CS 70: Homework #11

Abhijay Bhatnagar

November 9, 2018

1 Random Cuckoo Hashing **2**

2 Markov's Inequality and Chebyshev's Inequality **3**

3 Easy A's **4**

4 Confidence Interval Introduction **5**

1 Random Cuckoo Hashing

Cuckoo birds are parasitic beasts. They are known for hijacking the nests of other bird species and evicting the eggs already inside. Cuckoo hashing is inspired by this behavior. In cuckoo hashing, when we get a collision, the element that was already there gets evicted and rehashed.

We study a simple (but ineffective, as we'll see) version of cuckoo hashing, where all hashes are random. Let's say we want to hash n pieces of data $D_1, D_2, \ldots, D_n$ into n possible hash buckets labeled $1, \ldots, n$. We hash the $D_1, \ldots, D_n$ in that order. When hashing D_i, we assign it a random bucket chosen uniformly from $1, \ldots, n$. If there is no collision, then we place D_i into that bucket. If there is a collision with some other D_j, we evict D_j and assign it another random bucket uniformly from $1, \ldots, n$. (It is possible that D_j gets assigned back to the bucket it was just evicted from!) We again perform the eviction step if we get another collision. We keep doing this until there is no more collision, and we then introduce the next piece of data, D_{i+1} to the hash table.

a.) What is the probability that there are no collisions over the entire process of hashing $D_1, \ldots, D_n$ to buckets $1, \ldots, n$? What value does the probability tend towards as n grows very large?

Solution: Let A_n = event that there are no collisions in N buckets.
$\mathbb{P}[A_n] = \frac{(n)\times(n-1)\times\cdots\times(1)}{n^n} = \frac{n!}{n^n}$

$\lim_{n\to\infty} \mathbb{P}[A_n] = 0$ (the factorial is term by term less than n^n for every term $< n$).

b.) Assume we have already hashed $D_1, \ldots, D_{n-1}$, and they each occupy their own bucket. We now introduce D_n into our hash table. What is the expected number of collisions that we'll see while hashing D_n? (*Hint*: What happens when we hash D_n and get a collision, so we evict some other D_i and have to hash D_i? Are we at a situation that we've seen before?)

Solution: Let X be the number of collisions.

One method is an infinite geometric series: $\sum_{i=0}^{\infty} i \times (\frac{n-1}{n})^i$, but I don't know how to solve that so let's skip that.

The second method is to recognize that if it hits any of the $n-1$ occupied bins, it's the exact same problem after displacement:

Therefore, $\mathbb{E}[X] = 1 + \mathbb{P}[\text{collision}]\mathbb{E}[X] = 1 + \frac{n-1}{n}\mathbb{E}[X] \implies \mathbb{E}[X](1 - \frac{n-1}{n}) = 1 \implies \mathbb{E}[X] = n$

2 Markov's Inequality and Chebyshev's Inequality

A random variable X has variance $\text{Var}(X) = 9$ and expectation $\mathbb{E}[X] = 2$. Furthermore, the value of X is never greater than 10. Given this information, provide either a proof or a counterexample for the following statements.

a.) $\mathbb{E}[X^2] = 13$.

Solution: True.

Proof. $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \implies \mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2 = 9 + 2^2 = 13$.

□

b.) $\mathbb{P}[X \leq 1] \leq 8/9$.

Solution: True.

Proof. Let $Y = 10 - X$. Y is guaranteed to be nonnegative since $X \leq 10$. Therefore we can use Markov's Inequality on Y. $\mathbb{P}[X \leq 1]$ is equivalent to $\mathbb{P}[Y \geq (10-1)] = \mathbb{P}[Y \geq 9] = \frac{\mathbb{E}[Y]}{9}$. We can find $\mathbb{E}[Y] = \mathbb{E}[10 - X] = 10 - \mathbb{E}[X] = 8$. Plugging this back in gives $\mathbb{P}[X \leq 1] = \mathbb{P}[Y \geq 9] = \frac{8}{9}$, as desired. □

c.) $\mathbb{P}[X \geq 6] \leq 9/16$.

Solution: True.

Proof. Let $Y = (X - \mu)^2$. We know $\mathbb{E}[Y] = \text{Var}(X)$ and $\mathbb{P}[X \geq 6] = \mathbb{P}[X - \mu \geq 6 - \mu] = \mathbb{P}[Y \geq 4] = \frac{\mathbb{E}[Y]}{4^2} = \frac{\text{Var}(X)}{16} = \frac{9}{16}$. □

d.) $\mathbb{P}[X \geq 6] \leq 9/32$.

3 Easy A's

A friend tells you about a course called "Laziness in Modern Society" that requires almost no work. You hope to take this course next semester to give yourself a well-deserved break after working hard in CS 70. At the first lecture, the professor announces that grades will depend only on two homework assignments. Homework 1 will consist of three questions, each worth 10 points, and Homework 2 will consist of four questions, also each worth 10 points. He will give an A to any student who gets at least 60 of the possible 70 points.

However, speaking with the professor in office hours you hear some very disturbing news. He tells you that, in the spirit of the class, the GSIs are very lazy, and to save time the grading will be done as follows. For each student's Homework 1, the GSIs will choose an integer randomly from a distribution with mean $\mu = 5$ and variance $\sigma^2 = 1$. They'll mark each of the three questions with that score. To grade Homework 2, they'll again choose a random number from the same distribution, independently of the first number, and will mark all four questions with that score.

If you take the class, what will the mean and variance of your total class score be? Use Chebyshev's inequality to conclude that you have less than a 5% chance of getting an A when the grades are randomly chosen this way.

Solution: Let T = total score of all problems, and S_i = the score from the ith homework assignment. Each problem has the same distributions, i.e. $\mathbb{E}[S_i] = \mu = 5$, and $\text{Var}(S_i) = 1$.

Start by recognizing the score from each homework problem is independent from each other. This allows us to simplify some of the combined values:

$$\mathbb{E}[T] = \mathbb{E}[S_1 + ... + S_7] = \sum_{i=1}^{7} \mathbb{E}[S_i] = 5 \times 7 = 35$$

$$\text{Var}(T) = \text{Var}(S_1 + ... + S_7) = \sum \text{Var}(S_i) = 7$$

From these, we can proceed to find the probability that you get an A, i.e. $\mathbb{P}[T \geq 60]$. Let $Y = (T - \mathbb{E}[T])^2$, and we proceed with the stronger Chebyshev's Inequality:

$$\begin{aligned}
\mathbb{P}[T \geq 60] &= \mathbb{P}[T - \mu \geq 60 - \mu] \\
&= \mathbb{P}[T - \mu \geq 60 - 35] \\
&= \mathbb{P}[Y \geq 25^2] \\
&\leq \frac{\text{Var}(T)}{25^2} \\
&\leq \frac{7}{25^2} \\
&< 1.2\% \\
&< 5\%
\end{aligned}$$

4 Confidence Interval Introduction

We observe a random variable X which has mean μ and standard deviation $\sigma \in (0, \infty)$. Assume that the mean μ is unknown, but σ is known.

We would like to give a 95% confidence interval for the unknown mean μ. In other words, we want to give a random interval (a, b) (it is random because it depends on the random observation X) such that the probability that μ lies in (a, b) is at least 95%.

We will use a confidence interval of the form $(X - \varepsilon, X + \varepsilon)$, where $\varepsilon > 0$ is the width of the confidence interval. When ε is smaller, it means that the confidence interval is narrower, i.e., we are giving a more *precise* estimate of μ.

a.) Using Chebyshev's Inequality, calculate an upper bound on $\Pr\{|X - \mu| \geq \varepsilon\}$.

Solution: Upper bound $= \frac{\text{Var}(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$

b.) Explain why $\Pr\{|X - \mu| < \varepsilon\}$ is the same as $\Pr\{\mu \in (X - \varepsilon, X + \varepsilon)\}$.

Solution: This is just the definition of absolute value. Consider the two cases:

(a) $(X - \mu)$ is positive, then $X - \mu < \varepsilon \implies X < \varepsilon + \mu \implies \mu > X - \varepsilon$

(b) $(X - \mu)$ is negative, then $X - \mu > -\varepsilon \implies X > -\varepsilon + \mu \implies \mu < X + \varepsilon$

c.) Using the previous two parts, choose the width of the confidence interval ε to be large enough so that $\Pr\{\mu \in (X - \varepsilon, X + \varepsilon)\}$ is guaranteed to exceed 95%.

[Note: Your confidence interval is allowed to depend on X, which is observed, and σ, which is known. Your confidence interval is not allowed to depend on μ, which is unknown.]

Solution: $1 - \frac{\sigma^2}{\varepsilon^2} > 0.95 \implies 1 - 0.95 > \frac{\sigma^2}{\varepsilon^2} \implies \varepsilon > \frac{\sigma}{\sqrt{1-0.95}}$

d.) The previous three parts dealt with the case when you observe one sample X. Now, let n be a positive integer and let $X_1, \ldots, X_n$ be i.i.d. samples, each with mean μ and standard deviation $\sigma \in (0, \infty)$. As before, assume that μ is unknown but σ is known.

Here, a good estimator for μ is the *sample mean* $\bar{X} := n^{-1} \sum_{i=1}^{n} X_i$. Calculate the mean and variance of $\bar{X}$.

Solution:

$$\mathbb{E}[\bar{X}] = \mathbb{E}[n^{-1} \sum_{i=1}^{n} X_i] = n^{-1} \mathbb{E}[\sum_{i=1}^{n} X_i] = n^{-1} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mu$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(n^{-1} \sum_{i=1}^{n} X_i) = n^{-2} \mathrm{Var}(\sum_{i=1}^{n} X_i) = n^{-2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = n^{-2}(n \times \sigma^2) = \frac{\sigma^2}{n}$$

e.) We will now use a confidence interval of the form $(\bar{X} - \varepsilon, \bar{X} + \varepsilon)$ where $\varepsilon > 0$ again represents the width of the confidence interval. Imitate the steps of (a) through (c) to choose the width ε to be large enough so that $\Pr\{\mu \in (\bar{X} - \varepsilon, \bar{X} + \varepsilon)\}$ is guaranteed to exceed 95%.

To check your answer, your confidence interval should be *smaller* when n is larger. Intuitively, if you collect more samples, then you should be able to give a more *precise* estimate of μ.

Solution: Upper bound $= \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$.

$\Pr\{\mu \in (\bar{X} - \varepsilon, \bar{X} + \varepsilon)\} = 1 - \frac{\sigma^2}{n\varepsilon^2} > 0.95 \implies \varepsilon > \frac{\sigma}{\sqrt{n(1-0.95)}}$