8

Intro to Probability, Kolmogorov Complexity

Computer Science Mentors 70

Mar 29 - Apr 2
Spring 2021

Amogh Gupta, Sylvia Jin, Aekus Bhathal, Abinav Routhu, Debayan Bandyopadhyay
Roast us here: `https://tinyurl.com/csm70-feedback20`

Introduction to Kolmogorov Complexity

1. Recall from HW 7 that the *Kolmogorov complexity* of a string $K(x)$ is the length of an optimally-compressed copy of x; that is, $K(x)$ is the length of the shortest program that returns x. Let's explore this definition a little bit.

   (a) True or false? The Kolmogorov complexity of any string is always smaller than the length of the string itself.

   > **Solution:** False. By HW 7 Q8 part (b), there is always at least one string of n bits that is incompressible, so there is at least 1 string of length n whose Kolmogorov complexity is at least n.

   (b) True or false? When you use a program to "zip" files on your computer into a compressed archive, the resulting .zip file has a smaller Kolmogorov complexity than the original files.

   > **Solution:** False, they have the same Kolmogorov complexity. The zip file is (usually) just better-compressed.

   (c) In plain English, write the shortest possible representation of the following strings. There are multiple possible solutions because of the constraints of the English language.

      (a) `cs70cs70cs70cs70cs70cs70cs70`

      > **Solution:** "cs70 7 times", "cs70 7x", or similar.

      (b) `3.14159265358979323846264...`

      > **Solution:** "The ratio of a circle's circumference to its diameter". Just "pi" may also work but is not preferred, since "pi" can also be used as a variable or have other quantities associated with it. Note that although the length of the string is infinite, its Kolmogorov complexity is very small!

      (c) `We're no strangers to love / You know the rules and so do I / A full commitment's what I'm thinking of / You wouldn't get this from any other guy / I just wanna tell you how I'm feeling / Gotta make you understand / Never gonna give you up / Never gonna let you down / Never gonna run around and desert you ...`

      > **Solution:** "the lyrics to Rick Astley's song *Never Gonna Give You Up*", "rickroll song", or similar. This demonstrates that strings with some degree of repetition usually have a lower Kolmogorov complexity than a random string of the same length.

Introduction to Probability

1. Suppose two integers a and b are drawn uniformly from $[-n \,..\, n]$, that is $a, b \in \mathbb{Z}$ and $-n \leq a, b \leq n$.

   (a) Define a probability space for (a, b). Does each sample point occur with uniform probability?

Solution: A probability space is defined by a sample space and probabilities for each sample point. The sample space is defined as

$$\Omega = \{(i, j) : -n \leq i, j \leq n\}$$

Clearly $\mathbb{P}[(i, j)] = \mathbb{P}[(k, l)]$ by uniform symmetry, so $\forall \omega \in \Omega$,

$$\mathbb{P}[\omega] = \frac{1}{|\Omega|} = \frac{1}{(2n+1)^2} \tag{1}$$

(b) Find the probability that $\max\{0, a\} = \min\{0, b\}$.

Solution: The problem is true if and only if $a \leq 0$ and $b \geq 0$. There are $n+1$ such values for a and $n+1$ such values for b, meaning there are $(n+1)^2$ sample points. Given each sample point occurs with equal probability, we simply divide the number of sample points by the cardinality of the sample space:

$$\frac{(n+1)^2}{(2n+1)^2}$$

(c) Find the probability that $|a - b| \leq k$. You may assume $k < \frac{n}{2}$.

Solution: We begin by finding a formula for arbitrary n.
If $a < -n + k$, then $a = -n + i$ where $i \in [0 \mathrel{..} k-1]$ then, b can be in $[-n \mathrel{..} -n+k+i]$, meaning there are $k + i + 1$ satisfying sample points for b.

If $-n + k \leq a \leq n - k$, then b can be in $[a - k \mathrel{..} a + k]$, meaning there are $2k + 1$ satisfying sample points for b.

If $a > n - k$, the case is symmetric to the first case and there are again $k + i + 1$ satisfying sample points for b.

Now, we find the number of sample points for (a, b) by summing over all possible values for a:

$$\begin{align}
|A| &= 2\Big(\sum_{i=0}^{k-1} k + i + 1\Big) + \Big(\sum_{i=1}^{(n-k)-(-n+k)+1} 2k + 1\Big) \tag{2}\\
&= 2\big(k^2 + \frac{k(k-1)}{2} + k\big) + (2n - 2k + 1)(2k + 1) \tag{3}\\
&= 2k(k+1) + k(k-1) + (2n - 2k + 1)(2k + 1) \tag{4}\\
&= -k^2 + k + 4nk + 2n + 1 \tag{5}
\end{align}$$

From (1) to (2), we use the arithmetic sum formula. Since each sample point occurs with equal probability,

$$\mathbb{P}(|A|) = \frac{-k^2 + k + 4nk + 2n + 1}{(2n+1)^2} \tag{6}$$

(d) Suppose we choose two closed intervals $u = [a \mathrel{..} b], v = [c \mathrel{..} d]$ uniformly at random from $[-n \mathrel{..} n]$. What is the probability that u is enveloped by v, meaning that $u \subset v$ and $c < a \leq b < d$. What happens to this probability as n approaches ∞?

Solution: First, we define our sample space as the set of all pairs of intervals, so that the probability of any sample point is equal.

Notice that any interval is uniquely defined by its starting and ending index. Hence, we place bars between elements, before the first element, and after the last element (e.g. $| -n| -n+1| \dots |n-1|n|$) and the number of intervals is the number of ways to choose 2 bars, which is the same as choosing 2 distinct elements from $2n+2$.

$$\binom{2n+2}{2} \tag{7}$$

Hence, the size of the sample space is $\binom{2n+2}{2}^2$.

Now, there is always exactly one arrangement of (a, b, c, d), (e, f, g, h) such that $e < f \leq g < h$. Once again, this problem can be modeled as selecting 4 bars from the bars between elements, giving us

$$|A| = \binom{2n+2}{4} \tag{8}$$

Dividing $|A|$ by $|\Omega|$ gives

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\binom{2n+2}{4}}{\binom{2n+2}{2}^2} = \frac{1}{6}\frac{2n!^2}{(2n-2)!(2n+2)!} = \frac{1}{6}\frac{2n(2n-1)}{(2n+2)(2n+1)} \tag{9}$$

As $n \longrightarrow \infty$, the probability converges to $\frac{1}{6}$

2. Alex and Shruti are playing Yahtzee, a game involving rolling 5 dice.

(a) First, define a probability space representing the possible outcome of Alex or Shruti's rolls of the 5 dice. Assume all dice are fair and labeled 1 through 6.

Solution: Our probability space can be represented by

$$\Omega = \{(d_1, d_2, d_3, d_4, d_5) \mid d_i \in \{1, 2, 3, 4, 5, 6\}\}$$

or similar, since there are 5 rolls and each one takes on a value of 1 through 6.

Alex and Shruti each roll 1 die to see who goes first. The person with the higher roll goes first, and in case of a tie, they both roll their die again.

(b) What's the chance Shruti rolls a higher number on the first roll?

Solution: Let p be the chance Shruti rolls a higher number. By symmetry, Alex has the same chance p of rolling the higher number. Therefore the chance of a tie (neither Alex nor Shruti wins) is $1-2p$. A tie happens when they both roll 1, 2, 3, 4, 5, or 6, which happens with probability $6/36 = 1/6$. Thus $1-2p = 1/6$. Solving for p, we get $p = 5/12$.

(c) What's the chance Shruti goes first?

Solution: Solution 1: By symmetry, it's $1/2$ since Shruti's chance of going first, even if rerolls are needed, is always the same as Alex's.
Solution 2: By the law of total probability, we sum up Shruti's chance of going first after $n = 1, 2, 3, \ldots$ rolls:

$$\begin{aligned}
P(1) &= \frac{5}{12} \\
P(2) &= \frac{1}{6} \cdot \frac{5}{12} \\
&\vdots \\
P(\text{Shruti wins}) &= \sum_{i=1}^{\infty} \frac{5}{12} \cdot \left(\frac{1}{6}\right)^{i-1} \\
&= \frac{5}{12} \cdot \frac{1}{1 - 1/6} \\
&= \frac{5}{12} \cdot \frac{6}{5} = \frac{1}{2}
\end{aligned}$$

(d) They finally begin playing. Partway through the game, Alex is missing the "three of a kind" category while Shruti is missing the "four of a kind" category. What is the probability of rolling...

1. exactly 3 of a kind?

Solution:

$$\frac{\binom{5}{3} \cdot 6 \cdot 5^2}{6^5}$$

2. exactly 4 of a kind?

Solution:

$$\frac{\binom{5}{4} \cdot 6 \cdot 5}{6^5}$$

3. Which one is more likely? 3 of a kind or 4 of a kind?

Solution: We can just compare the numerators of the 2 above parts, since the denominators are equal.

$$\binom{5}{3} \cdot 6 \cdot 5^2 = 10 \cdot 6 \cdot 5^2 = 1500$$

$$\binom{5}{4} \cdot 6 \cdot 6 = 5 \cdot 6^2 \cdot 5 = 900$$

Lining up with intuition, 3 of a kind is more likely.

Inclusion-Exclusion Principle, Bayes' Theorem

1. Tri-State Area is experiencing bad weather because of Doctor Doofenshmirtz "Gloomy - inator". It is always at least rainy, cloudy or windy, but because the inator is random we don't exactly know what it would be like. It rains with 0.5 probability, gets windy with 0.65 probability and gets cloudy with 0.45 probability. We experience at least 2 of these together with probability 0.45.

Help Agent P find the probability that all 3 of these happen together.

Solution: Let R be the event that it is Rainy and $P(R) = 0.5$
W be the event that it is Windy and $P(W) = 0.65$
C be the event that it is Cloudy and $P(C) = 0.45$.
Then **"always atleast"** one of these tells us $P(R \cup W \cup C) = 1$
By the Principle of Inclusion and Exclusion we have

$$P(R \cup W \cup C) = P(R) + P(W) + P(C) - P(R \cap W) - P(W \cap C) - P(C \cap R) + P(R \cap W \cap C)$$

We could write **"atleast two of these together"** as

$$\begin{aligned} P(R \cap W) + P(W \cap C) + P(C \cap R) - 2\dot{P}(R \cap W \cap C) &= 0.45 \\ P(R \cap W) + P(W \cap C) + P(C \cap R) &= 0.45 + 2\dot{P}(R \cap W \cap C) \end{aligned}$$

Notice the 2 comes from the fact that the probability of the three happening at the same time is being counted thrice when we sum pairwise intersection of events. Then substituting everything we get:

$$\begin{aligned} 0.5 + 0.65 + 0.45 - (0.45 + 2\dot{P}(R \cap W \cap C)) + P(R \cap W \cap C) &= 1 \\ 1.15 - P(R \cap W \cap C) &= 1 \\ P(R \cap W \cap C) &= 0.15 \end{aligned}$$

2. You own a pizzeria. You observe that one of your customers, Andy, buys a cheese pizza on Saturday with probability 0.3 and on Sunday with probability 0.6.

   (a) If Andy's pizza purchasing habits on Sunday is independent from his pizza purchasing habits on Saturday, what is the probability that he buys pizza on a given weekend?

   > **Solution:** We use the Inclusion-Exclusion principle. If event A is the situation Andy buys a pizza on Saturday, and event B is the situation in which Andy buys a pizza on Sunday, then the probability that Andy buys pizza a given weekend is $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.6 - 0.3 * 0.6 = 0.72$.

   (b) If Andy buying a pizza on Saturday means that he will not buy pizza on Sunday, what is the probability that he buys pizza on a given weekend (i.e if he buys pizza on one day, he is guaranteed to not buy a pizza the next day)? Note that the probability that he buys a pizza on Sunday, 0.6, is *not* a conditional probability, i.e. it is not conditioned on whether he buys a pizza on Saturday.

   > **Solution:** Now we know that $P(A \cap B) = 0$. Use the Inclusion-Exclusion principle: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.6 - 0 = 0.9$.

   (c) Suppose we don't know how Andy's pizza purchasing habits on Sundays depends on whether he bought a pizza on the preceding Saturday. Given that Andy buys pizza on a given weekend with probability 0.65, what is the probability that he buys pizza both days?

   > **Solution:** By the Inclusion-Exclusion principle, we know that $P(A) + P(B) - P(A \cap B)$. Solving, we find that $P(A \cap B) = 0.25$.

3. (*Note: before students attempt this problem, it is highly recommended that they attempt problem 1d*)

   Suppose Ezekiel places m rectangles on an nxn grid, each chosen uniformly at random and independent from all others. Lower bound the probability that no rectangle is enveloped by another rectangle (that is, without any edges touching, the area of one rectangle is located entirely inside the area of another rectangle). Assuming n is sufficiently large, express this bound as a function of m

   > **Solution:** First, we notice that for no rectangle to be enveloped by another rectangle, there must be no pairs (r_i, r_j) of rectangles r_i and r_j such that r_i envelopes r_j.
   >
   > Hence, let E_{ij} be the event where r_i envelopes r_j and N be the event wherein no rectangle is enveloped by another. Then,
   >
   > $$\mathbb{P}(N) = \mathbb{P}(\bigcap_{1 \le i \ne j \le m} \bar{E}_{ij}) \quad (10)$$
   > $$= 1 - \mathbb{P}(\bigcup_{1 \le i \ne j \le m} E_{ij}) \quad (11)$$
   >
   > Now, we can apply the union bound
   >
   > $$\mathbb{P}(N) \ge 1 - \sum_{1 \le i \ne j \le m} \mathbb{P}(E_{ij}) \quad (12)$$
   > $$= 1 - m(m-1)\mathbb{P}(E_{ij}) \quad (13)$$
   >
   > Here we make use of symmetry since $\mathbb{P}(E_{ij})$ should be the same regardless of the rectangles we choose.

Now, we calculate $\mathbb{P}(E_{ij})$. Notice first that any rectangle can be uniquely represented by a left edge, a right edge, a top edge, and a bottom edge. Also, notice that the location of the left edge and right edge are entirely independent of the locations of the top edge and bottom edge, since each rectangle appears with uniform probability.

Now, we can represent the index of the left and right edges as x_l and x_r, and the locations of the top and bottom edges as y_t and y_b. Thus, the sufficient and necessary conditions for E_{ij} to be satisfied are:

1) $x_{li} < x_{lj} < x_{rj} < x_{ri}$
2) $y_{bi} < y_{bj} < y_{tj} < y_{ti}$

Notice that both of these subproblems are the interval enveloping problem discussed in 1d). Only in this case, we have an interval length of $n - 1$ instead of $2n + 1$ (while the grid has length n, we require that rectangles have length at least 2, whereas intervals in 1d could have length 1. Suppose X_s is the event where the first condition is satisfied and Y_s is the condition where the second condition is satisfied. Then, by independence

$$\mathbb{P}(E_{ij}) = \mathbb{P}(X_s \cap Y_s) = \mathbb{P}(X_s)\mathbb{P}(Y_s) = \left(\frac{1}{6}\frac{(n-2)(n-3)}{n(n-1)}\right)^2 \tag{14}$$

$$= \frac{1}{36}\frac{(n-2)^2(n-3)^2}{n^2(n-1)^2} \tag{15}$$

Plugging this value back into (13) gives us the lower bound

$$\mathbb{P}(N) \geq 1 - \frac{m(m-1)}{36}\frac{(n-2)^2(n-3)^2}{n^2(n-1)^2} \tag{16}$$

The right term converges to 1, as $n \to \infty$ (proof isn't included here, but notice that the leading term of the expanded polynomial fraction has a coefficient of 1 in both the numerator and denominator).

Hence, we get rather a nice lower bound of

$$\mathbb{P}(N) \geq 1 - \frac{m(m-1)}{36} \tag{17}$$