中国科学技术大学
计算机科学与技术学院

《大数据算法》作业
2022 年春

截止日期：2022 年 5 月 6 日 23:59

Exercise 1 *20 分*

在 CountSketch 算法及其分析中，我们证明了如果选择 $w > 3k^2$，$d = \Omega(\log n)$，那么以 $1 - \frac{1}{n}$ 的概率，对于任意 $i \in [n]$，$|\tilde{x}_i - x_i| \leq \frac{\|x\|_2}{k}$。这个估计有可能在某些情况是比较坏的，例如当 $\|x\|_2$ 的值主要集中在少数几个坐标上的时候。

对于固定的整数 $\ell > 0$，对于任意 $i \in [n]$，定义向量 $y^{(i)} \in \mathbb{R}^n$ 如下：

$$y^{(i)}_j = \begin{cases} 0 & \text{如果 } j = i \text{ 或者 } j \text{ 是 } x \text{ 中（在绝对值意义下）最大的 } \ell \text{ 个值所对应的坐标之一,} \\ x_j & \text{否则} \end{cases}$$

证明对于 $\ell = k^2$，如果 $w = 6k^2$，$d = \Omega(\log n)$，那么以 $1 - \frac{1}{n}$ 的概率，对于任意 $i \in [n]$，$|\tilde{x}_i - x_i| \leq \frac{\|y^{(i)}\|_2}{k}$。

Exercise 2 *20 分*

假设 k_1, k_2 是两个核（kernel）函数。证明：

(a) 对于任意常数 $c \geq 0$，ck_1 是一个核函数。

(b) 对于任意标量 (scalar) 函数 f，$k_3(\mathrm{x}, \mathrm{y}) = f(\mathrm{x})f(\mathrm{y}) \cdot k_1(\mathrm{x}, \mathrm{y})$ 是一个核函数。

(c) $k_1 + k_2$ 是一个核函数。

(d) $k_1 \cdot k_2$ 是一个核函数。

Exercise 3 *20 分*

令 $X = \mathbb{R}^d$, 并定义 $\mathcal{H}$ 为 X 上的所有 axis-parallel boxes 所构成的集合。具体来说，$\mathcal{H} = \{h_{\mathrm{a,b}} \mid \mathrm{a}, \mathrm{b} \in X\}$。对于 $\mathrm{x} \in X$，$h_{\mathrm{a,b}}(\mathrm{x})$ 定义如下：

$$h_{\mathrm{a,b}}(\mathrm{x}) = \begin{cases} 1 & \text{如果 } a_i \leq x_i \leq b_i \text{ 对于任意的 } i = 1, \ldots, d, \\ -1 & \text{否则。} \end{cases}$$

选择一个可以被 $\mathcal{H}$ 打散 (shatter) 的点集 V，并

(a) 通过证明 V 是可以被 $\mathcal{H}$ 打散的，来证明 $\mathcal{H}$ 的 VC-维（VC-dimension）至少为 $|V|$；

(b) 通过证明不存在大小为 $|V| + 1$ 的点集是可以被 $\mathcal{H}$ 打散的，来证明 $\mathcal{H}$ 的 VC-维至多为 $|V|$。

Exercise 4 *20 分*

一个点集 $S \subseteq \mathbb{R}^d$ 被称为是 “可以被一个间隔（margin）为 γ 的线性分割子（linear separator）所打散的”，

如果对于 S 中所有点的任意一个分类标号（labelling）都是可以被某个间隔为 γ 的线性分割子来实现的。证明在单位球中，不存在一个大小为 $\frac{1}{\gamma^2}+1$ 且可以被一个间隔为 γ 的线性分割子所打散的集合。
提示：考虑感知机 (Perceptron) 算法；尝试反证法。

Exercise 5 *20 分*
令实例空间 (instance space)$X=\{0,1\}^d$，并令 $\mathcal{H}$ 为所有的 3-合取范式公式 (3-CNF formula) 所构成的类。具体来说，考虑所有的由至多 3 个文字 (literal) 的析取（即 OR）所构成的逻辑子句 (clause)，$\mathcal{H}$ 是所有的可以被描述成这样的子句的合取（conjunction）形式的概念（concepts）构成的集合。例如，目标概念 c^* 可能为 $(x_1 \vee \bar{x_2} \vee x_3) \wedge (x_2 \vee x_4) \wedge (\bar{x}_1 \vee x_3) \wedge (x_2 \vee x_3 \vee x_4)$。假设我们在 PAC-learning 的设定中：训练数据中的样本 (examples) 是根据某个分布 D 抽样出来的，它们是根据某个 3-合取范式公式 c^* 来被标号的。

(a) 给出样本个数 m 的一个下界，保证以至少 $1-\delta$ 的概率，对于所有的与训练数据一致 (consistent) 的 3-合取范式公式，其错误都不超过 ε，这里的错误是相对应于分布 D 而言的。

(b) 假设存在一个 3-合取范式公式与训练数据一致，给出一个多项式时间的算法来找到一个这样的公式。
