BIG DATA COMPUTING

Homework 2
Nearest Neighbours in High Dimension

Leonardo Di Nino : 1919479

November 20, 2023

Assignment 2

Question (a)

From the theory we know that for cosine distance between $x, y \in \mathbb{R}^d$ a well defined family of hashing functions, given $S^{d-1} = \{u \in \mathbb{R}^d : ||u|| = 1\}$, is the following:

$$H = \{h_u(w) = sign(u^T w) \mid u \sim Unif(S^{d-1})\} \tag{1}$$

We also know that the following hashing property holds:

$$\mathbb{P}(h_u(x) = h_u(y)) = 1 - \frac{\phi(x, y)}{\pi} \tag{2}$$

The following naturally arises from equation (2) leveraging probabilities for complementary events:

$$\frac{\phi(x, y)}{\pi} = \mathbb{P}(h_u(x) \neq h_u(y)) \tag{3}$$

So finally

$$\phi(x, y) = \pi \mathbb{P}(h_u(x) \neq h_u(y)) \tag{4}$$

In order to provide an estimator for $\phi(x, y)$ we need to estimate $\mathbb{P}(h_u(x) \neq h_u(y))$: we can do it through a usual repeated sampling scheme. Since we can always estimate a probability if we can define a proper succession of independent and similar binary random variable, we just leverage the sampling scheme of $u \sim Unif(S^{d-1})$ to build the following estimator: it is unbiased because of continuous mapping on an unbiased estimator.

$$\hat{\phi}(x, y) = \frac{\pi}{m} \sum_{i=1}^{m} \mathbb{1}\{h_{u_i}(x) \neq h_{u_i}(y)\} \tag{5}$$

Question (b)

We want to find the minimum value for the m repeated random samples from H such that the following inequality holds for $\epsilon, \delta \in [0, 1], \phi(x, y) > \theta$:

$$\mathbb{P}(|\hat{\phi}(x, y) - \phi(x, y)| > \epsilon \phi(x, y)) \leq \delta \tag{6}$$

In our case in order to build the estimator we defined a IID succession of binary random variables $\{S_i\}_{i=1}^m$ to estimate the probability $\mathbb{P}(h_u(x) \neq h_u(y))$ through repeated sample and averaging.

In order to retrieve a Chernoff bound is more useful to reason in terms of sum of random variables rather than sample average, so we can rewrite our bound as:

$$\mathbb{P}(|\frac{m}{\pi}\hat{\phi}(x, y) - \frac{m}{\pi}\phi(x, y)| > \frac{m}{\pi}\epsilon \phi(x, y)) \leq \delta \tag{7}$$

Defining $S = \frac{m}{\pi}\hat{\phi}(x, y) = \sum_{i=1}^m S_i$ and leveraging linearity of expectation, since $\mathbb{E}[\hat{\phi}(x, y)] = \phi(x, y)$, than $\mathbb{E}[S] = \mathbb{E}[\frac{m}{\pi}\hat{\phi}(x, y)] = \frac{m}{\pi}\phi(x, y)$.

Finally we can recognize a classical result in concentration of measures for sums of random variables using Chernoff bounds:

$$\mathbb{P}(|S - \mathbb{E}[S]| > \epsilon \mathbb{E}[S]) \leq \delta \tag{8}$$

$$\mathbb{P}(|S - \mathbb{E}[S]| > \epsilon \mathbb{E}[S]) \leq 2\exp\{-\frac{\epsilon^2 \mathbb{E}[S]}{3}\}$$

Plugging in our closed form for $\mathbb{E}[S] = \frac{m}{\pi}\phi(x, y)$, we now impose $\delta \geq 2\exp\{-\frac{\epsilon^2 \frac{m}{\pi}\phi(x,y)}{3}\}$.

In the end:

$$
\begin{aligned}
\delta &\geq 2\exp\{-\frac{\epsilon^2 m\phi(x,y)}{3\pi}\} \\
\frac{\epsilon^2 m\phi(x,y)}{3\pi} &\geq \log(\frac{2}{\delta}) \\
m &\geq \frac{3\pi\log(\frac{2}{\delta})}{\epsilon^2\phi(x,y)}
\end{aligned} \tag{9}
$$

Taking into account that this lower bound on m is decreasing in $\phi(x,y)$ and that clearly we have $\lim\limits_{\phi(x,y)\to 0} \frac{3\pi\log(\frac{2}{\delta})}{\epsilon^2\phi(x,y)} = +\infty$, we take as our best guess $\phi(x,y) = \theta$ since it ensures a working (even if overkilling) estimator for all angles greater than this threshold. At the very end we take

$$
m \geq \frac{3\pi\log(\frac{2}{\delta})}{\epsilon^2\theta} \tag{10}
$$

Question (c)

We want now to retrieve the minimum number of samples m such that given a set of n vectors $\{x_i\}_{i=1}^n \in \mathbb{R}^d$ the following holds:

$$
\mathbb{P}(\exists i,j \in [n] : |\hat{\phi}(x_i,x_j) - \phi(x_i,x_j)| > \epsilon\phi(x_i,x_j) | \forall i,j \in [n] : \phi(x_i,x_j) > \theta) \leq \delta \tag{11}
$$

In order to avoid an abuse of notation let's call the event E "A couple (i,j) exists satisfying the inequality". Subsequently E^c is "There exists no couple (i,j) satisfying the inequality" or equivalently "For all couples (i,j) the inequality is not satisfied". If we call A the event "A couple satisfies the inequality" clearly A^c is the event "A couple doesn't satisfy the inequality" we have the following:

$$
\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - (1 - \mathbb{P}(A))^{|\mathcal{C}_{n,2}|} \tag{12}
$$

where we are assuming all the vectors to be independent. We are going to leverage previous results since the bound we derived on m was in fact for $\mathbb{P}(A) \leq \delta$ If we assume n to be large we can also consider to approximate the total number of couples as $|\mathcal{C}_{n,2}| = \binom{n}{2} \approx \frac{n^2}{2}$. So we rewrite equation (13) as:

$$
\begin{aligned}
1 - (1 - \mathbb{P}(|\hat{\phi}(x,y) - \phi(x,y)| > \epsilon\phi(x,y)))^{\frac{n^2}{2}} &\leq \delta \\
(1 - \mathbb{P}(|\hat{\phi}(x,y) - \phi(x,y)| > \epsilon\phi(x,y)))^{\frac{n^2}{2}} &\geq 1-\delta \\
(1 - \mathbb{P}(|\hat{\phi}(x,y) - \phi(x,y)| > \epsilon\phi(x,y))) &\geq (1-\delta)^{\frac{2}{n^2}} \\
\mathbb{P}(|\hat{\phi}(x,y) - \phi(x,y)| > \epsilon\phi(x,y)) &\leq 1 - (1-\delta)^{\frac{2}{n^2}}
\end{aligned} \tag{13}
$$

So from the final result of the question (b) we have that equation (11) is satisfied if:

$$
m \geq \frac{3\pi\log(\frac{2}{1-(1-\delta)^{\frac{2}{n^2}}})}{\epsilon^2\theta} \tag{14}
$$