. The minimum of $p\lambda^2 + q\lambda$ is $-q^2/4p$, so we have

$$Pr\left(\sum_{i=1}^{n} X_i > t\right) \le \exp\left(\frac{-2t^2}{\sum(b_i-a_i)^2}\right)$$

■

Remark. Only one step of the proof required that these random variables X_i were bounded. In fact, there is a more general set called **sub-Gaussian distributions** which satisfy inequalities similar to Hoeffding's Lemma. The proof of Hoeffding's Inequality works just as well for all sub-Gaussian distributions.

The following corollary restates Hoeffding's Inequality in a slightly less general form from the perspective of finding the best t given a specified maximum probability of failure δ.

Corollary 4.5 *Let $X_1, \ldots, X_n$ be i.i.d. with mean μ, $-1 \le X_i - \mu \le 1$. Then for all $\delta > 0$, with probability at least $1-\delta$ we have*

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \le \sqrt{\frac{2\log(2/\delta)}{n}} \tag{4.5}$$

Proof: From Hoeffding's Inequality,

$$Pr\left(\left|\frac{1}{n}\sum(X_i-\mu)\right| > t\right) \le 2Pr\left(\sum_{i=1}^{n}(X_i-\mu) > tn\right) \le \exp\left(\frac{-2(tn)^2}{4n}\right) = 2\exp\left(\frac{-t^2 n}{2}\right) =: \delta$$

Now we just solve for t to get $t = \sqrt{\frac{2\log(2/\delta)}{n}}$. ■

4.2 Martingales

Martingales are a "generalization of sums of i.i.d. random variables". We will see that, although martingales are more general than sums of i.i.d. random variables, they obey a very similar concentration inequality.

Definition 4.6 *A sequence of random variables $Z_0, Z_1, \ldots, Z_n$ is a **martingale sequence** if $\forall i = 1, \ldots, n$, $\mathbb{E}[Z_i|Z_0, \ldots, Z_{i-1}] = Z_{i-1}$.*

Remark. Usually Z_0 will be a constant; e.g. your starting account balance.

Fact. If $Z_0, Z_1, \ldots, Z_n$ is a martingale sequence (and Z_0 is constant), then

$$\mathbb{E}[Z_n] = \mathbb{E}[\mathbb{E}[Z_n | Z_1, \ldots, Z_{n-1}]] = \mathbb{E}[Z_{n-1}] = \cdots = \mathbb{E}[Z_1] = Z_0 \tag{4.6}$$

Example. Let $X_1, \ldots, X_n$ be i.i.d. fair coin tosses, $X_i = \pm 1$. Then the following are martingale sequences:

- $Z_n := \sum_{i=1}^n X_j$
- $Z_0 := c$, $Z_n := Z_{n-1} + \delta Z_{n-1} X_{n-1}$, where $c > 0$ and $\delta \in (0,1)$ are constants. This example represents a "betting strategy" where at each round n, you bet a fixed proportion δ of your current wealth Z_{n-1}.

Theorem 4.7 *(Azuma's Inequality) Let $Z_0, Z_1, \ldots, Z_n$ be a martingale sequence such that $\forall i$, $|Z_i - Z_{i-1}| \leq c_i$. Then*

$$Pr(Z_n - Z_0 > t) \leq \exp\left(\frac{-t^2}{2\sum c_i^2}\right) \tag{4.7}$$

We will prove this next class. The proof is almost identical to the proof of Hoeffding's Inequality.