Rozpoznávání – zápis z přednášky
VC dimenze

Jakub Ahmadyar
ČVUT FEL

31. května 2001

Obsah

Abstrakt

Po zopakování základních pojmů ze Statistické teorie učení je nastíněn pojem VC dimenze. Následuje jeho přesná matematická definice. Význam a způsob výpočtu VC dimenze je ilustrován na příkladech. Nakonec jsou diskutovány praktické výsledky statistické teorie učení, její výhody a omezení oproti ostatním metodám rozpoznávání.

1 Úvod – Statistická teorie učení

Tento text navazuje na zápis zpřednášky číslo devět – Statistická teorie učení, vněmž je možno nalézt dostatek studijního materiálu na dané téma. Pro výklad pojmu VC dimenze je klíčová nerovnice

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h.(\log \frac{2l}{h} + 1) - \log \frac{\eta}{4}}{l}}, \qquad (1)$$

kde $R(\alpha)$ je riziko (na neznámých datech), $R_{emp}(\alpha)$ empirické riziko (na zkušebních datech), l je délka dat, $(1-\eta)$ představuje pravděpodobnost, že uvedená nerovnice bude splněna a $h$je VC dimenze množiny klasifikátorů.

2 VC dimenze

Nerovnice (1) umožňuje spravděpodobností $(1-\eta)$ stanovit horní mez strukturálního rizika jako součet R_{emp} a *VC důvěry*. Pozoruhodné je, že ktomu nejen není nutná znalost $P(x,y)$, ale dokonce nemusí být znám ani konkrétní klasifikátor $f : f(x) = y$. Tedy, vlastnosti generalizace ztrénovací množiny na neznámá data se dají obecně zkoumat pro celé třídy klasifikátorů. Lze intuitivně očekávat, že pro dané množství trénovacích dat je třeba použít klasifikátor sjakousi právě optimální schopností se datům přizpůsobit, zachytit dostatečně komplexní závislosti, ale zároveň ignorovat lokální jevy, které nemají obecný charakter. Mírou této schopnosti je zjistého úhlu pohledu právě VC dimenze. V následujícím odstavci je uvedena její matematická definice, převzatáz[1].

2.1 VC dimenze třídy konceptů

Poznámka 1 *Definice uvedené vtomto odstavci nejsou vsoučasné době součástí přednášek zpředmětu Rozpoznávání.*

Nejdříve definujeme pojmy koncept a třída konceptů a definujeme jejich vztah ke klasifikátorům zminulého odstavce.

Definice 1 *(Pro zopakování) Potenční množina množiny* X, $2^X \stackrel{def}{=} \{Z \mid Z \subset X\}$.

Definice 2 *Nechť* X *je libovolná množina. Potom* $c \subset X$ *nazveme* konceptem *(nad množinou* X*). Neprázdnou množinu* $C \subset 2^X$ *nazveme* třídou konceptů *(nad množinou* X*). Prvky této množiny* $c \in C$ *budeme nazývat koncepty třídy* C.

Jedním ze způsobů popisu konceptu je jeho zadání pomocí zobrazení. Formálně definujme pro booleovskou funkci $f$koncept c_f.

Definice 3 *Nechť* f*je zobrazení z množiny* X*do dvouprvkové množiny* $\{-1,1\}$. *Potom definujeme koncept* c_f*jako množinu* $c_f \stackrel{def}{=} \{x \in X \mid f(x) = 1\}$.

Předchozí definici lze zřejmě rozšířit zdefinice konceptu na definici celé třídy konceptů, odpovídající dané množině zobrazení.

Definice 4 *Předpokládejme, že Fje libovolná množina funkcí definovaných na množině Xsoborem hodnot $\{-1,1\}$. Pak říkáme, že F*reprezentuje třídu konceptů

$$C_f \stackrel{def}{=} \{A \subset X \mid (\exists f \in F)\,(A = c_f)\}\,.$$

Nyní se konečně dostáváme kdefinici Vapnik-Červoněnkovy dimenze, která vyjadřuje schopnost pevně daného systému množin vymezit vjiné množině libovolnou podmnožinu.

Definice 5 *Nechť Xje daná množina, $Y \subset X$, $C \subset 2^X$. Potom říkáme, že* C rozbíjí (shatters) Y, *právě když platí:*

$$(\forall z \subset Y)\,(\exists c \in C)\,(c \cap Y = z)$$

Dále definujme pojem Vapnik-Červoněnkova dimenze systému množin Cjako

$$VC_{\dim}(C) \stackrel{def}{=} \max\{|Y| \mid C \textit{ rozděluje } Y\}\,,$$

nebo jako $+\infty$ vpřípadě, že toto maximum neexistuje.

2.1.1 Definice VC dimenze – shrnutí

Každá funkce $f : X \to \{-1,1\}$ zmnožiny Fbooleovských funkcí je vzorem pro jednu podmnožinu X. Pro množinu $Y \subset X$ platí, že Frozbíjí Yprávě tehdy, když pro každou podmnožinu Ylze najít nějakou fci $f \in F$, která je jejím vzorem. VC dimenze množiny Fje definována jako maximum zhodností množin, které F rozbíjí.

2.2 Výpočet VC dimenze

Význam a způsob výpočtu VC dimenze množiny klasifikátorů půjde nejlépe ozřejmit na příkladech.

Příklad 1 *Nechť $X \in \mathbf{R}$ a množina Fje tvořena všemi funkcemi tvaru*

$$f(x,\alpha) = \Big< \begin{array}{l} +1,\ x \in (\alpha,+\infty)\,,\ \alpha \in \mathbf{R} \\ -1 \textit{ jinak} \end{array}\,.$$

Pak zřejmě neexistuje pro žádné dva body $\{\mathrm{a},\mathrm{b} \in \mathbf{R} \mid \mathrm{a} < \mathrm{b}\}$ takové α, aby interval $(\alpha,+\infty)$ zahrnoval bod a *a zároveň nezahrnoval bod* b*. VC dimenze F je tedy rovna jedné, $VC_{\dim}(F) = 1$.*

Velmi užitečná je znalost VC dimenze množiny všech orientovaných nadrovin prostoru $\mathbf{R}^n$. Vztah pro její výpočet je uveden v následujícím tvrzení.

Tvrzení 1 *Nechť $X \in \mathbf{R}^n$ a nechť F je množina všech orientovaných nadrovin vprostoru $\mathbf{R}^n$. Potom $VC_{\dim}(F) = n + 1$.*

Důkaz. Formální důkaz tohoto tvrzení lze nalézt například v[2], zde jej pouze zhruba nastíníme.

Mějme množinu m bodů (a jejich směrových vektorů) $M \subset \mathbf{R}^n$ takovou, že Frozbíjí M. Zároveň požadujme, aby hodnost Mbyla maximální možná a pokusme se tuto hodnost určit.

Začneme tím, že zjistíme, kdy můžeme dvě libovolné množiny bodů S_1a S_2rozdělit orientovanou nadrovinou procházející počátkem. Počátek soustavy souřadnic položme do bodu o. Nazvěme $\mathbf{L}_1$a $\mathbf{L}_2$konvexní obaly S_1a S_2. Vést bodem o rovinu oddělující S_1od S_2bude zřejmě možné jen tehdy, pokud $\mathbf{L}_1$a $\mathbf{L}_2$mají jediný společný bod, a to bod o. (Pro porozumění si to lze představit tak, že $\mathbf{L}_1$ a $\mathbf{L}_2$ mají tvar jakýchsi n-rozměrných jehlanovitých geometrických útvarů, dotýkajících se svými vrcholy vpočátku. Aby tyto bylo možno rozdělit, nesmí mít více než právě jeden společný bod.)

Pokud chceme být schopni rozbít množinu Morientovanými nadrovinami, umístíme jeden zjejích prvků do počátku. Všechny ostatní body (jejich směrové vektory) zM musí být navzájem lineárně nezávislé. Pak pro jakákoli dvě disjunktní S_1, S_2 zM je průnik $\mathbf{L}_1$a $\mathbf{L}_2$roven $\{o\}$. Pokud by totiž body $s \in M$ byly LZ, platilo by

$$\sum_{i=1}^{m} s_i.k_i = 0, \ \ k_i \in \mathbf{R}, \ \textstyle\sum_i k_i = 1.$$

Členy sumy se záporným k_i převedeme na levou stranu rovnice:

$$\sum_{i=1}^{m} s_i.k_i^+ = \sum_{i=1}^{m} s_i.\left|k_i^-\right|, \ \ k_i \in \mathbf{R}, \ \textstyle\sum_i k_i = 1.$$

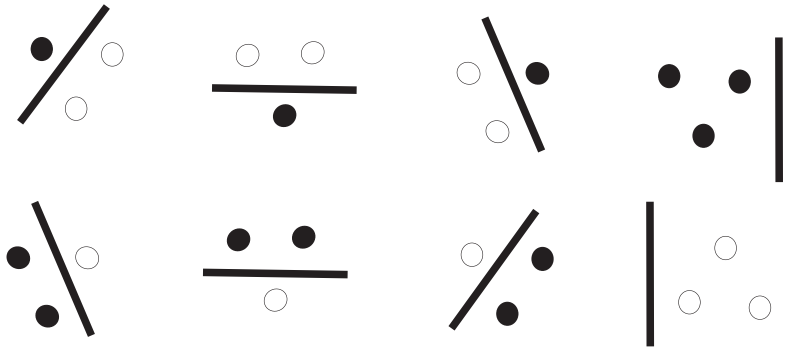
Tato rovnost říká, že konvexní obaly bodů zlevé a zpravé strany rovnice mají více než jeden společný bod, nejde je tedy rozdělit.

Lineárně nezávislých bodů v$\mathbf{R}^n$ lze nalézt maximálně n, jeden bod je vpočátku, ztoho dostáváme $VC_{\dim}(F) = hod(M) = n + 1$.

■

Příklad 2 *Předchozí tvrzení si zkusíme ověřit pro lineární prostor dimenze 2. Nechť $X \in \mathbf{R}^2$ a množina F je tvořena orientovanými přímkami. Všechny* *body na jedné straně přímky budou ohodnoceny +1 a všechny body na druhé straně -1.*

Celkem lehce můžeme najít tři body, které mohou být přímkami rozbity, ovšem rozbít přímkou body čtyři je již nemožné. Ztoho dostáváme $VC_{\dim}(F) = 3$, což souhlasí snaším očekáváním.



Příklad 3 *Zkusme určit VC dimenzi množiny klasifikátorů F, tvořené elipsami*

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = r^2.$$

Body uvnitř elipsy se ohodnotí +1, body vně -1.

Řešení. I vtomto případě se dá využít vztahu pro VC dimenzi orientovaných nadrovin. Ktomu je ovšem třeba analitický vztah elipsy linearizovat:

$$\begin{aligned}
&(bx)^2 + (ay)^2 - (abr)^2 = 0 \\
&w_1 = a^2 \\
&w_2 = a^2y^2 \\
&w_3 = b^2 \\
&w_4 = b^2x^2 \\
&w_5 = (abr)^2 \\
&w_4 + w_2 - w_5 = 0
\end{aligned}$$

Lineární nadprostor parametrů elipsy má dimenzi 5, VC dimenze množiny všech elips se středem v počátku $VC_{\dim}(F) = 6$.

■

Na závěr uvedeme učebnicový příklad třídy klasifikátorů sjediným parametrem a nekonečnou VC dimenzí.

Příklad 4 *Definujme skokovou funkci* $\theta(x)$, $x \in \mathbf{R} : \{\theta(x) = 1\ \forall x > 0;\ \theta(x) = -1\ \forall x \leq 0\}$. *Dále uvažujme množinu funkcí* $f(x, \alpha) = \theta(\sin(\alpha x));\ x, \alpha \in \mathbf{R}$. *Zvolme* n *bodů,* $x_1 = 10^{-1}$, $x_2 = 10^{-2}$, $\ldots$, $x_n = 10^{-n}$. *Pak jejich libovolné ohodnocení* $y_1, y_2, \ldots, y_n$, $y_i \in \{+1, -1\}$ *lze realizovat pomocí funkce* $f(x, \alpha)$, *kde*

$$\alpha = \pi \left(1 + \sum_{i=1}^{n} \frac{(1 - y_i)\,.10^i}{2} \right).$$

Na maximální počet bodů x_i*neklademe žádné požadavky, VC dimenze třídy klasifikátorů* $f(x, \alpha)$ *je tedy nekonečná.*

VC dimenze funkcí zminulého příkladu je nekonečná. Zároveň bychom ale dokázali najít čtyři body, které nejdou žádnou zfunkcí $f(x, \alpha)$ rozbít. Stejně tak žádná orientovaná nadrovina vn-rozměrném prostoru nerozbije tři body, ležící na jedné přímce. Dále uvidíme, že u některých tříd klasifikátorů vztah mezi VC dimenzí a jejich schopností generalizovat není tak jednoznačný, jak by se mohlo na první pohled zdát.

3 Využití Statistické teorie učení

Vztah (1) představuje prozatím jedinou možnost, jak sdanou jistotou určit horní mez rizika, když neznáme rozdělelení $P(x, y)$ vstupních dat. V praktických aplikacích je to obvyklá situace, o vztahu mezi pozorovanými příznaky a rozdělením objektů do tříd často nevíme vůbec nic. Pro daný úkol můžeme volit mezi metodami jako k-NN nebo Maximum Likelihood, a metodami založenými na minimalizaci strukturálního rizika (perceptron, SVM). Problémem klasických metod je, že sice většinou fungují, ale neposkytují žádné záruky své spolehlivosti.

Například metoda k-NN (knejbližších sousedů) může v praxi fungovat velmi dobře, přesto se vymyká předpovědím úspěšnosti na neznámých datech. Tato metoda je zajímavá i zhlediska výkladu statistické teorie učení. Platí totiž $VC_{\dim}(k\text{-}NN) = \infty$ a po učení $R_{emp}(k\text{-}NN) = 0$, což se může jevit jako příznak typický pro špatně generalizující třídu. Ve skutečnosti nekonečná VC dimenze znamená pouze to, že horní mez rizika k-NN klasifikátoru nelze pomocí nerovnice (1) stanovit.

Ostatní klasické metody jsou na tom se spolehlivostí podobně. Metody založené na odhadu parametrů rozdělení náhodné veličiny mohou při chybně zvoleném modelu divergovat. Navíc i při použití správného tvaru rozdělení je zaručena pouze asymptotická konvergence koptimu, takže po použití konečné trénovací množiny nevíme o spolehlivosti klasifikátoru opět nic. Pro zachycení vztahů ve více dimenzích nejsou tyto metody vhodné, zde jsme odkázáni na metody vzešlé ze STU, případně neuronové sítě, algoritmy UI apod.

Jak se dalo očekávat, metody minimalizující strukturální riziko mají kromě nesporných výhod také nevýhody. Na data ztrénovací množiny jsou stále kladeny požadavky známé ze statistiky: data musí reprezentovat navzájem nezávislá měření a průběh hustoty pravděpodobnosti se nesmí sčasem měnit. To jsou omezení, která pro konkrétní data těžko ověříme. Dále, pro některé třídy klasifikátorů nerovnice (1) prostě nefunguje. Kromě případů sVC_{dim} rovnou nekonečnu uvádí [2] ještě jeden zajímavý příklad.

Příklad 5 *Uvažujme klasifikátor tvořený pamětí skapacitou mpozorování. Ten je schopen mvzorkům přiřadit správnou třídu, pro všechna další pozorování bude rozhodovat stále stejně. Délka trénovací množiny lnechť je větší než m(abychom se vyhnuli triviálnímu vztahu). Pro prvních mpozorování bude empirický risk roven nule, pro ostatní $R_{emp} = 0,5$, skutečný risk $R = 0,5$ (předpokládáme náhodná data a třídy $\{+1,-1\}$ se stejným zastoupením). VC dimenze je rovna m. Dosadíme do nerovnice (1) a dostáváme*

$$\frac{m}{4l} \leq \ln\frac{2l}{m} + 1 - \frac{1}{m}\ln\frac{\eta}{4},$$

což je splněno pro všechna η, jak by ukázal další výpočet.

Nepříjemným nedostatkem metod minimalizujících strukturální riziko je také malá těsnost horní meze rizika$R$. Aby mohl řešitel problému garantovat sjistou pravděpodobností, že skutečná střední chyba se bude od chyby na trénovacích datech lišit nejvíce o stanovenou mez, musí jich požadovat mnohem více, než je ve skutečnosti zapotřebí. Přes tyto nedostatky je STU obrovským přínosem a metody na ní založené dnes prodělávají bouřlivý rozvoj.

Reference

[1] Hakl, F. a Holena, M., "Úvod do teorie neuronových sítí," *Skripta FJFI ČVUT,* 1997

[2] Burges, C. J. C., "A tutorial on SVM for pattern recognition," *Kluwer Academic Publishers, Boston (e-version)*