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

设 $C[m,s]$ 为 h_m 投影到 s 的计数器，$m \in [d]$，$s \in [w]$，

$M_i := \{ j \mid j \in x$ 中最大的 l 个值的对应坐标 或 $j = i \}$，$M_i^c = [n] \setminus M_i$

记 $Z_m = g_m(i)\, C[m, h_m(i)]$，$Y_j = \begin{cases} 1, & h_m(i) = h_m(j) \\ 0, & else \end{cases}$

则 $Z_m = x_i + \sum\limits_{j \in M_i \setminus \{i\}} g_m(j) g_m(i) Y_j x_j + \sum\limits_{j \in M_i^c} g_m(j) g_m(i) Y_j x_j$

设 $A = \{ \sum\limits_{j \in M_i \setminus \{i\}} g_m(j) g_m(i) Y_j x_j = 0 \}$

$P(A) = P(Y_j = 0 \text{ for all } j \in M_i \setminus \{i\})$

又 $P(h_m(i) = h_m(j)) = \frac{1}{w}$，$\therefore P(A) \geqslant (1 - \frac{1}{w})^l \geqslant 1 - \frac{l}{w} = \frac{5}{6}$

记 $Z_m' = \sum\limits_{j \in M_i^c} g_m(i) g_m(j) Y_j x_j$

则 $E[Z_m'] = 0$，$Var[Z_m'] < \frac{\|y^{(i)}\|^2}{w}$

记 $P(B) = P(|Z_m'| \le \frac{\|y^{(i)}\|}{k})$

由 Chebyshev's 不等式，$P(B) > 1 - \frac{Var[Z_m']}{\|y^{(i)}\|_2^2} \cdot k^2 = 1 - \frac{k^2}{w} = \frac{5}{6}$

若 A 和 B 同时发生，则有 $|Z_m - x_i| \le \frac{\|y^{(i)}\|}{k}$

$\therefore P(|Z_m - x_i| > \frac{\|y^{(i)}\|_2}{k}) \le P(\bar{A} \cup \bar{B}) \le P(\bar{A}) + P(\bar{B}) \le \frac{1}{3}$

$\therefore P(|Z_m - x_i| \le \frac{\|y^{(i)}\|}{k}) \geqslant \frac{2}{3}$

设 $\tilde{x}$ 为 $\{Z_1, Z_2, \cdots, Z_d\}$ 的中位数，则由 Chernoff bound，

对 $d = \Omega(\log n)$，$P(|\tilde{x} - x| > \frac{\|y^{(i)}\|}{k}) < e^{-cd} < \frac{1}{n}$

对于 $l = k^2$，如果 $w = 6k^2$，$d = \Omega(\log n)$，那么以 $1 - \frac{1}{n}$ 的概率 $|\tilde{x} - x| \le \frac{\|y^{(i)}\|}{k}$

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。

如果 $k(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$，则它为核函数.

设 $k_1(x, y) = \varphi_1(x)^T \varphi_1(y)$ ，$k_2(x, y) = \varphi_2(x)^T \varphi_2(y)$

(a) $ck_1(x, y) = c\varphi_1(x)^T \varphi_2(y) = (\sqrt{c}\,\varphi_1(x))^T (\sqrt{c}\,\varphi_2(y))$

故 $\forall\, c \geq 0$，ck_1 是一个核函数.

(b) $k_3(x, y) = f(x)f(y)k_1(x,y) = f(x)f(y)\varphi_1(x)^T\varphi_1(y)$

$= (f(x)\varphi_1(x))^T (f(y)\varphi_1(y))$

故 $\forall$ 标量函数 f，$k_3(x,y) = f(x)f(y)k_1(x,y)$ 是一个核函数.

(c) K 是核矩阵，即 $\exists$ 函数 φ s.t. $K_{ij} = \varphi(x_i)^T\varphi(x_j)$

$\Leftrightarrow$ K 是半正定矩阵

设 $K^{(1)}$ 是 k_1 的核矩阵，$K^{(2)}$ 是 k_2 的核矩阵，则 $K^{(1)} + K^{(2)}$ 是 $k_1 + k_2$ 的核矩阵.

有 $x^T K^{(1)} x \geq 0$ ，$x^T K^{(2)} x \geq 0$

则 $x^T (K^{(1)} + K^{(2)}) x \geq 0$

$\Rightarrow$ $K^{(1)} + K^{(2)}$ 是核矩阵 $\Rightarrow$ $k_1 + k_2$ 是核函数

(d) 记 $\varphi_1(x) = (\varphi_{11}(x), \varphi_{12}(x), \cdots, \varphi_{1d}(x))$，$\varphi_2(x) = (\varphi_{21}(x), \varphi_{22}(x), \cdots, \varphi_{2d}(x))$

记 $k_4(x,y) = k_1(x,y)\,k_2(x,y) = \varphi_1(x)^T\varphi_1(y) \cdot \varphi_2(x)^T\varphi_2(y)$

$= \left(\sum_{i=1}^{d} \varphi_{1i}(x)\varphi_{1i}(y)\right)\left(\sum_{j=1}^{d} \varphi_{2j}(x)\varphi_{2j}(y)\right)$

$= \sum_{i=1}^{d}\sum_{j=1}^{d} (\varphi_{1i}(x)\varphi_{2j}(x)\varphi_{1i}(y)\varphi_{2j}(y))$

记 $\varphi_{4ij}(x) = \varphi_{1i}(x)\varphi_{2j}(x)$，则 上式 $= \sum_{i=1}^{d}\sum_{j=1}^{d} \varphi_{4ij}(x)\varphi_{4ij}(y) = \varphi_4(x)^T\varphi_4(y)$

故 $k_4 = k_1 \cdot k_2$ 是核函数.

Exercise 3 *20 分*

考虑如下的在线学习（online learning）场景：在每个时刻 $t = 1, 2, \ldots,$, 下面两个事件依次发生：

(a) 算法看到任意一个样本 x_t，然后对它的标号（label）进行预测，令 ℓ_t' 为其预测值

(b) 接下来算法被告知样本的真实标号 ℓ_t，如果 $\ell_t' \neq \ell_t$，则我们称其犯了一个错误

我们的目标是设计一个在线的分类算法，使得其犯错次数越小越好。

在课堂上，我们分析了对于线性可分的 (linearly separable) 数据样本集，感知机 (Perceptron) 算法的正确性与效率。现在我们考虑如下的感知机算法：

- 令 $w = 0$，即 w 为全 0 向量。对于时刻 $t = 1, 2, \ldots$, 算法如下操作：
  (a) 对于样本 x_t，预测其标号为 $\mathrm{sgn}(x_t^T w)$
  (b) 如果预测值是错误的，那么更新 w 为 $w + x_t \ell_t$

证明下面的结论：

对于任意的样本序列 $x_1, x_2, \ldots$, 如果存在一个向量 w^* 满足对于任意的 $t \geq 1$，$x_t^T w^* \ell_t \geq 1$(即 $(w^*)^T x = 0$ 是一个间隔 (margin) 至少为 $\gamma = 1/\|w^*\|$ 的线性分割子)，那么上述的感知机算法犯错的次数不超过 $r^2\|w^*\|^2$ 次，这里的 $r = \max_t \|x_t\|$。

提示：参考课堂上 Perceptron 算法的分析。

$\because \exists\ w^*$ s.t. $\forall t \geq 1$, $x_t^T w^* l_t \geq 1$

$\therefore (w + x_t l_t)^T w^* = w^T w^* + x_t^T l_t w^* \geq w^T w^* + 1$

$\therefore$ 每次犯错会让 $w^T w^*$ 至少增加 1

$(w + x_t l_t)^T (w + x_t l_t) = \|w\|^2 + 2x_t^T l_t w + \|x_t\|^2$

$\because$ 犯错时 $x_t^T l_t w \leq 0$ $\therefore$ 上式 $\leq \|w\|^2 + \|x_t\|^2$

又 $\because r = \max_t \|x_t\|$ $\therefore$ 上式 $\leq \|w\|^2 + r^2$

$\therefore$ 每次犯错会让 $\|w\|^2$ 至多增加 r^2

设算法犯错的次数为 m，则

$$w^T w^* \geq m \quad , \|w\|^2 \leq m \cdot r^2$$

$\|x\|\|y\| \geq |\langle x, y\rangle|$ $\Rightarrow \|w\|\|w^*\| \geq m$, $\|w\| \leq \sqrt{m} \cdot r$

$$\Rightarrow \frac{m}{\|w^*\|} \leq \|w\| \leq \sqrt{m} \cdot r$$

$$\Rightarrow \sqrt{m} \leq r\|w^*\|$$

$$\Rightarrow m \leq r^2 \|w^*\|^2$$

$\therefore$ 犯错的次数不超过 $r^2\|w^*\|^2$.

Exercise 4 *20 分*

考虑数据样本集合不一定线性可分的情形。给定了数据样本（及其标号）集合 $(x_i, \ell_i), 1 \le i \le n$ 及一个向量 w^*，定义（样本 x_i 相对于 w^* 的）hinge 损失为

$$L_{hinge}(w^*, x_i) = \max\{0, 1 - x_i^T w^* \ell_i\}, \text{对任意的 } i \le n$$

证明下面的结论：对于一个样本序列 $S = x_1, x_2, \ldots,$，上题中所描述的（在线）感知机算法所犯错误最多为

$$\min_{w^*}(r^2\|w^*\|^2 + 2L_{hinge}(w^*, S)),$$

这里 $r = \max_t \|x_t\|$，$L_{hinge}(w^*, S) = \sum_{i=1}^n L_{hinge}(w^*, x_i)$。

提示：参考 Perceptron 算法的分析。注意到，对于一个标号为正的样本，每次 $w^T w^*$ 增加 $x_t^T w^*$，后者至少为 $1 - L_{hinge}(w^*, x_t)$。对于标号为负的情形类似。

对于一个正样本，每次犯错 $w^T w^*$ 增加 $x_t^T w \geq 1 - L_{hinge}(w^*, x_i)$

对于一个负样本，每次犯错 $w^T w^*$ 增加 $-x_t^T w \geq 1 - L_{hinge}(w^*, x_i)$

设算法犯错的次数为 m，则

$$w^T w^* \geq m - L_{hinge}(w^*, S), \quad \|w\|^2 \leq m \cdot r^2$$

$$\|w^T\| \|w^*\| \geq m - L_{hinge}(w^*, S)$$

$$\|w^T\|^2 \|w^*\|^2 \geq (m - L_{hinge}(w^*, S))^2$$

$$mr^2 \|w^*\|^2 \geq m^2 - 2m L_{hinge}(w^*, S) + L_{hinge}^2(w^*, S)$$

$$m \leq r^2 \|w^*\|^2 + 2 L_{hinge}(w^*, S) - \frac{1}{m} L_{hinge}^2(w^*, S)$$

$$m \leq r^2 \|w^*\|^2 + 2 L_{hinge}(w^*, S)$$

$\therefore$ 犯错的次数最多为 $\min_{w^*}(r^2\|w^*\|^2 + 2L_{hinge}(w^*, S))$

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) $X = \{0,1\}^d$，则最多有 $(2d)^3$ 个子句，

则共有 $2^{(2d)^3}$ 个可能的 3-CNF 公式，即 $|H| \leq 2^{(2d)^3}$

又有定理当样本个数 $m \geq \frac{1}{\varepsilon}(\ln|H| + \ln(\frac{1}{\delta}))$ 时，

可以保证以至少 $1-\delta$ 的概率，对于所有的与训练数据一致的 3-CNF 公式，

其错误相对应于分布 D 而言不超过 ε.

即 $m \geq \frac{1}{\varepsilon}\left(\ln(2^{(2d)^3}) + \ln(\frac{1}{\delta})\right) = \frac{1}{\varepsilon}\left((2d)^3\ln 2 + \ln\frac{1}{\delta}\right)$

(b) 设 $h = \varphi_1 \wedge \varphi_2 \wedge \cdots \wedge \varphi_{(2d)^3}$，其中 $\varphi_i = x_{i1} \vee x_{i2} \vee x_{i3}$

```
for i = 1, ..., n :
    if l_i == 1 :
        for j = 1, ..., (2d)^3 :
            if x⃗(i) leads φj = 0 :
                drop φj from h
return h
```

该算法包含 2 个 for 循环，时间复杂度为 $O(nd^3)$

*参考了教材《Foundation of Data Science》与老师讲义