Домашнее задание №6.

№1. Пусть у нас есть выборка $S = (x_i; y_i)$, $x_i \in \mathbb{R}$, $y_i \in \{-1, 1\}$ $i = \overline{1, m}$

①. Отсортируем выборку по возрастанию первой компоненты в паре $s_i \to x_i$.

Сортировка $O(m \log m)$.

②. Построим массив кумулятивных сумм a_i для наших отсортированных вторых компонент.

$O(m)$.

③ Далее нам требуется найти такие индексы i и j, что $s_j - s_{i-1}$ будет являться максимальным. Т.е. найти такой отрезок i, j с максимальной суммой. $O(m^2)$

⑤ Полученные i и j - это индексы объектов в выборке, которые требуется взять в качестве границ для интервала ERM_H.

Сложность: $O(m \log m) + O(m) + O(m^2) = O(m^2)$.

№2. H_i - последовательность классов гипотез для задачи бинарной классификации.

При выполнении предположения о реализуемости ERM_{H_n} за $O(n)$, empirical risk - $O(mn)$.

Если не выполняется предположение о реализуемости.

По условию дано, что ERM гипотеза в классе H_n определяется $O(n)$ примерами из тренировочной выборки. Тогда $\forall H_k \in H_n$ нам требуется количество элементов из выборки $\le kn$.

Тогда получаем, что при поиске ERM гипотезы нам придется рассмотреть все такие сочетания из тренировочной выборки, что их размер не превосходит kn.

$$\sum_{i=1}^{kn} |C_m^i| \leq m^{kn}$$ ← количество размещений с повторениями

Тогда в асимптотике получаем, что поиск оптимальной гипотезы будет занимать:

$m^{O(n)}$ — количество гипотез

$O(n)$ — строим гипотезу

$O(mn)$ — вычисляем риск

$$\Rightarrow O(m^{O(n)})\cdot\big(O(n)+O(mn)\big) = O(m^{O(n)}n).$$

№3.2. $X=\mathbb{R}^n$, H_k^n — класс пересеч. k лин. полупространств в $\mathbb{R}^n$

Воспользуемся подсказкой из условия и построим тренировочное мн-во:

$G=(V,E)$, $V=\{v_1,\dots v_n\}$, $S(G)\in(\mathbb{R}^n\times\{\pm1\})^m$, $m=|V|+|E|$.

1) $v_i\in V \Rightarrow e_i,\ y_i=-1$
2) $(v_i,v_j)\in E \Rightarrow \frac{e_i+e_j}{2}\quad y_{ij}=1$.

а) Если существует гипотеза $h\in H_k^n$ с нулевым эмпирическим риском, то граф k-раскрашиваем.

◇: Строим следующую раскраску: $f(v_i)=\arg\min_t\{h_t(e_i)=-1\}$, такое t существует по построению и $t\le k$ (т.к. у нас k гипотез в H_k^n). Покажем, что никакие две смежные вершины не окрашены в один цвет.

Пойдём от обратного.

Пусть v_i и v_j соединены ребром и $f(v_i) = f(v_j) = t$

$\Rightarrow h(e_i) = h(e_j) = -1.$

Но знаем, что $h\left(\frac{e_i + e_t}{2}\right) = 1$

$$h\left(\frac{e_i+e_j}{2}\right) = \operatorname{sign}\left(\langle w_t, \frac{e_i+e_j}{2}\rangle + b_t\right) =$$
$$= \operatorname{sign}\left((\langle w_t, e_i\rangle + b_t) + (\langle w_t, e_j\rangle + b_t)\right)$$

Учтём, что $h(e_i) = h(e_j) = \operatorname{sign}(\langle w_t, e_i\rangle + b_t) = -1.$

$$\Rightarrow \langle w_t, e_i\rangle + b_t < 0.$$

$$\Rightarrow h\left(\frac{e_i+e_j}{2}\right) < 0 \text{ и } h\left(\frac{e_i+e_j}{2}\right) = -1.$$

Противоречие. $\varnothing$

б) Если существует k-раскраска G, то существует гипотеза $h \in H_k^n$ с нулевым эмпирическим риском.

Подсказка: $b = 0.6$ и $t \le k \Rightarrow w_{t,i} = -1,\ f(v_i) = t$
$w_{t,i} = 0$ иначе

$\diamond$: Построим выборку: $e_{i,j} = \begin{cases} 1, & j = f(v_i) \\ -2 + \varepsilon_i, & j = f(v_i) \quad j \le k \\ 0, & k < j \end{cases}$

Пусть ε_i - различные для всех i.

Строим гипотезу: $b_t = 0.6$ и $w_{t,j} = \begin{cases} -1, & j = t \\ 0, & j \ne t \end{cases}$

Тогда $h_{f(v_i)}(e_i) = \operatorname{sign}(-0.6) = -1$ и т.к. $f(v_i) \ne f(v_j)$ для смежных вершин v_i и v_j, то

$$e_i + e_j = \begin{cases} -1, \{t = f(v_i)\} \cup \{t = f(v_j)\} \\ -4, \{t \neq f(v_i)\} \cap \{t \neq f(v_j)\} \cap \{t < k\} \\ 0, k < t \end{cases}$$

$$h\left(\frac{e_i + e_j}{2}\right) = \text{sign}\left(\langle w_t, \frac{e_i + e_j}{2} \rangle + 0.6\right) = 1$$

> 0 исходя из построения w_t и $e_i + e_j$

в) Показали, что задаче поиска гипотезы с нулевым риском над выборкой в H_k^n можно поставить в соответствие задачу о раскраске графа.

В случае выполнения предположения о реализуемости алгоритм будет находить гипотезу с нулевым риском. Получили, что задача $ERM_{H_k^n}$ является NP-трудной.