Probability

Sam Polgar

August 7, 2024

1. week 1

Definition 1.1. Expected Value

Outcomes of an experiment or random process are real numbers $a_1, \dots a_n$ with probabilities $p_1, \dots, p_n$

$$\sum_{k=1}^{n} a_k p_k = a_1 p_1 + \cdots + a_n p_n$$

Example 1.2. Lottery example: 500,000 people pay \$5 with winners 1 x \$1,000,000, 10 x \$1,000, 1000 x \$500, 10,000 x \$10. What's the expected value of the ticket?
$p_k = \frac{1}{500{,}000}$ for each $k = 1, \dots, 500000$ p_k is the probability of each outcome occuring. a_i is the net gain for a ticket a_i where $a_1 = 999,995$ the net gain for winning minus the \$5 cost. 2nd prize $a_2, \dots, a_{11} = 995$, 3rd = $a_{12}, \dots, a_{1011} = 495$ 4th = $a_{1011}, \dots, a_{1012} = 10$, 5th = $a_{1012}, \dots, a_{11011} = 10$, remainder $a_{11011}, \dots, a_{500000} = -5$ Expected value of a ticket is

$$\sum_{k=1}^{500000} a_k p_k. \; Given \; p_k = \frac{1}{500000}, then \frac{1}{500000} \cdot \sum_{k=1}^{500000} a_k$$

$$\frac{1}{500000}(999,995 + 10 \cdot 995 + 100 \cdot 495 + 10000 \cdot 5 + (-5) \cdot 488989) = -1.78$$

A person who plays this lottery will on average lose 1.78 per ticket!

Example 1.3. How many consecutive pairs of the same suit are expected in a deck of cards?

1. Define Indicator variables: Deck of cards = 52, 13 each suit. Let $X_i = 1$ if i and $i+1$ are same suit, 0 otherwise.
2. Calculate $P(X_i = 1)$. First $P = \frac{1}{52}$, Second card matches first = $\frac{12}{51}$ for each i.
3. Linearity of expectation: Total no. pairs

Experiment Shuffle a deck of cards, go through in order. How many times do 2 consecutive cards have the same suit?

1.1. Linearity of expectation

...The sum of each little thing

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

No assumption of independence or anything. Surprisingly useful.

Prove Expectation of selecting a card of type 1 1/13

sum of all expecations Xi depends on 2 cards. What's Pr

Probability of 52 C 13 E[Xi] = Pr[Xi = 1]

Note: difference between Expectation and Probability. Probability = the likelihood of the event e.g. selecting 2 consecutive suit cards from a deck of 52 Expectation = the average outcome. Multiply each outcome by it's probability.

need to do th

discuss differe

Monte Carlo, Las Vegas Running time Output quality

Question I have an array with n = 100 index of an even number what is time complexity of getting even numbers Theta n

- why isn't this constant? because you can create an algorithm that only selects

"on expectation", the las Vegas

expected time For all expectation [Ta] = sum from infiintiy i = 1, i Pr [takes i attempts to find an even number]

$$\mathbb{E}[T_a] = \sum_{i=1}^{\infty} i \cdot \Pr[\text{takes } i \text{ attempts to find an even number}]$$

$$= \sum_{i=1}^{\infty} \frac{i}{2^i} = O(1)$$

Why Randomization?

Faster, Simpler Algo'S - miller rabin, it's a monte carlo algo. Runs in $Otilden^2$

Algos Quicksort Expected running time Is it Las Vegas or Monte Carlo? It's always going to return the sorted array, so it's Las Vegas. Proof: T(n) = Expectations[runtime on array size n] T(n) = E[T(|A1|)] + E[T(|A2|)] + O(n) We know |A1| + |A2| = n-1 |Why n-1?

Expected time analysis vs worst case. expected time analysis: for randomized algorithms average time analysis: using input from a known probability distribution amortized analysis: reusing algorithm on a sequence of inputs, and look at the worst-case sequence of input for the algorithm divided by the length of the sequence

Tutorial 1

Expectation, Discrete Random Variable

Change to definitions here: https://proofwiki.org/wiki/Definition:Expectation

- $E[X]$ = expectation of random variable X
- X is a discrete random variable having probability mass function $p(x)$, then $E[X]$ is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

- Probability Mass Function (PMF) $p(x)$ gives Pr that a discrete random variable X is equal to some value x
- PMF: $p(x) \geqq 0$ for all x, $\sum p(x) = 1$ the sum of probabilities over all possible values

Expectation, Continuous Random Variable

- X is a continuous random variable having probability mass function $f(x)$, then $E[X]$ is defined by

$$E[X] = \int_{x:p(x)>0} xf(x)d(x)$$

- PDF = probability density function

Variance

- Def: average of the squared differences from the mean

$$X : Var(X) = \mathbb{E}[(X - \mathbb{E}[X]^2)] \text{ By Linearity } = \mathbb{E}[X^2] - (E[X])^2$$

- $Var(aX + b) = a^2 Var(X)$ for constants a and b
- For independent random variables: $Var(X + Y) = Var(X) + Var(Y)$
- Indicates how much a data point deviates from the mean, low variance = close to mean, high variance = far from mean, wider range
- Example: population $\sigma^2 = \sum(X - \mu)^2/N$. Sample: $s^2 = \sum \frac{(X-X)^2}{n-1}$.
- σ^2/s^2 is the variance, X is each value in a data set, $\mu or X$ is the mean, $Norn$ is number of data points
- Expressed as squared units of the original data, sometimes confusing
- Standard deviation is the square root of variance, in the same units of the original data
-

Complement Rule

For event A, $\Pr(A) + \Pr(\text{not } A) = 1 \equiv \Pr(\text{not } A) = 1 - P(A)$

Expectation of Binomial Distribution

X is a discrete random variable with binomial distribution parameters n, p for $n \in \mathbb{N}$ and $0 \leq p \leq 1$ then $\mathbb{E}[X] = np$ Proof: $https://proofwiki.org/wiki/Expectation_of_Binomial_Distribution$

Tutorial 1

Problem 1 Consider a deck of $4n$ cards with 'S', 'H', 'D', 'C', after shuffled randomly, what's the expected number of consecutive pairs of the same suit.

1. Define Indicator Variable X_i for each iteration required. We have $4n - 1$ because $4n$ cards, minus 1 match because the last card can't be matched with the null pointer next door. Let S = shuffled card

$$X_i = \begin{cases} 1 & \text{if } S_i = S_{i+1} \\ 0 & \text{else} \end{cases}$$

2. Sum the number of consecutive pairs found, i.e., sum all 1 cases

$$X = \sum_{i=1}^{4n-1} X_i \text{ we need to find } E[X] = E[\sum_{i=1}^{4n-1} X_i]$$

By linearity

$$= \sum_{i=1}^{4n-1} E[X_i] \; = \; \sum_{i=1}^{4n-1} \frac{\text{no. cards in a suit}}{\text{no. cards in a deck}}$$

We need to find $E[X_i]$ for each i = Pr that $S_i = S_{i+1}$:

$$\Pr(X_i = 1) = \frac{n-1}{4n-1} \; = E[X_i] = \frac{n-1}{4n-1}$$

Problem 2 Similar
Problem 3 Similar Variance part explained here https://claude.ai/chat/9210ca3a-e032-4137-80b1-455acfe9835a
Problem 4 explained here https://claude.ai/chat/fc539a53-9b45-4e59-9638-fcd3e8532612
Problem 5

Quiz 0

Q1: Handshaking Lemma states if $G = (V, E)$ is an undirected graph, $\sum_{v \in V} degv$ is equal to: $2|E|$.
Recall E0–v–0E

Q2: Linearity of expectation means if X, Y are 2 arbitrary random variables and a, b are 2 arbitrary random numbers then $\mathbb{E}[aX + bY] = \mathbb{E}[aX] + \mathbb{E}[bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ if:

1. as long as both expectations are defined
2. only if X, Y are independent
3. only if a, b are positive
4. only if X, Y are uncorrelated

Quiz 1

1. if X is the indicator variable of event E, then $\mathbb{E}[X]$ is...
   $\Pr[E]$ because by def. $\mathbb{E}[X]$ is the sum of the probability of each event happening times its outcome. An indicator variable is related to 1 single event.
2. The $\Omega(n \log n)$ worst-case lower bound for comparison based sorting algorithms doesn't apply to randomised algorithms.
   False. Why?
3. What type of randomized algorithm is randomized QuickSort?
   Las Vegas, will produce the correct result with randomized running time. Monte Carlo is 0.99 percent correct with fixed running time.
   Randomized QuickSort fits Las Vegas because it optimizes for correctness not speed, running time varies. QuickSort needs to arrange elements in order and needs correctness (the ordering can't be wrong)
   Efficiency = expected O(n log n), worst case O^2
4. if a random variable has well-defined variance and expectation, then
   $Var[X] \leq \mathbb{E}[X^2]$ as per definition of X with finite variance Question: What's the difference between $\mathbb{E}[X^2] and \mathbb{E}[X]^2$ - something to do with constants? X is a random variable not a constant. Exp[X] is a constant.
5. The expected number of comparisons by Randomized QuickSort is at most
   $\mathcal{O}(n \log n)$ - why?
6. For a given input x a randomied algo A always gives different answers when given different random strings
   False: (I guessed True). Why?

   What is the purpose of randomization!? Many cases it's to improve average case performance or avoid worst-case performance, not to produce varying output. Often there is consistent output for a given

input. Randomness is used not necessarily as encryption but to enable the algorithm to find the solution faster. E.g. Randomized Quicksort: the final sorted array is the same but choice of pivots is randomized

7. We have a much faster polynomial-time randomised algorithm than the best known deterministic algorithm for:
   integer primality testing: Miller Rabin or the other one Clément mentioned that's n^6

8. Suppose I'm given a randommized algorithm A which takes integer n as input and outputs a random not necessarily uniform English sentence of exactly n words. If I run the algorithm twice with input n with the same random bits, will I get the same sentence?
   Yes because the source of randomness is the same.

   randomString = "10110101" fn A(n) -> sentence

9. What is a possible limitation to keep in mind when using randomised algorithms? They assume access to good randomness (not they're harder to analyse, my answer)

10. If the expected running time of an algorithm of input size n is $\mathcal{O}(1)$ then worst-case running time is always $\mathcal{O}(1)$
    False! it would be the proper worse case

2. week 2

Summary Concentration Inequalities, Markov first moment method, Chebyshev second moment method, Chernoff/Hoefding Union bound: with linearity of expectation Probabilty amplification: majority vote, median trick Sampling

Markov Inequality Proof Application from Las Vegas to Monte Carlo Chebyshev + reduce to Markov + Application Chernoff/Hoeffding bounds Concentration tools

Random [0,1], (0,1] notation = inclusive/uninclusive

Assignment Problem Handchecking Lemma 52 minutes into the lecture

Concentration Inequalities How likely a random variable X is to deviate from its expected value (mean). concentration = how tight are the values around the mean? Inequalities mean they provide upper bounds on probabilities, not exact values.

What they do: provide useful bounds on the probability of events occurring. Markov: the probability of random variable X is no more than ... Chebyshev: the probability of deviating from its mean by more than some amount is at most ...

When is it useful? Useful information about a random variable used when we know the mean and variance but not the PMF/PDF e.g. you take some sample statistics from the distribution Knowing mean = knowing $\mathbb{E}[X]$ because they imply each other Knowing variance = knowing the spread of values around the mean, that is, Significance of knowing PMF/PDF: this is a function that describes the distribution e.g. Normal (Gaussian), Exponential, Poisson

Used in the analysis of randomized algorithms ‘’

If you want the math to appear in its own line, the standard way is to use:

```
\[
    \sqrt{x+y}
\]
```