Topologická analýza dat

Anotace

Topologická analýza dat (zkráceně TDA) je jedno z nejnovějších aplikací současné matematiky, jehož počátky se datují k začátku 21. století. Ačkoli ještě není využití TDA příliš prozkoumané, už nyní má zajímavé výsledky na poli medicíny, kdy metody TDA klasifikovaly několik nových druhů rakoviny prsu, či strojového učení, kdy se pomocí těchto nástrojů snižuje komplexita vstupu do neuronové sítě při zachování důležitých informací. V přednášce se naučíme základní koncept za těmito metodami a ukážeme si, jak je aplikovat tak, aby nám řekly něco užitečného.

Motivace

Na počátku dostaneme velkou spostu dat a chtěli bychom zjistit, co ty údaje mají společného. Jestli data tvoří dvojice čísel, pak si je můžeme snadno zobrazit jako dvojrozměrné pole. Při troše štěstí si můžeme povšimnout nějakého vzoru, například, že formují útvar ve tvaru osmičky. Avšak, pokud mají data 42 rozměrů, rozpoznat, jaký útvar tvoří pouhým pohledem, je nemožné. Pro tento účel si vybudujeme teorii TDA.

Simpliciální komplexy

Abstraktní simpliciální komplex je přirozené rozšíření pojmu graf. Bude se jednat o dvojici $\mathcal{K} = (X, F)$, kde X je množina vrcholů a F je množina stěn, tj. podmnožin X takových, že pokud $\sigma \in F$ a $\tau \subset \sigma$, pak $\tau \in F$.

Příklad: Plný trojúhelník, tetrahedron, platónská tělesa, hyperkrychle, ...

Dimenzí simpliciální stěny $\sigma \in F$ nazýváme hodnotu

$$\dim(\sigma) = |\sigma| - 1,$$

tedy počet vrcholů mínus jedna.
Dimenzí komplexu $\mathcal{K}$ pak rozumíme

$$\dim(\mathcal{K}) = \max_{\sigma \in F} \dim(\sigma).$$

Příklad:
- Vrchol má dimenzi 0 - Hrana má dimenzi 1 - Trojúhelník má dimenzi 2 - Čtyřstěn má dimenzi 3.

Subkomplexem $\mathcal{L} \subseteq \mathcal{K}$ rozumíme simpliciální komplex, jehož vrcholy jsou podmnožinou vrcholů $\mathcal{K}$ a jeho stěny jsou podmnožinou stěn $\mathcal{K}$. Formálně:

$$\mathcal{L} = (Y, G), \quad Y \subseteq X,\ G \subseteq F.$$

Uzávěr stěny $\sigma \in \mathcal{K}$ je definován jako

$$\overline{\sigma} = \{\tau \in \mathcal{K} \mid \tau \subseteq \sigma\},$$

tedy množina všech jejích podstěn. Uzávěr komplexu $\mathcal{K}$ je sjednocení uzávěrů všech jeho stěn.

Lidsky, vybrané stěny uzavřeme do komplexu jako do klícky, ve které žije. Zároveň je ta klícka co nejmenší.

Hvězdou stěny σ rozumíme podkomplex:

$$\mathrm{st}(\sigma) = \{\tau \in \mathcal{K} \mid \sigma \subseteq \tau\},$$

tedy všechny simpliciální stěny, které obsahují σ.

Intuitivně jde o „okoli“ σ v rámci komplexu. Tedy např. když vezmeme vrchol, tak si vezmeme všechno, co ho obsahuje.

Geometrickou realizací komplexu $|\mathcal{K}|$ rozumíme přiřazení takové, že každému vrcholu přiřadíme bod v $\mathbb{R}^n$ a pro každou stěnu $\tau \in \mathcal{K}$ vezmeme její **konvexní obal** v $\mathbb{R}^n$. Platí, že průnik každých dvou stěn je opět stěna našeho komplexu nebo prázdná množina.

Příklad:
Komplex tří vrcholů $\{a, b, c\}$, kde stěna je každá podmnožina, má např. realizaci jako konvexní obal tří afinně nezávislých bodů.

Konstrukce komplexů

Říkáme, že je množina $A \subset \mathbb{R}^n$ *konvexní*, pokud splňuje, že pro každé dva body $a, b \in A$ a $\lambda \in [0, 1]$ platí $\lambda a + (1 - \lambda)b \in A$.

Nyní si definujeme, co je to *nerv* nějakého topologického prostoru. Netrapme se tím, co je topologický prostor, bude to krabička, kterou nepoužijeme. Budou nám stačit otevřené koule a podmnožiny $\mathbb{R}^n$. Vezmeme $K \subset \mathbb{R}^\kappa$ omezená množinu, kterou pokryjeme koulemi o libovolných nenulových poloměrech, tj sjednocení koulí je rovno K. Poté nerv je simpliciální komplex takový, že vrcholy jsou středy těch koulí a uděláme stěnu z těch středů, pokud všechny dané středy mají společný neprázdný průnik. Pro soubor soubor koulí $\mathcal{U}$ značíme nerv jako $\mathcal{N}(\mathcal{U})$.

Věta o nervu (Čech, 1932):
Nechť $\mathcal{U}$ je soubor otevřených koulí v $\mathbb{R}^n$. Pak platí, že $|\mathcal{N}(\mathcal{U})| \simeq \bigcup \mathcal{U}$.

Pro množinu bodů $P \subset \mathbb{R}^n$ a poloměr $r > 0$ definujeme **Čechův komplex** $\mathcal{C}_r(P)$ takto:
- Každý bod $p \in P$ je vrchol.
- Podmnožina $\sigma \subseteq P$ je stěna právě tehdy, když

$$\bigcap_{p \in \sigma} B(p, r) \neq \emptyset,$$

kde $B(p, r)$ je otevřená koule.

Pro množinu bodů P a poloměr $r > 0$ definujeme **Vietoriho žebra** $\mathcal{R}_r(P)$:
- Vrcholy jsou body P.
- Podmnožina $\sigma \subseteq P$ je stěna, pokud pro všechna $p, q \in \sigma$ platí:

$$\|p - q\| \leq r.$$

$\mathcal{R}_r(P)$ je tedy jednodušší na výpočet než Čechův komplex, protože kontrolujeme pouze dvojice, ne všechny průniky.

Deloneho komplex (Delaunayho triangulace) je definován pro množinu bodů $P \subset \mathbb{R}^n$ takto: simpliciální stěna $\sigma \subseteq P$ je v komplexu právě tehdy, když existuje koule, která obsahuje všechny body σ na svém povrchu a žádný jiný bod z P uvnitř. Ten zkonstruujeme tak, že každý bod $\mathbb{R}^n$ přiřadíme k nejbližšímu datu, které máme.

Alpha komplex $\mathcal{A}_\alpha(P)$ je podkomplexem Deloneho komplexu, který se získá filtrováním podle poloměru α. Obsahuje právě ty stěny, jejichž kružnice nebo koule mají poloměr $\leq \alpha$.

Topologické vlastnosti

Eulerova charakteristika $\chi(\mathcal{K})$ simpliciálního komplexu je definována jako:

$$\chi(\mathcal{K}) = \sum_{k=0}^{\dim \mathcal{K}} (-1)^k \cdot f_k,$$

kde f_k je počet k-dimenzionálních stěn.

Když se omezíme na maximálně 2-dimenzionální objekty, dostaneme klasickou eulerovu charakteristiku z grafů.

Pro simpliciální komplex $\mathcal{K}$ definujeme **řetězové vektorové prostory** C_k jako vektorové prostory nad oblíbeným tělesem T formálně generované všemi k-simplicemi.

Definujeme hranový operátor

$$\partial_k : C_k \to C_{k-1},$$

který z k-stěny vymaže vždy jeden vrchol a střídá znaménka. Konkrétně si to definujeme následovně:

$$\partial_k = \sum_{i=1}^{k} (-1)^i (v_0, \ldots, \hat{v_i}, \ldots, v_k).$$

Pak k-tá homologie je:

$$H_k(\mathcal{K}) = \ker(\partial_k) / \operatorname{im}(\partial_{k+1}).$$

Bettiho čísla β_k jsou dimenze homologií:

$$\beta_k = \dim H_k(\mathcal{K}).$$

- β_0 = počet souvislých komponent.
- β_1 = počet „děr“ nebo smyček.
- β_2 = počet „dutých“ prostorů (např. uvnitř koule).