Machine Learning Notes

This note is based on Machine Learning at DIKU

December 5, 2021

Chapter 2. Concentration of Measure Inequalities

Concentration of measure inequalities are one of the main tools for analysing learning algorithms. This chapter is devoted to a number of concentration of measure inequalities that form the basis for the results discussed in later chapters.

2.1 Markov's Inequality

Markov's inequality is the simplest and relatively weak concentration inequality.

Theorem 2.1 (Markov's Inequality). *For any non-negative random variable X and $\varepsilon > 0$,*

$$\mathbf{Pr}(X \geqslant \varepsilon) \leqslant \frac{\mathbf{E}[X]}{\varepsilon}. \tag{2.1}$$

Proof. Define a random variable $Y = \mathbb{I}(X \geqslant \varepsilon)$ to be the indicator function of whether X exceeds ε. Then $Y \leqslant \frac{X}{\varepsilon}$. Since Y is a Bernoulli random variable, $\mathbf{E}[Y] = \mathbf{Pr}(Y = 1)$. We have:

$$\mathbf{Pr}(X \geqslant \varepsilon) = \mathbf{Pr}(Y = 1) = \mathbf{E}[Y] \leqslant \mathbf{E}\left[\frac{X}{\varepsilon}\right] = \frac{\mathbf{E}[X]}{\varepsilon}. \tag{2.2}$$

NB. $Y = \mathbb{I}\left(\frac{X}{\varepsilon} \geqslant 1\right) \in [0,1]$, and that $\frac{X}{\varepsilon} \in [0, +\infty)$ due to the non-negativity of X.

1) If $\frac{X}{\varepsilon} \in [0,1)$, then $Y = 0$, and $Y \leqslant \frac{X}{\varepsilon}$;
2) If $\frac{X}{\varepsilon} = 1$, then $Y = 1$, and $Y = \frac{X}{\varepsilon}$;
3) If $\frac{X}{\varepsilon} \in (1, +\infty)$, then $Y = 1$, and $Y < \frac{X}{\varepsilon}$.

Overall, $Y \leqslant \frac{X}{\varepsilon}$. Then $\mathbf{Pr}\big(X \geqslant \frac{1}{\delta}\mathbf{E}[X]\big) \leqslant \frac{\mathbf{E}[X]}{\delta \mathbf{E}[X]} = \delta$. ■

By denoting the right hand side of Markov's inequality by δ we obtain the following equivalent statement. For any non-negative random variable X,

$$\mathbf{Pr}\left(X \geqslant \frac{1}{\delta}\mathbf{E}[X]\right) \leqslant \varepsilon. \tag{2.3}$$

We note that even though Markov's inequality is weak, there are situations in which it is tight.

2.2 Chebyshev's Inequality

Our next stop is Chebyshev's inequality, which exploits variance to obtain tighter concentration.

Theorem 2.2 (Chebyshev's inequality). *For any $\varepsilon > 0$,*

$$\mathbf{Pr}(|X - \mathbf{E}[X]| \geqslant \varepsilon) \leqslant \frac{\mathbf{Var}[X]}{\varepsilon^2}. \tag{2.4}$$

Proof. The proof uses a transformation of a random variable. We have that $\mathbf{Pr}(|X-\mathbf{E}[X]|\geqslant\varepsilon) = \mathbf{Pr}\big((X-\mathbf{E}[X])^2\geqslant\varepsilon^2\big)$, because the first statement holds if and only if the second holds. In addition, using Markov's inequality and the fact that $(X-\mathbf{E}[X])^2$ is a non-negative random variable we have

$$\mathbf{Pr}(|X-\mathbf{E}[X]|\geqslant\varepsilon) = \mathbf{Pr}\left((X-\mathbf{E}[X])^2\geqslant\varepsilon^2\right) \leqslant \frac{\mathbf{E}\left[(X-\mathbf{E}[X])^2\right]}{\varepsilon^2} = \frac{\mathbf{Var}[X]}{\varepsilon^2}\,. \tag{2.5}$$

NB. By definition, for a random variable X, its expectation is given by $\mathbf{E}[X]=\sum_{x\in\mathcal{X}} x\mathbf{Pr}(X=x)$, and its variance is defined by $\mathbf{Var}[X] = \mathbf{E}\left[\left(X-\mathbf{E}[X]\right)^2\right] = \mathbf{E}[X^2]-\left(\mathbf{E}[X]\right)^2$. ■

In order to illustrate the relative advantage of Chebyshev's inequality compared to Markov's consider the following example. Let $X_1,\ldots,X_n$ be n independent identically distributed Bernoulli random variables and let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be their average. We would like to bound the probability that $\hat{\mu}_n$ deviates from $\mathbf{E}[\hat{\mu}_n]$ by more than ε (this is the central question in machine learning). We have $\mathbf{E}[\hat{\mu}_n] = \mathbf{E}[X_1] = \mu$ and by independence of X_i-s and Theorem B.26[1] we have $\mathbf{Var}[\hat{\mu}_n] = \frac{1}{n^2}\mathbf{Var}[n\hat{\mu}_n] = \frac{1}{n^2}\sum_{i=1}^n \mathbf{Var}[X_i] = \frac{1}{n}\mathbf{Var}[X_1]$. By Markov's inequality

$$\mathbf{Pr}(\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n]\geqslant\varepsilon) = \mathbf{Pr}(\hat{\mu}_n \geqslant \mathbf{E}[\hat{\mu}_n]+\varepsilon) \leqslant \frac{\mathbf{E}[\hat{\mu}_n]}{\mathbf{E}[\hat{\mu}_n]+\varepsilon} = \frac{\mathbf{E}[X_1]}{\mathbf{E}[X_1]+\varepsilon}\,. \tag{2.9}$$

Note that as n grows the inequality stays the same. By Chebyshev's inequality we have

$$\mathbf{Pr}(\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n]\geqslant\varepsilon) \leqslant \mathbf{Pr}(|\hat{\mu}_n - \mathbf{E}[\hat{\mu}_n]|\geqslant\varepsilon) \leqslant \frac{\mathbf{Var}[\hat{\mu}_n]}{\varepsilon^2} = \frac{\mathbf{Var}[X_1]}{n\varepsilon^2}\,. \tag{2.10}$$

Note that as n grows the right hand side of the inequality decreases at the rate of $\frac{1}{n}$. Thus, in this case Chebyshev's inequality is much tighter than Markov's and it illustrates that as the number of random variables grows the probability that their average significantly deviates from the expectation decreases. In the next section we show that this probability actually decreases at an exponential rate.

[1] We cite Theorem B.26 here.

Theorem (B.26)**.** *If $X_1,\ldots,X_n$ are independent random variables then*

$$\mathbf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{Var}[X_i]\,. \tag{2.6}$$

The proof is based on Theorem B.23 and the result does not necessarily hold when X_i-s are not independent.

Theorem (B.23)**.** *If X and Y are independent random variables, then*

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]\,. \tag{2.7}$$

We emphasize that in contrast with Theorem B.22, this property does not hold in the general case (if X and Y are not independent).

Theorem (B.22 Linearity)**.** *For any pair of random variables X and Y, not necessarily independent,*

$$\mathbf{E}[X+Y] = \mathbf{E}[X]+\mathbf{E}[Y]\,. \tag{2.8}$$