Sapienza University of Rome

Big Data Computing

Homework 2

Student

Clara Lecce, 1796575

Academic Year 2023/2024

Assingment 1

The all assignment was done on Colab in the file `1796575-Lecce_HW2_2023.ipynb`, including the report requested.

Assingment 2

a)

The goal for this assignment is to find an estimator, let's call it $\hat{\phi}(x, y)$, of the angle $\phi(x, y)$. We can define the family of hash function as the *sign* function: $h_u(x) = sign(u \cdot x)$, where u is the normal vector to a random hyperplane that intersects x and y. If the hyperplane lies "between" the two vectors (meaning the vectors are on different side of the hyperplane), then the dot product between u and x and y respectively will have different signs. Otherwise, if x and y lie on the same side of the hyperplane, then the dot products $u.x$ and $u.y$ will have the same signs. Let's introduce a random variable $Z_i \sim Ber(p)$:

$$Z_i = \begin{cases} 1 & h_u(x) \neq h_u(y) \\ 0 & h_u(x) = h_u(y) \end{cases}$$

We know from the slides given during class that $\mathbb{P}(Z = 1) = \mathbb{P}(h_u(x) \neq h_u(y)) = \frac{\phi(x,y)}{\pi} = p$, so $\phi(x, y) = p\pi$. Hence, an estimator of this value can be defined by $\hat{\phi}(x, y) = \hat{p}\pi$. So, at the end, the value we finally want to estimate is p, and this can be done by averaging Bernoulli trials:

$$\hat{\phi}(x, y) = \hat{p}\pi = \frac{\pi}{m} \sum_{i=1}^{m} Z_i$$

b)

Using what we have achieved from the previous point, we have the indicator random variable $Z_i \sim Ber\left(p = \frac{\phi(x,y)}{\pi}\right)$, where $\mathbb{E}\left(\sum_{i=1}^{m} Z_i\right) = m\mathbb{E}(Z_1 = 1) = mp = m\frac{\phi(x,y)}{\pi}$ and we can prove that $\hat{\phi}(x, y)$ is an unbiased estimator:

$$\mathbb{E}\left[\hat{\phi}(x,y)\right] = \mathbb{E}\left(\frac{\pi}{m}\sum_{i=1}^{m} Z_i\right) = \frac{\pi}{m}\mathbb{E}\left(\sum_{i=1}^{m} Z_i\right) = \frac{\pi}{m}m\frac{\phi(x,y)}{\pi} = \phi(x,y)$$

So, **to compute the minimum value of m** we can use the **Hint** given, which suggests to use the Chernoff Bounds:

$$\mathbb{P}(|X - \mu| > \varepsilon\mu) \leq 2e^{-\frac{\mu\varepsilon^2}{3}}$$

Developing the formula suggested in the description of the assignment:

$$\begin{aligned}
\mathbb{P}(|\hat{\phi}(x,y) - \phi(x,y)| > \varepsilon\phi(x,y)) &= \mathbb{P}\left(\left|\frac{\pi}{m}\sum_{i=1}^{m} Z_i - \phi(x,y)\right| > \varepsilon\phi(x,y)\right) \\
&= \mathbb{P}\left(\left|\sum_{i=1}^{m} Z_i - m\frac{\phi(x,y)}{\pi}\right| > \varepsilon m\frac{\phi(x,y)}{\pi}\right)
\end{aligned}$$

In our case $X = \sum_{i=1}^{m} Z_i$ and $\mu = \mathbb{E}\left(\sum_{i=1}^{m} Z_i\right) = m\frac{\phi(x,y)}{\pi}$, so we can rewrite the bounds as the following, taking into account that $\phi(x, y) > \theta$:

$$\begin{aligned}
\mathbb{P}(|X - \mu| > \varepsilon\mu) &\leq 2e^{-\frac{\mu\varepsilon^2}{3}} \\
\mathbb{P}\left(\left|\sum_{i=1}^{m} Z_i - m\frac{\phi(x,y)}{\pi}\right| > \varepsilon m\frac{\phi(x,y)}{\pi}\right) &\leq 2e^{-m\frac{\phi(x,y)}{\pi}\frac{\varepsilon^2}{3}} \\
&< 2e^{-m\frac{\theta}{\pi}\frac{\varepsilon^2}{3}} \leq \delta
\end{aligned}$$

Now we can isolate m and calculate its lower bound:

$$\begin{aligned}
2e^{-m\frac{\theta}{\pi}\frac{\varepsilon^2}{3}} &\leq \delta \\
-m\frac{\theta}{\pi}\frac{\varepsilon^2}{3} &\leq \ln\left(\frac{\delta}{2}\right) \\
\ln\left(\frac{2}{\delta}\right) &\leq m\frac{\theta}{\pi}\frac{\varepsilon^2}{3} \\
\ln\left(\frac{2}{\delta}\right)\frac{\pi}{\theta}\frac{3}{\varepsilon^2} &\leq m
\end{aligned}$$

c)

As the final part of this assignment, we want **to find the minimum value of m** such that we have:

$$\mathbb{P}\left(\exists i, j \in \{1, \ldots, n\} : |\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| > \varepsilon\phi(x_i, x_j)\right) \leq \delta$$

The opposite of this assertion is that there aren't **any** pairs of i, j such that the quantity on the left of the inequality is greater than the one on the right, meaning that for every pair of i, j the former is less or equal than the latter, in formulas:

$$1 - \mathbb{P}\left(\forall i, j \in \{1, \ldots, n\} : |\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| \leq \varepsilon\phi(x_i, x_j)\right).$$

But this probability can be seen as the probability of the condition occurring for one couple of vectors, raised to all the possible couples of vectors $x_i, x_j \in \{1, \ldots, n\}$, so:

$$1 - \left(\mathbb{P}\left(|\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| \leq \varepsilon\phi(x_i, x_j)\right)\right)^{\binom{n}{2}}$$

Now, if we want to use the conditions from point b), we can rewrite the probability above as its complementary:

$$1 - \left(1 - \mathbb{P}\left(|\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| > \varepsilon\phi(x_i, x_j)\right)\right)^{\binom{n}{2}}$$

At this point should be easy to finally compute the minimum bound for m:

$$\begin{aligned}
1 - \left(1 - \mathbb{P}\left(|\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| > \varepsilon\phi(x_i, x_j)\right)\right)^{\binom{n}{2}} &\leq \delta \\
\left(1 - \mathbb{P}\left(|\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| > \varepsilon\phi(x_i, x_j)\right)\right)^{\binom{n}{2}} &\geq 1 - \delta \\
1 - \mathbb{P}\left(|\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| > \varepsilon\phi(x_i, x_j)\right) &\geq (1 - \delta)^{-\binom{n}{2}} \\
\mathbb{P}\left(|\hat{\phi}(x_i, x_j) - \phi(x_i, x_j)| > \varepsilon\phi(x_i, x_j)\right) &\leq 1 - (1 - \delta)^{-\binom{n}{2}}
\end{aligned}$$

This is basically the same inequality from point b), so we can use an alias for the "new" δ, let's call it $\psi = 1 - (1 - \delta)^{-\binom{n}{2}}$, and we have the lower bound for m:

$$\ln\left(\frac{2}{\psi}\right)\frac{\pi}{\theta}\frac{3}{\varepsilon^2} \leq m$$