Scalable Algorithms for Data Analysis CS-6713

Theory Assignment

Each question consist of ten marks. Total: 30 Marks.

1. Suppose we have two data points let $\vec{a} = \left[a_1, a_2, ..., a_d\right]$ and $\vec{b} = \left[b_1, b_2, ..., b_d\right]$ compressed to $\vec{\alpha} = \left[\alpha_1, \alpha_2, ..., \alpha_k\right]$ and $\vec{\beta} = \left[\beta_1, \beta_2, ..., \beta_k\right]$ using Feature Hashing/Count Sketch Algorithm. Then, prove the following:

$$\begin{aligned}\mathbb{E}\left[\langle\vec{\alpha}, \vec{\beta}\rangle\right] &= \langle\vec{a}, \vec{b}\rangle. \\ \mathrm{Var}\left[\langle\vec{\alpha}, \vec{\beta}\rangle\right] &= \frac{1}{k} \sum_{i \neq i', i, i' \in [d]} \left[{a_i}^2 {b_{i'}}^2 + a_i b_{i'} a_{i'} b_i\right]. \\ &= \frac{1}{k}\left[\|\vec{a}\|_2 \cdot \|\vec{b}\|_2 + \langle\vec{a}, \vec{b}\rangle^2 - 2\sum_{i=1}^{d} {a_i}^2 {b_i}^2\right].\end{aligned}$$

2+8=10 Marks.

2. Suppose we have two data points let $\vec{a} = \left[a_1, a_2, ..., a_d\right]$ and $\vec{b} = \left[b_1, b_2, ..., b_d\right]$ compressed to $\vec{\alpha} = \left[\alpha_1, \alpha_2, ..., \alpha_k\right]$ and $\vec{\beta} = \left[\beta_1, \beta_2, ..., \beta_k\right]$ using Johnson Lindenstrauss lemma (*a.k.a.* random projection). Then, prove the following:

$$\begin{aligned}\mathbb{E}\left[\langle\vec{\alpha}, \vec{\beta}\rangle\right] &= \langle\vec{a}, \vec{b}\rangle. \\ \mathrm{Var}\left[\langle\vec{\alpha}, \vec{\beta}\rangle\right] &= \frac{1}{k}\left[\|\vec{a}\|_2 \cdot \|\vec{b}\|_2 + \langle\vec{a}, \vec{b}\rangle^2\right].\end{aligned}$$

2+8=10 Marks.

3. Recall the problem of *set membership query* that given a set of elements $\mathcal{S} = \{x_1, x_2, \ldots x_n\}$, where $x_i \in \mathbb{R}$ as input the aim is to prepare a data structure that can answer the queries of the following form:

   - Is query item $q \in \mathcal{S}$?

   We know that the Bloom filter gives an elegant solution to this problem.

   We are interested in solving an analogous problem – **approximate set membership query** that given a set of elements $\mathcal{S} = \{x_1, x_2, \ldots x_n\}$, where $x_i \in \mathbb{R}$, and an error parameter $\epsilon > 0$ as input, the aim is to prepare a data structure that can answer the queries of the following form:

   - Is the query item q is sufficiently close to some elements in $\mathcal{S}$. That is,

$$\exists x_i \in \mathcal{S} \text{ such that } ||x_i - q||_2^2 \leq \epsilon,$$

     where $\epsilon > 0$ is a parameter. Construct a data structure for this problem and suggest a trade-off among its size, false positive rate, and the error parameter ϵ.

10 Marks.