Algorithms for Data Science

Barna Saha

1 Probability Review

Sample space: the set Σ of all possible outcomes of a random process.
Example: Tossing two coins. $\Sigma = \{HH, HT, TH, TT\}$.

Event: a subset of a sample space Σ.
Example: event that both the coins have the same output. $E = (HH, TT)$

Lemma 1 (Union Bound)**.** *For any finite or countably infinite sequence of events E_1, E_2,....,*

$$\Pr[\bigcup_{i\geq 1} E_i] \leq \sum_{i\geq 1} \Pr[E_i]$$

If the events above are pairwise mutually disjoint then

$$\Pr[\bigcup_{i\geq 1} E_i] = \sum_{i\geq 1} \Pr[E_i]$$

Definition. Events E_1, E_2,..., E_k are mutually independent if and only if, for any subset $I \subseteq [1, k]$,

$$\Pr[\bigcap_{i\in I} E_i] = \prod_{i\in I} \Pr[E_i]$$

The conditional probability that event E occurs given that even F has already occurred

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)}$$

If E and F are mutually independent then $\Pr(E \mid F) = \Pr(E)$.

We will be mostly consider discrete probability space, thus the sample space will be finite or countably infinite.

Random Variable.

Definition. A random variable on a sample space Σ is a real-valued function on Σ; that is $X : \Sigma \to \mathbb{R}$. A discrete random variable is a random variable that taken on only a finite or countably infinite number of values.

Let X be a random variable.

Definition (Expectation)**.** For a discrete random variable X, the expectation of X, $\mathbf{E}[X]$ is

$$\mathbf{E}[X] = \sum_a a \Pr[X = a]$$

For a continuous random variable X, the expectation of X, $\mathbf{E}[X]$, is

$$\mathbf{E}[X] = \int a\phi(a)da$$

where ϕ is the probability density function of X.

Lemma 2 (Linearity of Expectation)**.** *Let X and Y be two random variables.* $\boldsymbol{E}[X+Y] = \boldsymbol{E}[X] + \boldsymbol{E}[Y]$.

Lemma 3. *If X and Y are two independent random variables, then* $\boldsymbol{E}[XY] = \boldsymbol{E}[X]\boldsymbol{E}[Y]$

Lemma 4 (Markov's inequality)**.** *Let X be a non-negative random variable. For all $\lambda > 0$,*

$$\Pr[X > \lambda] \leq \frac{\boldsymbol{E}[X]}{\lambda}$$

Definition (Variance)**.** The variance of a random variable X, denoted $\text{var}[X]$, is

$$\text{var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

Lemma 5 (Linearity of Variance)**.** *Let $X_1, X_2, ..., X_n$ be mutually independent random variables. Then*

$$var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} var[X_i]$$

Lemma 6 (Chebyshev Inequality)**.** *For all $\lambda > 0$,*

$$\Pr[|X - \boldsymbol{E}[X]| > \lambda] \leq \frac{var[X]}{\lambda^2}$$

Coin Tossing Example Consider tossing n fair coins, that is each coin has equal probability $\frac{1}{2}$ of returning a head or a tail. Obtain an upper bound on the probability of obtaining more than $\frac{3n}{4}$ heads.

Define the following indicator random variable.

$$X_i = \begin{cases} 1 & \text{if the } i\text{th coin flip is a head} \\ 0 & \text{otherwise} \end{cases}$$

Define

$$Y = \sum_{i=1}^{n} X_i$$

Then Y counts the number of heads in a sequence of n tosses.

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \frac{n}{2}.$$

Then, applying Markov's Inequality

$$\Pr[Y \geq \frac{3n}{4}] \leq \frac{n/2}{3n/4} = \frac{2}{3}.$$

Let us now compute the variance of Y. Since X_is are all independent random variables, we have

$$\text{var}[Y] = \text{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} var[X_i] = \sum_{i=1}^{n} E[X_i^2] - E[X_i]^2 = \frac{n}{4}.$$

Then, by applying Chebyshev's inequality,

$$\Pr[Y \geq \frac{3n}{4}] \leq \Pr[|Y - \frac{n}{2}| \geq \frac{n}{4}] \leq \frac{n/4}{n^2/16} = \frac{4}{n}.$$

Lemma 7 (The Chernoff Bound: Upper bound)**.** *Let $X_1, X_2, ..., X_n$ be independent random variables taking values in $\{0, 1\}$ with $\mathbb{E}[X_i] = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then the following holds*

1. *For any $\delta > 0$,*
$$\Pr[X \geq (1+\delta)\mu] < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \right)^{\mu}$$
2. *For $0 < \delta \leq 1$,*
$$\Pr[X \geq (1+\delta)\mu] \leq e^{-\frac{\mu\delta^2}{3}}$$

Lemma 8 (The Chernoff Bound: Lower bound)**.** *Let $X_1, X_2, ..., X_n$ be independent random variables taking values in $\{0, 1\}$ with $\mathbb{E}[X_i] = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then the following holds*

1. *For any $\delta > 0$,*
$$\Pr[X \leq (1-\delta)\mu] < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \right)^{\mu}$$
2. *For $0 < \delta \leq 1$,*
$$\Pr[X \leq (1-\delta)\mu] \leq e^{-\frac{\mu\delta^2}{2}}$$

Exercises.

Exercise 1. *We flip a fair coin ten times. Find the probability of the following events.*

(a) The number of heads and the number of tails are equal.

(b) There are more heads than tails.

(c) The ith flip and the $(11-i)$th flip are the same for $i = 1, ..., 5$.

(d) We flip at least four consecutive heads.

Exercise 2. *Suppose we roll a fair k-sided die with the numbers 1 through k on the die's faces. If X is the number that appears, what is $\mathbb{E}[X]$?*

Exercise 3. *A monkey types on a 26-letter keyboard that has lowercase letters only. Each letter is chosen independently and uniformly at random from the alphabet. If the monkey types $1,000,000$ letters, what is the expected number of times the sequence "proof" appears?*

Exercise 4. *Suppose you play a simple game with your friend where you flip a coin. If the coin is heads, your friend pays you a dollar. If it's tails, you pay your friend a dollar.*

(a) Suppose you play the game 100 times, what is your expected pay off?

(b) Suppose your friend decides to trick you, and swaps the fair coin for a biased coin that comes up tails with probability 0.7. What is your expected pay off if you play 100 times?

(c) Use Markov Inequality to give an upper bound on the probability that your friend gets more than 50 after 100 rounds.

Exercise 5. *Let X be a number chosen uniformly at random from $[1, n]$. Find var$[X]$.*

Exercise 6. *Suppose that we roll a standard fair die 100 times, Let X be the random variable denoting the sum of numbers that appear over the 100 rolls. Use Chebyshev's inequality to bound $\Pr[|X - 350| \geq 50]$.*

Exercise 7. *Suppose you throw m balls into n bins, each ball equally likely to go into any of the n bins; imagine $m \geq n$. Let random variable B_i denote the number of balls in bin i. What is $E[B_i]$?*

- *Suppose $m = 100n \ln n$. Use the Chernoff bound to show that the number of balls in bin i does not differ from the expectation by more than (say) $25 \ln n$ with probability at least $1 - \frac{1}{n^2}$. Hence, show that the load of the heaviest and lightest bins differ by at most a constant factor with probability at least $1 - \frac{1}{n}$.*
- *For general $m = \Omega(n \ln n)$, show that the number of balls in all the bins lie in the range $\frac{m}{n} \pm O(\sqrt{\frac{m}{n} \ln n})$ with probability at least $1 - \frac{1}{n}$.*
- *Now suppose $m = n$. Show that the height of the heaviest bin is $O(\frac{\ln n}{\ln \ln n})$ with probability $1 - o(1)$.*