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 $i \in [n]$，$|\tilde{x}_i - x_i| \le \frac{\|x\|_2}{k}$。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：

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

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 $i \in [n]$，$|\tilde{x}_i - x_i| \le \frac{\|y^{(i)}\|_2}{k}$。

证明. For $i' \in [n]$, let $Y_{i'}$ be the indicator random variable, that is 1 if $h_l(i) = h_l(i')$,then we have

$$Z_l = g_l(i)C[l, h_l(i)] = x_i + \sum_{i' \neq i} g_l(i)g_l(i')x_{i'}Y_{i'}$$

Let $H \subset [n]$ denote the indices of top ℓ entries in magnitude in x, and $T := [n] \setminus H$ be the remaining indices. We have

$$E_r = Z_l - x_i = \sum_{i' \neq i} g_l(i)g_l(i')x_{i'}Y_{i'} = E_1 + E_2$$

E_1 and E_2 are defined as

$$E_1 = \sum_{j \in H \setminus \{i\}} g_l(i)g_l(j)x_jY_j$$
$$E_2 = \sum_{j \in T \setminus \{i\}} g_l(i)g_l(j)x_jY_j$$

On the one hand, since h_l is 2-wise independent, by union bound, we have

$$\begin{aligned} \Pr[E_1 = 0] &= \Pr[\bigwedge_{j \in H \setminus \{i\}} Y_j = 0] \\ &\ge 1 - \frac{\ell}{w} \\ &= \frac{5}{6} \end{aligned}$$

On the other hand, since g_l is also 2-wise independent, we have

$$\begin{aligned}E[|E_2|] &= E[\sum_{j\in T\setminus\{i\}} g_l(i)g_l(j)x_jY_j] \\ &= \sum_{j\in T\setminus\{i\}} E[g_l(i)g_l(j)]E[Y_j]x_j \\ &= 0\end{aligned}$$

therefore

$$\begin{aligned}\mathrm{Var}[|E_2|] &= E[E_2^2] \\ &= E\left[\left(\sum_{j\in T\setminus\{i\}} g_l(i)g_l(j)x_jY_j\right)^2\right] \\ &= E\left[\sum_{j\in T\setminus\{i\}} x_j^2Y_j^2 + \sum_{j_1,j_2\in T\setminus\{i\}}^{j_1\neq j_2} x_{j_1}x_{j_2}g_l(j_1)g_l(j_2)Y_{j_1}Y_{j_2}\right] \\ &= \sum_{j\in T\setminus\{i\}} x_j^2E[Y_j^2] \\ &= \frac{\|y^{(i)}\|_2^2}{w} \\ &= \frac{\|y^{(i)}\|_2^2}{6k^2}\end{aligned}$$

By Chebyshev bound, we have

$$\Pr\left[|E_2| \geq \frac{\|y^{(i)}\|_2}{k}\right] \leq \frac{\frac{\|y^{(i)}\|_2^2}{6k^2}}{\frac{\|y^{(i)}\|_2^2}{k^2}} \leq \frac{1}{6}$$

that is

$$\Pr\left[|E_2| \leq \frac{\|y^{(i)}\|_2}{k}\right] \geq \frac{5}{6}$$

Therefore, by union bound, we have

$$\begin{aligned}\Pr\left[|E_r| \leq \frac{\|y^{(i)}\|_2}{k}\right] &\geq \Pr\left[E_1 = 0 \bigwedge |E_2| \leq \frac{\|y^{(i)}\|_2}{k}\right] \\ &\geq 1 - (\frac{1}{6} + \frac{1}{6}) \\ &= \frac{2}{3}\end{aligned}$$

Via the Chernoff bound, we can conclude that

$$\Pr\left[|\tilde{x}_i - x_i| \leq \frac{\|y^{(i)}\|_2}{k}\right] \geq 1 - \frac{1}{n}$$

Q.E.D. □

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。

证明. Because k_1 and k_2 are kernel functions, their corresponding kernel matrices K_1 and K_2 are PSD. Consequently, $\forall \alpha \in \mathbb{R}^n$, we have

$$\alpha^T K_1 \alpha \geq 0$$

Since $c \geq 0$, we have

$$\alpha^T (cK_1)\alpha = c(\alpha^T K_1 \alpha) \geq 0$$

So cK_1 is PSD. Therefore, the corresponding function ck_1 is a kernel function.
Similarly, we have

$$\begin{aligned}\alpha^T (K_1 + K_2)\alpha &= \alpha^T K_1 \alpha + \alpha^T K_2 \alpha \\ &\geq 0\end{aligned}$$

Thus, the corresponding kernel function $k_1 + k_2$ is also a kernel function.
Since k_1 is a kernel function, for each entry $k_1(x_i, x_j)$ of its kernel matrix K_1, we have

$$k_1(x_i, x_j) = (\psi_1(x_i))^T \psi_1(x_j)$$

Consequently, we have

$$\begin{aligned}k_3(x_i, x_j) &= f(x_i)f(x_j)k_1(x_i, x_j) \\ &= f(x_i)f(x_j)(\psi_1(x_i))^T \psi_1(x_j) \\ &= (f(x_i)(\psi_1(x_i))^T)(f(x_j)\psi_1(x_j)) \\ &= (f(x_i)(\psi_1(x_i)))^T (f(x_j)\psi_1(x_j))\end{aligned}$$

Let $\psi_3(x) = f(x)\psi_1(x)$, the above equation can be written as

$$k_3(x_i, x_j) = (\psi_3(x_i))^T \psi_3(x_j)$$

So k_3 is a kernel function.
Let $\lambda_1 \dots \lambda_n \geq 0$ be eigenvalues of K_1 and $\mu_1 \dots \mu_n \geq 0$ be eigenvalues of K_2, we have

$$\begin{aligned}K_1 &= \sum_{i=1}^{n} \lambda_i u_i u_i^T \\ K_2 &= \sum_{i=1}^{n} \mu_i v_i v_i^T\end{aligned}$$

Define notation $A \odot B$ as the entry-wise product of two matrices A and B. We have

$$\begin{aligned}K_1 \odot K_2 &= (\sum_{i=1}^{n} \lambda_i u_i u_i^T) \odot (\sum_{i=1}^{n} \mu_i v_i v_i^T) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j (u_i u_i^T) \odot (v_j v_j^T) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} (\lambda_i \mu_j)(u_i \odot v_j)(u_i \odot v_j)^T\end{aligned}$$

Because $\lambda_i \mu_i \geq 0$, $K_1 \odot K_2$ is PSD. Hence the corresponding function $k_1 \cdot k_2$ is a kernel function.
Q.E.D. □

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 $|V|$。

(a)

证明. Let $V = \{v \in \mathbb{R}^d | \exists 0 \leq i \leq d-1, v_i = \pm 1, \bigwedge_{j \neq i} v_j = 0\}$, then for any subset $S = \{v_{s_1}, v_{s_2} \ldots v_{s_k}\}$ of V, generate a and b of $h_{a,b}$ by the following data streaming algorithm:

(a) Initialize $a_i = 0$ and $b_i = 0$ for each $0 \leq i \leq d-1$;

(b) For each item v_{s_i}, by definition, there is some $0 \leq w \leq d-1$ s.t. $v_{s_i w} = \pm 1$ and $\bigwedge_{j \neq w} v_{s_i j} = 0$. Set $a_w = -1$ if $v_{s_i w} = -1$; set $b_w = 1$ if $v_{s_i w} = 1$;

(c) Output a and b.

On the one hand, $\forall x \in S$, since a_i and b_i are set as the smallest and the largest value of all x_i respectively, $a_i \leq x_i \leq b_i$ holds for $\forall 0 \leq i \leq d-1$. Hence $\forall x \in S$, we have $h_{a,b}(x) = 1$.
On the other hand, $\forall x \notin S$, there must be some index i s.t. $x_i = 1$ or $x_i = -1$ while $\forall y \in S, y_i \leq 0$ or $y_i \geq 0$. What is meant by that is, there is either $x_i > b_i$ or $x_i < a_i$. Therefore, we have $h_{a,b}(x) = -1$ holds for $\forall x \notin S$.
In conclusion, $\forall S \subset V$, S can be expressed as $h \cap V$ for some $h \in \mathcal{H}$. In other words, V can be shattered by $\mathcal{H}$. Consequently, we have

$$\text{VC-dimension}(\mathcal{H}) \geq |V| = 2d$$

Q.E.D. □

(b)

证明. Assuming that there exits V' of size $|V'| = 2d+1$ s.t. V' can be shattered by $\mathcal{H}$. Consider the smallest axis-parallel box $h^*_{a^*,b^*}$ s.t. $\forall x \in V'$, we have $h^*_{a^*,b^*}(x) = 1$.
If there exits at least one point x^* in the interior of box h^*, then we can not find a $h_{a,b} \in \mathcal{H}$ s.t. $h_{a,b}(x^*) = -1$ while for all point at the boundary of h^* we have $h_{a,b}(x) = 1$. Otherwise, there will be $h^*_{a^*,b^*}(x^*) = -1$ as well, which contradicts to our assumption.

Otherwise, if none of the points are in the interior of box h^*, i.e. $\forall x \in V'$, x is at the boundary of box h^*. Define the *face* of the box h^* as

$$f_{a_j^*}(x) = \begin{cases} 1 & x_j = a_j^*, \bigwedge_{i \neq j} a_i^* \leq x_i \leq b_i^* \\ -1 & \text{o.w.} \end{cases}$$

and

$$f_{b_j^*}(x) = \begin{cases} 1 & x_j = b_j^*, \bigwedge_{i \neq j} a_i^* \leq x_i \leq b_i^* \\ -1 & \text{o.w.} \end{cases}$$

Obviously, there are $2d$ faces of box h^* in total. By the pigeonhole principle, since there are $2d+1$ points in V', at least 2 of them (x_1 and x_2) are in the same face. Assume this face is $f_{a_j^*}$, if there exists at least one point (x_3) in a face other than $f_{b_j^*}$, then we will fail to find either a box $h_1 \in \mathcal{H}$ s.t. $h_1 \cap V' = \{x_1, x_3\}$ or a box $h_2 \in \mathcal{H}$ s.t. $h_2 \cap V' = \{x_2, x_3\}$. Otherwise, if all points are in face $f_{a_j^*}$ and face $f_{b_j^*}$, the problem will degenerate into shattering $2d+1$ points in $(d-1)$-dimensional space. By induction, this is impossible.
In conclusion, in a d-dimensional space, there doesn't exist a set V' of size $|V'| = 2d+1$ s.t. it can be shattered by $\mathcal{H}$. Therefore, the VC-dimension of $\mathcal{H}$ is $2d$.
Q.E.D. □

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。

证明. Assuming that there exists a set of size $\frac{1}{\gamma^2}+1$ that can be shattered by a linear separator whose margin is γ. Thus, by the Perceptron algorithm, for any given set of labels $\{l_i | 1 \leq i \leq \frac{1}{\gamma^2}+1\}$, we can obtain such a linear separator

$$w = \sum_{i=1}^{\frac{1}{\gamma^2}+1} \alpha_i x_i l_i$$

where $a_i \in \mathbb{N}, \forall 1 \leq i \leq \frac{1}{\gamma^2}+1$ after at most

$$\left(\frac{R}{\gamma}\right)^2 = \frac{1}{\gamma^2}$$

updates where γ is the margin and $R = \max_i \|x_i\| = 1$ in a unit sphere for any execution order. However, for any set of size $\frac{1}{\gamma^2}+1$, considering a labeling

$$l_i = \begin{cases} 1 & i = 1 \\ -\text{sgn}((\sum_{k=1}^{i-1} x_k l_k)^T x_i) & \text{o.w.} \end{cases}$$

Therefore, when the execution order is $x_1, x_2 \ldots$, $w^T x_i l_i \leq 0$ will hold for the first $\frac{1}{\gamma^2}+1$ steps. In other words, the number of updates in this execution order will be at least $\frac{1}{\gamma^2}+1$, which is bigger than $\left(\frac{R}{\gamma}\right)^2 = \frac{1}{\gamma^2}$, contradicting to the update number bound of the Perceptron algorithm. Consequently, there doesn't exist a set of size $\frac{1}{\gamma^2}+1$ that can be shattered by a linear separator whose margin is γ.

Q.E.D. $\square$

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。

(a)

Solution. Considering a training method of 3-CNF formula by eliminating elements, based on samples in the training set, from the conjunction of all possible clauses constructed by at most 3 literals, we have

$$|\mathcal{H}| = 2^{2^3\binom{d}{3}+2^2\binom{d}{2}+2\binom{d}{1}}$$

Since c^* is consistent with the training set, i.e. the training error $\mathrm{err}_s(h)=0$, we have

$$\begin{aligned} m &\geq \frac{1}{\varepsilon}(\ln(|\mathcal{H}|)+\ln(\frac{1}{\delta})) \\ &= \frac{1}{\varepsilon}((2^3\binom{d}{3}+2^2\binom{d}{2}+2\binom{d}{1})\ln(2)+\ln(\frac{1}{\delta})) \end{aligned}$$

s.t. the true error $\mathrm{err}_D(h)\leq\varepsilon$ w.p. at least $1-\delta$.

(b)

Solution.

(a) Start with the 3-CNF formula that is constructed by the conjunction of all possible clauses of size at most 3;

(b) $\forall x\in\{0,1\}^d$ in the training set, calculate its corresponding values in all clauses of size at most 3, denoting them as $\{c_i(x) | 1\leq i\leq 2^3\binom{d}{3}+2^2\binom{d}{2}+2\binom{d}{1}\}$;

(c) $\forall x\in\{0,1\}^d$ in the training set, if it is labelled as 1, eliminate c_i from the original 3-CNF formula if $c_i(x)=0$;

(d) Output the 3-CNF formula after the above eliminations.

The time complexity of the algorithm is

$$\begin{aligned} T &= \mathcal{O}(m(2^3\binom{d}{3}+2^2\binom{d}{2}+2\binom{d}{1})) \\ &= \mathcal{O}(md^3) \end{aligned}$$

Therefore, the algorithm has a polynomial time complexity.