CMPS218 – Homework #1

Zayd Hammoudeh

March 8, 2018

Chapter #2.9, Problem #33

Two points are selected at random on a straight line segment of length 1. What is the probability that a triangle can be constructed out of the three resulting segments?

For the two randomly selected points, x_1 and x_2, define $L_{\max}$ as the length of the longest segment. By the triangle inequality theorem, it is clear that a triangle can only be formed from these three resulting segments if $L_{\max} < 0.5$.

Consider the tuple (x_1, x_2). Each point in the domain $(0,1)^2$ is equally likely since the points are selected randomly. Figure 1 shows the portions of this domain (highlighted in light blue) where the lines segments induced by x_1 and x_2 can be used to form a triangle.

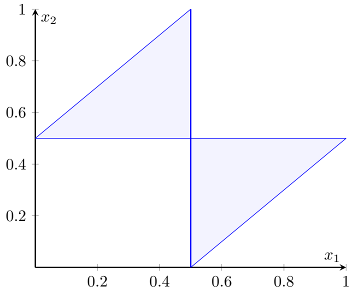

Figure 1: Regions of the domain where the line segments induced by (x_1, x_2) form a triangle.

Since the probability of selecting any point from the above domain is uniform, the probability of selecting points that form a triangle is found via:

$$\begin{aligned}\Pr(\text{Form a triangle}) &= \frac{\text{Shaded Area}}{\text{Total Area}} \\ &= \frac{2 \cdot \left(\frac{1}{2}\right)^3}{1} \\ &= \boxed{\frac{1}{4}}\end{aligned}$$

Chapter #5.8, Problem #30
Scientific American **carried the following puzzle in 1975.**

The poisoned glass: *'Mathematicians are curious birds,' the police commissioner said to his wife. 'You see, we had all those partly filled glasses lined up in rows on a table in the hotel kitchen. Only one contained poison, and we wanted to know which one before searching the glass for fingerprints. Our lab could test the liquid in each glass, but the tests take time and money, so we wanted to make as few of them as possible by simultaneously testing mixtures of small samples from groups of glasses. The university sent over a mathematics professor to help us. He counted the glasses, smiled and said:*
"Pick any glass you want, Commissioner. We'll test it first."
"But won't that waste a test?" I asked.
"No," he said. "it's part of the best procedure. We can test one glass first. It doesn't matter which one."
'How many glasses were there to start with?' the commissioner's wife asked.
'I don't remember. Somewhere between 100 and 200.'
What was the exact number of glasses?

Solve this puzzle and then explain why the professor is in fact wrong and the commissioner was right. What is in fact the optimal procedure for identifying the one poisoned glass? What is the expected waste relative to this optimum if one followed the professor's strategy? Explain the relationship to symbol coding.

The test for poison has a binary outcome, i.e., the sample either has poison or not. Therefore, assuming each cup has poison with equal probability, the size of the remaining set of glasses is, on average, cut in half with each test.

If the number of glasses, n, is a power of 2, then the number of tests required is $\lg n$, where lg is the base-2 logarithm. Note that the only power of 2 between 100 and 200 is 128. There was one extra glass that the professor tested separately. Therefore, there was $\boxed{\text{129 glasses}}$.

Glass ID	Probability of Poison	# Tests
1	1/129	1
2-129	128/129	1 + 7 = 8

Table 1: Number of tests required using the professor's strategy

Table 1 shows the number of tests required when using the professor's strategy. Glass #1 represents the first glass tested, i.e., the one selected at random. In the unlikely event that glass has the poison, no additional testing is required. In contrast, if the poison is in one of the other 128 glasses, seven tests (plus the additional one for the first glass) are required. Using this strategy, the expected number of tests is:

$$\begin{aligned}\mathbb{E}(\text{Professor's Strategy}) &= \frac{1}{129} \cdot 1 + \frac{128}{129} \cdot 8 \\ &\approx \boxed{7.946}.\end{aligned}$$

In contrast, the optimal strategy is:

1. Select one glass at random and leave it off to the side.

2. Test a sample that combines wine from half of the remaining glasses (excluding the one off to the side).
3. If poison is observed in this tested sample, discard the untested glasses. Otherwise, discard the tested glasses.
4. Repeat steps #2 and #3 until only a single glass remains (excluding the one off to the side).
5. If poison was ever observed in any of the previous tests, then the remaining glass not off to the side has the poison, and no additional testing is required.
6. If poison was never observed in any test, then test just the remaining glass not off to the side. If poison is detected, then the answer is clear, and the tested glass has poison; otherwise, the glass off to the side has the poison.

Table 2 shows the number of tests required using this optimum strategy. Note that Glass# 129 entails the remaining glass in step #6 where no sample tests positive for poison up to the last remaining glass not off to the side.

Glass ID	Probability of Poison	# Tests
1	1/129	7 + 1
2-128	127/129	7
129	1/129	7 + 1

Table 2: Number of tests required using the optimum strategy

Using this optimum strategy, the expected number of tests is:

$$\begin{aligned}\mathbb{E}(\text{Optimum Strategy}) &= \frac{2}{129} \cdot 8 + \frac{127}{129} \cdot 7 \\ &\approx \boxed{7.016}.\end{aligned}$$

It is clear then that the expected waste of the professor's strategy is $\boxed{0.93}$ tests.

Maximum compression of a symbol code is achieved by assigning shorter codes (i.e., with less bits) to outcomes with higher probability. In contrast, the professor prioritized the least likely outcome by testing the randomly selected glass first. The optimal strategy described above always tests the most likely outcome (i.e., more glasses at once) similar to how symbol codes are encoded.

Chapter #20.2, Problem #2

Show that as the stiffness β goes to ∞, the soft K-means algorithm becomes identical to the original hard K-means algorithm except for the way in which means assigned no points behave. Describe what those means do instead of sitting still.

In the standard or "hard" K-means algorithm, each point is assigned to exactly one cluster. As such, each of a cluster's points have equal membership.

Soft K-means reduces the rigidity of the standard K-means by introducing a new "stiffness" hyperparameter, β. Rather than each point being a member of exclusively one cluster, the *responsibility* for that point is shared (generally unevenly) among all K clusters. For cluster k and point $\mathbf{x}^{(n)}$, the responsibility, $r_k^{(n)}$ is:

$$r_k^{(n)} = \frac{\exp(-\beta d(\mathbf{m}^{k'}, \mathbf{x}^{(n)}))}{\sum_k \exp(-\beta d(\mathbf{m}^{k'}, \mathbf{x}^{(n)}))} \tag{1}$$

where d is the distance metric, and m^k is the center of cluster k.

As β increases, then even small differences in d can cause massive changes in responsibility. As $\beta \rightarrow \infty$, all responsibility for a point will be assigned to its nearest cluster. This behavior is exactly the same as standard, "hard" K-means where points belong to only the cluster's whose centroid is closest.

When performing Soft K-means with $\beta \rightarrow \infty$, it is possible that some clusters may have no responsibility for no points. In which case, the centroid approaches the zero vector, $\vec{0}$.

Chapter #22.5, Problem #15

The seven scientists. N datapoints $\{x_n\}$ are drawn from N distributions, all of which are Gaussian with a common mean μ but with different unknown standard deviations σ_n. What are the maximum likelihood parameter μ,$\{\sigma_n\}$ given the data? For example, seven scientists (A, B, C, D, E, F, G) with wildly-differing experimental skills to measure μ. You expect some of them to do accurate work (i.e., to have small σ_n), and some of them to turn in wildly inaccurate results (i.e., to have enormous σ_n). Table 3 shows their seven results. What is the μ, and how reliable is each scientist?

Scientist	x_n
A	−27.020
B	3.570
C	8.191
D	9.898
E	9.603
F	9.945
G	10.056

Table 3: Seven measurements $\{x_n\}$ of a parameter μ by seven scientists each having his own noise-level σ_n.

I hope that you agree that, intuitively, it looks pretty certain that A and B are both inept measurers, that D-G are better, and that the true value μ is somewhere close to 10. But what does maximizing the likelihood tell you?

Given n observers that each make a single obserberation with a common mean μ and standard deviations $\{\sigma_n\}$, the maximum likelihood mean, $\bar{x}$, is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

In the case of the seven scientists, the sample mean is:

$$\begin{aligned}\bar{x} &= \frac{-27.020 + 3.570 + 8.191 + 9.898 + 9.603 + 9.945 + 10.056}{7} \\ &= \boxed{3.463}\end{aligned}$$

For observer i where $1 \leq i \leq n$, the maximum likelihood standard deviation is:

$$\sigma_i = |x_i - \mu|$$

since there is only one sample per distribution. Table 4 lists the maximum likelihood standard deviations for the seven scientists.

From the data, it appears that scientists D–G are reliable. Scientist C appears less reliable than them, but better than A and B which appear to be the worst.

Scientist	x_n
A	30.483
B	0.107
C	4.728
D	6.435
E	6.140
F	6.482
G	6.593

Table 4: Maximum likelihood σ for the seven scientists

Clearly for this problem, maximizing the likelihood does not yield the most plausible outcome. There are four observers that essentially measure the same value; it is theoretically possible these measurements are due to coincidence. However, that it is unlikely given the problem description. Therefore, relying blindly on the maximum likelihood calculation may lead to poor conclusions.