Спецкурс 2020/2021: “Геометрические и комбинаторные свойства матриц и аппроксимация” Блок лекций “Сложность матриц и аппроксимация” Лекция 5: “Реализация матриц с большим отступом (margin)”

24 ноября 2020 г.

Определение

Пусть $S \in \{-1, 1\}^{m\times n}$ — сигнум-матрица.
Определим “максимальный отступ” матрицы:

$$\mathrm{margin}(S) := \max_{\{x_i\},\{y_j\}} \min_{i,j} \frac{|\langle x_i, y_j\rangle|}{|x_i|\cdot|y_j|},$$

где максимум берётся по всем *реализациям* матрицы, т.е. таким наборам векторов $x_1, \dots, x_m$, $y_1, \dots, y_n$, что $S_{i,j} = \mathrm{sign}\langle x_i, y_j\rangle \quad \forall i, j$.
Смысл в том, что равенство $S_{i,j} = \mathrm{sign}\langle x_i, y_j\rangle$ должно выполняться с большим “запасом”.

Margin — отступ, зазор.

Определим сложность реализации с отступом (*margin complexity*) сигнум матрицы S как

$$\mathrm{mc}(S) := \mathrm{margin}^{-1}(S).$$

(Чем сложнее реализовать матрицу, тем больше сложность.)

Базовые свойства

Оцените $\mathrm{mc}(S)$:

- "Диагональная" матрица $n \times n$: $S_{i,i} = 1$, $S_{i,j} = -1$ при $i \neq j$.
- $S_{i,i} = -1$, $S_{i,j} = 1$ при $i \neq j$.
- $S_{1,1} = 1$, $S_{i,j} = -1$ при остальных i, j.

Утверждение

Для любой матрицы $S \in \{-1, 1\}^{m \times n}$ имеем

$$1 \leqslant \mathrm{mc}(S) \leqslant \min\{\sqrt{m}, \sqrt{n}\}.$$

Докажите оценку снизу. $\mathrm{margin}(S) \leqslant 1$, т.к. $|\langle x_i, y_j \rangle| \leqslant |x_i||y_j|$.

Докажите оценку сверху. Пусть $m \leqslant n$. Возьмём в качестве $\{y_j\}_{j=1}^n$ столбцы матрицы S, в качестве $\{x_i\}_{i=1}^n$ — базисные вектора. Тогда $|x_i| \cdot |y_j| \leqslant m^{1/2}$ и $\langle x_i, y_j \rangle = S_{i,j}$.

Классификация с максимальным отступом

Пусть $(x_1, t_1), \ldots, (x_n, t_n)$ — выборка из некоторого множества объектов. Каждый объект принадлежит одному из двух классов: $t_i = 1$ или $t_i = -1$. Вектор $x_i \in \mathbb{R}^d$ состоит из d чисел (признаков), описывающих i-й объект.

Задача классификации состоит в построении функции $\hat{f} \colon \mathbb{R}^d \to \{-1, 1\}$, позволяющей отличать объекты разных классов. То есть, ошибка $\mathrm{Err} = \mathsf{P}(t \neq \hat{f}(x))$ должна быть мала.

В линейной классификации функция $\hat{f}$ строится с помощью линейной: $\hat{f}(x) = \mathsf{sign}\langle b, x\rangle$, $b \in \mathbb{R}^d$. (Случай аффинной функции $b_0 + \langle b, x\rangle$ можно свести к линейному, добавим признак, тождественно равный 1.)

Пространство $\mathbb{R}^d$ гиперплоскостью $\langle b, x\rangle = 0$ разделяется на два полупространства: $\langle b, x\rangle > 0$, классифицируется как $t = 1$, и $\langle b, x\rangle < 0$ (соответственно, $t = -1$).

Классификация с максимальным отступом

Разделяющая гиперплоскость $\langle b, x \rangle = 0$, строится по обучающей выборке $(x_1, t_1), \ldots, (x_n, t_n)$.

Предположим, мы можем получить на обучении нулевую ошибку: существует ненулевой вектор b, такой что

$$t_i = \operatorname{sign}\langle b, x_i \rangle, \quad i = 1, \ldots, n.$$

Какой b взять, чтобы ошибка на тестовой выборке была поменьше?

Потребуем, чтобы b обеспечивал правильную классификацию с максимальным запасом, *отступом*:

$$\begin{cases} \min\limits_i |\langle b, x_i \rangle| \to \max, \\ t_i = \operatorname{sign}\langle b, x_i \rangle, \quad i = 1, \ldots, n, \\ |b| = 1. \end{cases}$$

Эквивалентная формулировка: мы проводим разделяющую гиперплоскость так, чтобы максимизировать расстояние от точек до края.

Опорные вектора

Можно показать, что b является линейной комбинацией *опорных векторов* x_i, для которых $|\langle b, x_i \rangle|$ минимально.
Отсюда название *Метод Опорных Векторов*, Support Vector Machine (SVM). На практике в SVM используется так называемый *soft margin*, когда мы не требуем, чтобы гиперплоскоть правильно разделяла обучающую выборку; разрешаем "залезать за край", но "штрафуем" за это:

$$\begin{cases} |b|^2 + C\sum_{i=1}^{n} \xi_i \to \min, \\ t_i\langle b, x_i \rangle \geqslant 1 - \xi_i, \quad i = 1, \dots, n \\ \xi_i \geqslant 0, \quad i = 1, \dots, n. \end{cases}$$

Классификация с максимальным отступом

Вернёмся к классификации с максимальным отступом. Пусть $\mathcal{O} = \{O_1, \ldots, O_n\}$ — множество объектов, $\Phi\colon \mathcal{O} \to \mathbb{R}^d$ — отображение в пространство признаков (feature map), $x_i = \Phi(O_i)$. Класс объекта задаётся (неизвестной нам) функцией $f\colon \mathcal{O} \to \{-1, 1\}$: $t_i = f(O_i)$. Обозначим через $\mathrm{margin}_\Phi(f)$ величину максимального отступа:

$$\begin{cases} \min\limits_i |\langle b, x_i\rangle| \to \max, \\ t_i = \mathrm{sign}\langle b, x_i\rangle, \quad i = 1, \ldots, n,\ |b| = 1. \end{cases}$$

Теперь предположим, есть целый класс возможных функций: $\mathcal{F} = \{f_1, \ldots, f_m\}$. Минимальный отступ

$$\mathrm{margin}_\Phi(\mathcal{F}) = \min(\mathrm{margin}_\Phi(f_1), \ldots, \mathrm{margin}_\Phi(f_m))$$

характеризует сложность задачи классификации произвольной функции из $\mathcal{F}$.

Классификация с максимальным отступом

Если теперь минимизировать $\mathrm{margin}_\Phi(\mathcal{F})$ по всевозможным отображениям $\Phi\colon \mathcal{O} \to S^{d-1}$ (нормируем признаки):

$$\begin{cases} \min_{i,j} |\langle b^j, x_i\rangle| \to \max, \\ t_i^j = \mathrm{sign}\langle b^j, x_i\rangle, \quad i = 1, \dots, n,\ j = 1, \dots, m, \\ |b^j| = 1, |x_i| = 1, \end{cases}$$

мы, получим в точности величину $\mathrm{margin}(S)$ сигнум-матрицы $S = \{f_j(O_i)\}_{\substack{i=1,\dots,n \\ j=1,\dots,m}}$.

Таблица различных мер сложности

Напомним определения:

$$\mathrm{margin}(S) := \max\{\min_{i,j} \frac{|\langle x_i, y_j\rangle|}{|x_i| \cdot |y_j|} : \ \mathrm{sign}\langle x_i, y_j\rangle = S_{i,j}\},$$

$$\mathrm{mc}(S) := \mathrm{margin}^{-1}(S).$$

	равенство	**знак**
размерность	rank	$\mathrm{rank}_{\pm}$
длина	γ_2	mc

Величина $\log \mathrm{rank}_{\pm}$ эквивалентна коммуникационной сложности в вероятностной модели с неограниченной ошибкой.

Величина γ_2 возникла в функциональном анализе (факторизация операторов через гильбертовы пространства).

Theorem (Forster, 2002)

Для $S \in \{-1, 1\}^{m \times n}$ *имеет место неравенство*

$$\mathrm{mc}(S) \geqslant \frac{\sqrt{mn}}{\|S\|_{2\to 2}}.$$

Пусть $\{x_i\}$, $\{y_j\}$ — реализация S с помощью единичных векторов. Рассмотрим величину

$$D := \sum_{j=1}^{n} (\sum_{i=1}^{m} |\langle x_i, y_j \rangle|)^2.$$

Ранее было доказано (лекция №2), что $D \leqslant m\|S\|_{2\to 2}^2$.
С другой стороны, если это реализация с максимальным отступом, то $|\langle x_i, y_j \rangle| \geqslant \mathrm{margin}(S)$, поэтому $D \geqslant nm^2 \,\mathrm{margin}^2(S)$. Отсюда

$$nm^2 \,\mathrm{margin}^2(S) \leqslant m\|S\|_{2\to 2}^2, \quad \text{ч.т.д.}$$

Пусть H — матрица Адамара, т.е. $n \times n$ сигнум-матрица с ортогональными строками. Чему равно $\mathrm{mc}(H)$? Воспользуемся теоремой Форстера. Оценка снизу: $n^{1/2}$, сверху тоже $n^{1/2}$, следовательно, $\mathrm{mc}(H) = n^{1/2}$.

Связь γ_2 и mc

Утверждение

$$\mathrm{mc}(S) = \min\{\gamma_2(A)\colon A_{i,j}S_{i,j} \geqslant 1\ \forall i,j\}.$$

Доказательство. Пусть есть реализация $S_{i,j} = \operatorname{sign}\langle x_i, y_j\rangle$ с векторами $|x_i| = 1$, $|y_j| = 1$ и максимальным отступом $\operatorname{margin}(S) = \min|\langle x_i, y_j\rangle|$. Матрица $A_1 = (\langle x_i, y_j\rangle)$ имеет $\gamma_2(A_1) \leqslant 1$. Матрица $A = A_1/\operatorname{margin}(S)$ имеет $\gamma_2(A) \leqslant \mathrm{mc}(S)$, при этом

$$A_{i,j}S_{i,j} = \frac{1}{\operatorname{margin}(S)}\langle x_i, y_j\rangle \operatorname{sign}\langle x_i, y_j\rangle \geqslant 1.$$

Мы доказали оценку $\min\gamma_2(A) \leqslant \mathrm{mc}(S)$.
Обратно, пусть $A_{i,j}S_{i,j} \geqslant 1$. Запишем $A_{i,j} = \langle x_i, y_j\rangle$ и $|x_i|\cdot|y_j| \leqslant \gamma_2(A)$. Тогда $\operatorname{sign}\langle x_i, y_j\rangle = S_{i,j}$ и

$$\min\frac{|\langle x_i, y_j\rangle|}{|x_i|\cdot|y_j|} \geqslant \frac{|A_{i,j}|}{\gamma_2(A)} \geqslant \gamma_2^{-1}(A).$$

Следовательно, $\operatorname{margin}(S) \geqslant \gamma_2^{-1}(A)$, т.е. $\mathrm{mc}(S) \leqslant \gamma_2(A)$.

Утверждение

$\mathrm{mc}(S) \geqslant mn/\gamma_2^*(S)$.

Пусть $\mathrm{mc}(S) = \gamma_2(A)$, $A_{i,j}S_{i,j} \geqslant 1$.
Нормируем $A' = A/\gamma_2(A)$, тогда

$$\gamma_2^*(S) \geqslant \langle S, A' \rangle = \sum S_{i,j}A'_{i,j} \geqslant \gamma_2(A)^{-1} = \mathrm{mc}(S)^{-1}.$$

Ч.т.д. Оказывается, эта оценка усиливает теорему Форстера ($\mathrm{mc}(S) \geqslant \sqrt{mn}/\|S\|_{2\to 2}$), поскольку $\gamma_2^*(S) \leqslant \sqrt{mn}\|S\|_{2\to 2}$ (**Упражнение**).

Мы доказали, что величины mc, γ_2 и rank связаны следующим образом:

$$\mathrm{mc}(S) \leqslant \gamma_2(S) \leqslant \sqrt{\mathrm{rank}(S)}.$$

Для матриц Адамара достигается равенство.

Установим связь между mc и $\mathrm{rank}_{\pm}$.

Утверждение

$$\mathrm{rank}_{\pm}(S) \leqslant C\, \mathrm{mc}(S)^2 \log(m+n),$$

где C — абсолютная постоянная.

Johnson–Lindenstrauss

Нам потребуется лемма Johnson–Lindenstrauss.

Утверждение

Пусть R — матрица $d \times N$ со стандартными гауссовыми элементами, т.е. $R_{i,j} \sim \mathcal{N}(0,1)$. Тогда для любых векторов $x, y \in \mathbb{R}^N$, $|x|, |y| \leqslant 1$ и $\varepsilon \in (0, 1/2)$ имеем

$$\mathrm{P}(|\langle \frac{1}{\sqrt{d}} Rx, \frac{1}{\sqrt{d}} Ry \rangle - \langle x, y \rangle| \geqslant \varepsilon) \leqslant 2 \exp(-d\varepsilon^2/8).$$

Отметим, что от размерности N ничего не зависит.
Следствие (Johnson–Lindenstrauss): M точек в единичном шаре можно (линейно) отобразить в пространство размерности $d \asymp \log(M)\varepsilon^{-2}$, изменив попарные расстояние не более чем на ε. Действительно, нам требуется сохранить M^2 скалярных произведений; если $2M^2 \exp(-d\varepsilon^2/8) < 1$, то случайная матрица делает это с положительной вероятностью.

Докажем теперь оценку $\mathrm{rank}_{\pm}(S) \ll \mathrm{mc}(S)^2 \log(m+n)$.

Возьмём реализацию $S_{i,j} = \mathrm{sign}\langle x_i, y_j\rangle$ с $|x_i| = |y_j| = 1$ и максимальным отступом $\mathrm{margin}(S) = \min|\langle x_i, y_j\rangle|$. Применим лемму Johnson–Lindenstrauss:

$$\mathsf{P}(|\langle x_i', y_j'\rangle - \langle x_i, y_j\rangle| \geqslant \varepsilon) \leqslant 2\exp(-d\varepsilon^2/8). \qquad (*)$$

Если $2mn\exp(-d\varepsilon^2/8) < 1$, то найдутся вектора $\{x_i'\}, \{y_i'\}$ в d-мерном пространстве, такие что (*) выполнено для всех $i = 1, \dots, m$, $j = 1, \dots, n$.
При этом, если $\varepsilon < \mathrm{margin}(S)$, например, $\varepsilon = \frac{1}{2}\mathrm{margin}(S)$, то знак $\langle x_i, y_j\rangle$ не поменяется: $\mathrm{sign}\langle x_i', y_j'\rangle = S_{i,j}$. Тогда $\mathrm{rank}_{\pm}(S) \leqslant d$. Условие на d:

$$d > 8\varepsilon^{-2}\ln(2mn) \asymp \mathrm{mc}(S)^2 \log(2mn).$$

Дискрепанс

Напомним понятие дискрепанса матрицы $S \in \{-1, 1\}^{m \times n}$.

Пусть μ — мера на множестве индексов $[m] \times [n]$. Для $R \subset [m] \times [n]$ обозначим через R_+ множество индексов $(i, j) \in R$ в которых $S_{i,j} = 1$, и R_- соответственно. Тогда

$$\mathrm{disc}_\mu(S) := \max_R |\mu(R_+) - \mu(R_-)|,$$

где максимум берётся по всевозможным комбинаторным прямоугольникам $R = R' \times R'' \subset [m] \times [n]$. Положим также $\mathrm{disc}(S) := \min_\mu \mathrm{disc}_\mu(S)$, минимум берётся по всем *вероятностным* распределениям μ.
В лекции №1 мы установили неравенство $C(f) \geqslant \log_2(1/\,\mathrm{disc}(f))$ для коммуникационной сложности в детерминированной модели ($\chi \leqslant 2^C$, $\mathrm{disc} \geqslant 1/\chi$).

Дискрепанс и CUT-норма

В случае равномерной меры u величина $\mathrm{disc}_u(S)$ выражается следующим образом:

$$mn\,\mathrm{disc}_u(S) = \max_{R',R''} \Big| \sum_{i\in R',j\in R''} S_{i,j} \Big|.$$

Величина в правой части равенства называется также CUT-нормой: $\|S\|_{\mathrm{CUT}}$. CUT — т.к. мы “вырезаем” из матрицы прямоугольник и суммируем элементы. Можно считать её дискрепансом по отношению к считающей мере $\mu_{i,j} \equiv 1$. Кроме того, CUT-норма определена для всех вещественных матриц (и является нормой).

Комбинаторный дискрепанс

Ранее нам уже встречался дискрепанс матрицы M вида $\mathrm{disc}_{old}(M) = \min_{x\in\{-1,1\}^n} \|Mx\|_\infty$. В чём связь между этими понятиями?

$\mathrm{disc}_u(S)$ и $\mathrm{disc}_{old}(M)$ происходят из общего понятия — *комбинаторного дискрепанса*.

Пусть Ω — множество и $\mathcal{A}$ — некоторое семейство его подмножеств. Задача: раскрасить Ω в два цвета так, чтобы в каждом $A \in \mathcal{A}$ было примерно поровну точек обоих цветов. Раскраска $\chi\colon \Omega \to \{-1,1\}$ имеет дискрепанс

$$\mathrm{disc}(\chi,\mathcal{A}) = \max_{A\in\mathcal{A}} \Big|\sum_{x\in A} \chi(x)\Big|.$$

Дискрепанс семейства $\mathcal{A}$ определяется как $\min_\chi \mathrm{disc}(\chi,\mathcal{A})$.

$\mathrm{disc}_u(S)$ это дискрепанс для множества $\Omega = [m]\times[n]$, семейства $\mathcal{R} = \{R'\times R''\}$ комбинаторных прямоугольников и $\chi = S$.

$\mathrm{disc}_{old}(S)$ это дискрепанс семейства множеств $A_i \subset \Omega = [n]$, задаваемых строками S_i (т.е. $j \in A_i$, если $S_{i,j} = 1$).

Дискрепанс и mc

Theorem (Linial, Shraibman, 2008)

$$\frac{1}{8}\operatorname{margin}(S) \leqslant \operatorname{disc}(S) \leqslant 8\operatorname{margin}(S).$$

Доказательство.

1 шаг. Заменим в определении margin произвольные вектора $\{x_i\}, \{y_j\} \in \mathbb{R}^N$ на знаковые $\{x_i\}, \{y_j\} \in \{-1, 1\}^N$. Получится величина $\operatorname{margin}_\pm(S) \leqslant \operatorname{margin}(S)$.

Мы докажем (используя неравенство Гротендика), что эти величины эквивалентны:

$$\operatorname{margin}_\pm(S) \leqslant \operatorname{margin}(S) \leqslant K_G \operatorname{margin}_\pm(S).$$

Левое неравенство очевидно (строже требования к x_i, y_j, меньше отступ).

Доказательство disc $\asymp$ margin (продолжение)

Рассмотрим матрицу $B = (\frac{\langle x_i, y_j \rangle}{|x_i||y_j|})$, где $x_i, y_j \in \{-1,1\}^N$. Имеем $B_{i,j} = N^{-1} \sum_{p=1}^{N} x_{i,p} y_{j,p}$. Тем самым, B есть выпуклая комбинация одноранговых сигнум матриц $B_p = (x_{i,p} y_{j,p})_{i,j}$.

Обочначим через $M_\pm$ выпуклую оболочку всех одноранговых $m \times n$ сигнум-матриц. Тогда $B \in M_\pm$. Обратно, любую матрицу из выпуклой оболочки можно приблизить такой матрицей B (с коэффициентами $1/N$).

Следовательно,

$$\mathrm{margin}_\pm(S) = \max_{B \in M_\pm} \min_{i,j} S_{i,j} B_{i,j}.$$

По определению,

$$\mathrm{margin}(S) = \max_{\gamma_2(A) \leqslant 1} \min S_{i,j} A_{i,j}.$$

Из неравенства Гротендика мы вывели, что $\gamma_2^*(A) \leqslant K_G \|A\|_{\infty \to 1}$. Следовательно, для сопряжённой нормы $\gamma_2(A) \geqslant K_G^{-1} \|A\|_\nu$. Нетрудно проверить, что $M_\pm$ это шар в норме $\|A\|_\nu$, сопряженной к $\|A\|_{\infty \to 1}$.

Доказательство disc $\asymp$ margin (продолжение)

2 шаг. Докажем, что $\mathrm{disc}_\mu(S) \leqslant \|S \circ \mu\|_{\infty\to 1} \leqslant 4\,\mathrm{disc}_\mu(S)$.
Здесь $(S \circ \mu)_{i,j} = S_{i,j}\mu_{i,j}$. Ясно, что неравенство сводится к матрице $T = (S \circ \mu)$.
Имеем $\mathrm{disc}_\mu(S) = \|T\|_{\mathrm{CUT}}$, нужно доказать:

$$\|T\|_{\mathrm{CUT}} \leqslant \|T\|_{\infty\to 1} \leqslant 4\|T\|_{\mathrm{CUT}}.$$

Левое неравенство: $\sum_{i\in R',j\in R''} T_{i,j} = \langle T\mathbf{1}_{R''}, \mathbf{1}_{R'}\rangle \leqslant \|T\|_{\infty\to 1}$. Правое неравенство: для любых $x_i, y_j \in \{-1, 1\}$ имеем

$$\sum t_{i,j}x_iy_j = \sum_{\substack{x_i=1\\ y_j=1}} t_{i,j} - \sum_{\substack{x_i=1\\ y_j=-1}} t_{i,j} - \sum_{\substack{x_i=-1\\ y_j=1}} t_{i,j} + \sum_{\substack{x_i=-1\\ y_j=-1}} t_{i,j}.$$

Ясно, что эта величина не превосходит $4\|T\|_{\mathrm{CUT}}$.

Доказательство disc $\asymp$ margin (продолжение)

Итак, $\mathrm{disc}_\mu(S) \leqslant \|S \circ \mu\|_{\infty\to 1} \leqslant 4\,\mathrm{disc}_\mu(S)$, следовательно,

$$\mathrm{disc}(S) \leqslant \inf_\mu \|S \circ \mu\|_{\infty\to 1} \leqslant 4\,\mathrm{disc}(S).$$

3 шаг. Остаётся доказать, что

$$\mathrm{margin}_\pm(S) = \inf_\mu \|S \circ \mu\|_{\infty\to 1}.$$

Мы выяснили, что $\mathrm{margin}_\pm(S) = \max_{B\in M_\pm} \min_{i,j} S_{i,j}B_{i,j}$. Множество $M_\pm$ это многогранник, вершины которого – одноранговые сингум-матрицы. Будем обозначать такие матрицы как X^q, $q \in Q$. Таким образом, $B = \sum_{q\in Q} \lambda_q X^q$, где $\sum \lambda_q = 1$, $\lambda_q \geqslant 0$.

Обозначим $\delta := \min_{i,j} S_{i,j}B_{i,j}$. Нам нужно максимизировать δ при условии $S_{i,j} \sum_q \lambda_q X^q_{i,j} \geqslant \delta$.

Доказательство disc $\asymp$ margin (окончание)

Мы приходим к задаче линейного программирования:

$$\begin{cases} \delta \to \max, \\ \sum_q \ \lambda_q (X^q \circ S)_{i,j} \geqslant \delta, \\ \sum_q \lambda_q = 1, \ \lambda_q \geqslant 0. \end{cases}$$

Упражнение: найдите двойственную задачу и докажите, что её значение равно $\|S\|_{\infty \to 1}$.

- N. Linial, S. Mendelson, G. Schechtman, A. Shraibman, "COMPLEXITY MEASURES OF SIGN MATRICES" (2007).
- S.V. Lokam, "Complexity Lower Bounds using Linear Algebra" (2009).