VC-dimenze

Zdeněk Dvořák

18. listopadu 2019

Definice 1. *Nechť $\mathcal{F}$ je systém množin. Pro množinu X definujme $X \cap \mathcal{F}$ jako systém množin $\{F \cap X : F \in \mathcal{F}\}$, a říkejme, že $\mathcal{F}$* rozbíjí *X, jestliže $X \cap \mathcal{F}$ obsahuje všechny podmnožiny X.* Vapnik-Chervonenkisova dimenze *(*VC-dimenze*) systému $\mathcal{F}$ je maximální k takové, že existuje množina velikosti k, kterou $\mathcal{F}$ rozbíjí.*

Příklady:

- Systém všech polorovin v rovině má VC-dimenzi 3. Snadno ověříme, že existuje 3-bodová množina, kterou lze rozbít (stačí, aby její prvky nebyly na společné přímce). Naopak, uvažme libovolnou 4-prvkovou množinu X. Obsahuje-li bod $x \in X$, který je ve vnitřku konvexního obalu zbývajících tří bodů, pak žádná polorovina neprotíná X přesně v $X \setminus \{x\}$. Leží-li tři z bodu X na přímce v pořadí x_1, x_2, x_3, pak žádná polorovina neprotíná X v podmnožině obsahující x_1 a x_3 a neobsahující x_2. Zbývá tedy možnost, že body X jsou v konvexní pozici; nechť body X v pořadí na konvexním obalu X jsou x_1, x_2, x_3, x_4. Pak žádná polorovina neprotíná X právě v $\{x_1, x_3\}$.

- Systém všech osově orientovaných obdélníků v rovině má VC-dimenzi nejvýše 5: Uvažme libovolnou 6-prvkovou množinu X, a nechť x_1 je její nejlevější bod. Bez újmy na obecnosti můžeme předpokládat, že existuje $X' \subseteq X \setminus \{x_1\}$ tž. všechny body z X' mají alespoň tak velkou y-ovou souřadnici jako x_1 a $|X'| \geq 3$. Nechť $x_2 \in X'$ má největší y-ovou souřadnici a $x_3 \in X'$ má největší x-ovou souřadnici. Pak neexistuje osově orientovaný obdélník obsahující x_1, x_2, x_3 a neobsahující bod(y) z $X' \setminus \{x_2, x_3\}$.

- Systém všech konvexních mnohoúhelníků v rovině má nekonečnou VC-dimenzi: Jestliže X je libovolná množina bodů v konvexní poloze, pak konvexní obal $X' \subseteq X$ protíná X právě v X'.

Lemma 1. *Každý systém množin $\mathcal{F}$ rozbíjí alespoň $|\mathcal{F}|$ podmnožin $\bigcup \mathcal{F}$.*

Důkaz. Indukcí dle $|\mathcal{F}|$. Jestliže $\mathcal{F}$ je prázdný, tvrzení je triviální. Jestliže $|\mathcal{F}| = 1$, stačí si povšimnout, že $\mathcal{F}$ rozbíjí prázdnou množinu. Předpokládejme tedy, že $|\mathcal{F}| = 2$, a speciálně $\bigcup \mathcal{F} \neq \emptyset$. Zvolme prvek $c \in \bigcup \mathcal{F}$, který neleží ve všech množinách systému $\mathcal{F}$, a označme $\mathcal{F}_1 = \{F \in \mathcal{F} : c \in F\}$ a $\mathcal{F}_2 = \{F \in \mathcal{F} : c \notin F\}$. Dle volby c máme $|\mathcal{F}_1| < |\mathcal{F}|$ a $|\mathcal{F}_2| < |\mathcal{F}|$. Povšimněme si, že jestliže $c \in X$, pak ani $\mathcal{F}_1$ ani $\mathcal{F}_2$ nerozbíjí X.

Pro $i = 1, 2$ označme jako $\mathcal{R}_i$ systém podmnožin $\bigcup \mathcal{F}_i$ rozbitých $\mathcal{F}_i$; dle indukčního předpokladu máme $|\mathcal{R}_i| \geq |\mathcal{F}_i|$. Dále nechť $\mathcal{R}_3 = \{X \cup \{c\} : X \in \mathcal{R}_1 \cap \mathcal{R}_2\}$. Dle pozorování na konci minulého odstavce máme $\mathcal{R}_3 \cap \mathcal{R}_1 = \emptyset$, $\mathcal{R}_3 \cap \mathcal{R}_2 = \emptyset$ a $|\mathcal{R}_3| = |\mathcal{R}_1 \cap \mathcal{R}_2|$. Navíc, jestliže $X' \in \mathcal{R}_3$, pak $X' \cap \mathcal{F}_1$ obsahuje všechny podmnžiny X' obsahující c a $X' \cap \mathcal{F}_2$ obsahuje všechny podmnžiny X' neobsahující c, a proto $\mathcal{F}$ rozbíjí X'. Každá množina, kterou rozbije $\mathcal{F}_1$ nebo $\mathcal{F}_2$, je také rozbita $\mathcal{F}$. Počet množin rozbitých $\mathcal{F}$ je tedy alespoň $|\mathcal{R}_1 \cup \mathcal{R}_2| + |\mathcal{R}_3| = (|\mathcal{R}_1| + |\mathcal{R}_2| - |\mathcal{R}_1 \cap \mathcal{R}_2|) + |\mathcal{R}_1 \cap \mathcal{R}_2| \geq |\mathcal{F}_1| + |\mathcal{F}_2| = |\mathcal{F}|$. □

Důsledek 2. *Má-li systém množin $\mathcal{F}$ VC-dimezi nejvýše k, pak pro každou množinu X platí*

$$|X \cap \mathcal{F}| \leq \sum_{i=0}^{k} \binom{|X|}{i},$$

a tedy je-li $k, |X| \geq 2$, pak $|X \cap \mathcal{F}| \leq |X|^k$.

Důkaz. VC-dimenze $X \cap \mathcal{F}$ je menší nebo rovna VC-dimenzi $\mathcal{F}$, a je tedy nejvýše k. Tedy $X \cap \mathcal{F}$ může rozbíjet pouze množiny velikosti nejvýše k, kterých je $\sum_{i=0}^{k} \binom{|X|}{i}$. Dle Lemma 1 dostáváme požadovaný odhad na $|X \cap \mathcal{F}|$. □

Definice 2. *Nechť μ je míra, Y je měřitelná množina konečné míry a $\mathcal{F}$ je systém měřitelných množin. Nechť $\varepsilon > 0$ je reálné číslo. Pak $N \subseteq Y$ je ε-síť, jestliže pro každé $F \in \mathcal{F}$ tž. $\mu(F \cap Y) \geq \varepsilon\mu(Y)$ platí $F \cap N \neq \emptyset$.*

Příklad: Nechť Y je osově orientovaný čtverec o hraně délky 1 a $\mathcal{F}$ je systém všech osově orientovaných obdélníků obsažených v F. Pak N je ε-síť, jestliže N protíná každý obdélník $D \subseteq Y$ obsahu alespoň ε. Povšimněme si, že obdélník obsahu alespoň ε obsažený v F musí mít obě hrany délky alespoň ε. Jako N tedy lze zvolit pravidelnou síť bodů ve vzdálenosi ε od sebe; pak máme $|N| = \varepsilon^{-2}$. Jak uvidíme ve Větě 4, existuje i asymptoticky menší ε-síť.

Bez důkazu využijeme následující vlastnost binomického rozdělení.

Lemma 3. *Nechť $0 < p \leq p_1 \leq 1$ jsou reálná čísla. Nechť $X_1, \dots, X_t$ jsou nezávislé náhodné proměnné, každá z nich 1 s pravděpodobností p_1 a 0 jinak. Pak $Pr[X_1 + \ldots + X_t \geq \lfloor pt \rfloor]$ je alespoň 1/2.*

Věta 4. *Nechť μ je míra, Y je měřitelná množina konečné míry a $\mathcal{F}$ je systém měřitelných množin VC-dimenze nejvýše $k \geq 2$. Nechť $0 < \varepsilon \leq 1$ je reálné číslo tž. $k/\varepsilon \geq 15000$. Nechť N je množina $\lceil 3\frac{k}{\varepsilon} \log \frac{k}{\varepsilon} \rceil$ bodů, zvolených nezávisle z pravděpodobnostního rozdělení π na Y definovaného vztahem $\pi(X) = \mu(X)/\mu(Y)$ pro každou měřitelnou $X \subseteq Y$. Pak s nenulovou pravděpodobností je N ε-síť.*

Důkaz. Bez újmy na obecnosti můžeme předpokládat, že $\mu(Y) = 1$ (a tedy $\pi = \mu$) a že všechny prvky $\mathcal{F}$ jsou podmnožiny Y míry alespoň ε (jinak nahraďme $\mathcal{F}$ systémem $Y \cap \mathcal{F}$ a zahoďme z něj množiny míry menší než ε, které nemají vliv na to, zda N je ε-síť). Nechť $t = \lceil 3\frac{k}{\varepsilon} \log \frac{k}{\varepsilon} \rceil$ a nechť x_1, …, x_t jsou body zvolené při volbě N (hodnoty se mohou opakovat). Zvolme nezávisle ze stejného rozdělení dalších t bodů $y_1, \dots, y_t$ (opět s opakováními). Nechť M je multimnožina bodů $y_1, \dots, y_t$.

Nechť $p \geq \varepsilon$ je infimum z $\mu(F)$ pro $F \in \mathcal{F}$ a $m = \lfloor pt \rfloor$. Pro každé $F \in \mathcal{F}$ má každý bod $y_1, \dots, y_t$ pravděpodobnost alespoň p, že bude patřit do F, z Lemma 3 je tedy $|F \cap M| \geq m$ s pravděpodobností alespoň 1/2. Máme tedy

$$\Pr[(\exists F \in \mathcal{F}) F \cap N = \emptyset \wedge |F \cap M| \geq m] \geq \frac{1}{2} \Pr[(\exists F \in \mathcal{F}) F \cap N = \emptyset]. \quad (1)$$

Nechť Q je multimnožina $2t$ prvků z Y. Jako A_Q označme jev, že $Q = \{x_1, \dots, x_t, y_1, \dots, y_t\}$. Vyberme si libovolnou množinu F, a odhadněme podmíněnou pravděpodobnost že nastane jev $B_F \equiv F \cap N = \emptyset \wedge |F \cap M| \geq m$ za předpokladu, že platí A_Q. Nechť Q obsahuje q prvků z F. Jestliže $q < m$, pak je tato pravděpodobnost nulová. Jinak B_F nastane právě tehdy, když žádný z těchto q prvků není mezi t volbami pro $x_1, \dots, x_t$. Máme tedy

$$\begin{aligned}
\Pr[B_F | A_Q] &= \frac{(2t-q)(2t-q-1)\ldots(t+1-q)}{(2t)(2t-1)\ldots(t+1)} \\
&\leq \frac{(2t-m)(2t-1-m)\ldots(t+1-m)}{(2t)(2t-1)\ldots(t+1)} \\
&< \left(1 - \frac{m}{2t}\right)^t \leq e^{-m/2}.
\end{aligned}$$

Dle Důsledku 2 máme $|Q \cap \mathcal{F}| \leq |Q|^k = (2t)^k$. Proto

$$\Pr[(\exists F \in \mathcal{F}) B_F | A_Q] = \Pr[(\exists F' \in Q \cap \mathcal{F}) B_{F'} | A_Q] < (2t)^k e^{-m/2}.$$

Jelikož tato nerovnost platí pro každé Q, máme

$$\Pr[(\exists F \in \mathcal{F}) B_F] < (2t)^k e^{-m/2}.$$

Spolu s (1) tedy dostáváme

$$\Pr[(\exists F \in \mathcal{F}) F \cap N = \emptyset] < 2(2t)^k e^{-m/2}.$$

Jelikož $m = \lfloor pt \rfloor \geq \varepsilon t - 1$, dostáváme že množina N není ε-síť s pravděpodobností méně než

$$2(2t)^k e^{1/2 - \varepsilon t/2} = e^{\log 2 + 1/2 + k \log(2t) - \varepsilon t/2} \leq e^{k \log(4t) - \varepsilon t/2} \leq 1.$$

Proto N je s nenulovou pravděpodobností ε-síť. □

Pro systém množin $\mathcal{F}$ označme jako $\tau(\mathcal{F})$ velikost nejmenší množiny $X \subseteq \bigcup \mathcal{F}$, která protne všechny prvky $\mathcal{F}$ (tedy např. jsou-li prvky $\mathcal{F}$ hrany grafu, pak $\tau(\mathcal{F})$ je velikost nejmenšího vrcholového pokrytí). Jako $\tau^\star(\mathcal{F})$ označme zlomkovou verzi tohoto parametru, tedy minimum z

$$\sum_{v \in \bigcup \mathcal{F}} x_v$$

přes hodnoty $x_v \geq 0$ tž

$$\sum_{v \in F} x_v \geq 1$$

pro každou množinu $F \in \mathcal{F}$.

Důsledek 5. *Nechť $\mathcal{F}$ je systém množin s konečným sjednocením. Má-li $\mathcal{F}$ VC-dimenzi nejvýše $k \geq 2$, pak $\tau(\mathcal{F}) = O(k\tau^\star(\mathcal{F}) \log(k\tau^\star(\mathcal{F})))$.*

Důkaz. Označme $Y = \bigcup \mathcal{F}$ a $\tau^\star = \tau^\star(\mathcal{F})$. Nechť x_v pro $v \in Y$ jsou hodnoty proměnných v optimálním řešení lineárního programu definujícího $\tau^\star(\mathcal{F})$. Pro $X \subseteq Y$ definujme $\mu(X) = \sum_{v \in X} x_v$. Pak μ je míra na Y a $\mu(Y) = \tau^\star$. Navíc $\mu(F) \geq 1$ pro každé $F \in \mathcal{F}$. Dle Věty 4 existuje $1/\tau^\star$-síť N velikosti $O(k\tau^\star \log(k\tau^\star))$; pak N protíná všechny množiny z $\mathcal{F}$. □

Příklad: Mějme dánu množinu osově orientovaných obdélníků $\mathcal{F}$ v rovině a konečnou množinu X bodů, které protnou všechny z nich. Chceme najít nejmenší podmnožinu $S_{\mathrm{opt}} \subseteq X$, která protne všechny obdélníky z $\mathcal{F}$. Vyřešením lineárního programu a použitím Důsledku 5 pro systém $Y \cap \mathcal{F}$ dokážeme najít alespoň podmnožinu velikosti $O(|S_{\mathrm{opt}}| \log |S_{\mathrm{opt}}|)$.

Nechť G je graf, v jeho vrchol, a r přirozené číslo. Pak jako $B_G(v, r)$ označme množinu vrcholů G ve vzdálenosti nejvýše r od v, a položme $\mathcal{B}_G = \{B_G(v, r) : v \in V(G), 0 \leq r \leq |V(G)|\}$.

Lemma 6. *Jestliže G neobsahuje K_t jako minor, pak $\mathcal{B}_G$ má VC-dimenzi nejvýše $t-1$.*

Důkaz. Pro spor předpokládejme, že nějaká podmnožina $X = \{v_1, \ldots, v_t\}$ vrcholů G je rozbitá $\mathcal{B}_G$. Speciálně pro $1 \leq i < j \leq t$ existují $v_{ij} \in V(G)$ a $r_{ij} \in \{1, \ldots, |V(G)|\}$ tž. $B_G(v_{ij}, r_{ij}) \cap X = \{v_i, v_j\}$. Zvolme tyto vrcholy v_{ij} tž. r_{ij} je minimální. V G existují nejkratší cesty $P_{ij,1}$ a $P_{ij,2}$ z v_{ij} do v_i a v_j, délky nejvýše r_{ij}. Z minimality r_{ij} vyvodíme, že $P_{ij,1}$ a $P_{ij,2}$ se protínají pouze ve v_{ij}; označme jejich sjednocení P_{ij} – to je tedy cesta mezi v_i a v_j obsahující v_{ij} (ve vzdálenosti nejvýše r_{ij} od obou konců této cesty). Pro $i > j$ definujme $P_{ij} = P_{ji}$, $v_{ij} = v_{ji}$ a $r_{ij} = r_{ji}$.

Nechť se vrchol $y \neq v_{ij}$ nachází na cestě P_{ij} mezi v_i a v_{ij}. Pak tvrdíme, že $d(v_i, y) < d(v_j, y)$: kdyby $d(v_j, y) \leq d(v_i, y) = r$, pak si povšimněme, že $r < r_{ij}$: podcesta $P_{ij,1}$ z v_{ij} do v_i je nejkratší, její délka je tedy nejvýše r_{ij}, a její podcesta z y do v_i délky r je ještě kratší. Pak ale $B_G(y, r) \cap X = \{v_i, v_j\}$, ve sporu s minimalitou r_{ij}. Tedy

($\star$) všechny vrcholy před v_{ij} jsou blíže k v_i než k v_j, a symetricky vrcholy za v_{ij} jsou blíže k v_j než k v_i.

Nechť se vrchol x nachází v průniku dvou takových cest $P_{i_1j_1}$ a $P_{i_2j_2}$. Ze symetrie můžeme předpokládat $d(x, v_{i_s}) \leq d(x, v_{j_s})$ pro $s \in \{1, 2\}$, a dle ($\star$) tedy x leží na nejkratší cestě z $v_{i_sj_s}$ do v_{i_s}; proto $d(v_{i_sj_s}, x) + d(x, v_{i_s}) = d(v_{i_sj_s}, v_{i_s})$. Dále ze symetrie můžeme předpokládat, že $d(x, v_{i_1}) \leq d(x, v_{i_2})$. Trojúhelníková nerovnost pak dává

$$d(v_{i_2j_2}, v_{i_1}) \leq d(v_{i_2j_2}, x) + d(x, v_{i_1}) \leq d(v_{i_2j_2}, x) + d(x, v_{i_2}) = d(v_{i_2j_2}, v_{i_2}) \leq r_{i_2j_2}.$$

Proto $v_{i_1} \in B_G(v_{i_2j_2}, r_{i_2j_2}) \cap X = \{v_{i_2}, v_{j_2}\}$. Jestliže $d(x, v_{i_2}) < d(x, v_{j_2})$, pak $d(x, v_{i_1}) < d(x, v_{j_2})$, a tedy $i_1 \neq j_2$ a $i_1 = i_2$. Jestliže $d(x, v_{i_2}) = d(x, v_{j_2})$, pak můžeme předpokládat $i_1 = i_2$ ze symetrie. Položme $i = i_1 = i_2$. Kdyby $d(x, v_i) = d(x, v_{j_s})$ pro nějaké $s \in \{1, 2\}$, pak

$$d(v_{ij_{3-s}}, v_{j_s}) \leq d(v_{ij_{3-s}}, x) + d(x, v_{j_s}) = d(v_{ij_{3-s}}, x) + d(x, v_i) = d(v_{ij_{3-s}}, v_i) \leq r_{ij_{3-s}}$$

a dostáváme $v_{j_s} \in B_G(v_{ij_{3-s}}, r_{ij_{3-s}}) \cap X$, což je spor. Máme tedy následující:

($\star\star$) Jestliže $x \in V(P_{i_1j_1}) \cap V(P_{i_2j_2})$, pak $i_1 = i_2$ a $d(x, v_{i_s}) < d(x, v_{j_s})$ pro $s \in \{1, 2\}$.

Pro $i = 1, \ldots, t$ položme

$$X_i = \{x \in V(P_{ij}) : j \in [t] \setminus \{i\}, d(x, v_i) \leq d(x, v_j), \text{a když } d(x, v_i) = d(x, v_j), \text{ pak } i < j\}.$$

Dle ($\star$) indukují množiny X_i souvislé podgrafy v G a dle ($\star\star$) jsou množiny X_i navzájem disjunktní. Jejich kontrakcí dostaneme minor K_t v G, což je spor. □

Uvažujme následující zobecnění dominující množiny. Nechť $r : V(G) \to \mathbf{Z}_0^+$ je libovolná funkce. Pak jako $\mathrm{dom}_r(G)$ označme velikost nejmenší množiny $X \subseteq V(G)$ tž. $d(v, X) \leq r(v)$ pro každé $v \in V(G)$.

Důsledek 7. *Pro každé t existuje algoritmus s polynomiální časovou složitostí, který pro libovolný graf G neobsahující K_t jako minor a libovolnou funkci $r : V(G) \to \mathbf{Z}_0^+$ vrátí číslo d takové, že $d \leq \mathrm{dom}_r(G) = O(td \log(td))$.*

Důkaz. Uvažme systém množin $\mathcal{F} = \{B_G(v, r(v)) : v \in V(G)\}$. Máme $\mathcal{F} \subseteq \mathcal{B}_G$, a dle Lemma 6 má $\mathcal{F}$ VC-dimenzi nejvýše $t - 1$. Dále si povšimněme, že $\mathrm{dom}_r(G) = \tau(\mathcal{F})$. Dle Důsledku 5 můžeme tedy položit $d = \tau^\star(\mathcal{F})$ (d lze určit v polynomiálním čase vyřešením lineárního programu). $\square$