Strojové učení

J. Mayer, Gymnázium, Brno, Vídeňská 47, m1024@seznam.cz
M.Hynčica, Biskupské gymnázium Vansdorf, gandalph@centrum.cz
J. Novotný, Gymnázium Mikulov, novotny.jann@centrum.cz
M. Feigl, Gymnázium U Libeňského zámku, dragon_king@post.cz

Abstrakt:
Seznámení s principy algoritmů, díky nimž se počítačová schopnost uvažovat přibližuje mílovými kroky schopnostem lidským

1 Co strojové učení zahrnuje?

Strojové učení, anglicky Machine Learning (ML), patří mezi nejperspektivnější odvětví informatiky. Hlavním cílem ML je přimět počítače, aby samy získavaly zkušenosti a učily se svévolně novým věcem, stejně jako to dělají živočichové již po několik miliónů let. Zatím je tento cíl někde v nepříliš vzdálené budoucnosti, ale již v dnešní době je strojové učení hojně využívano například pro rozpoznávání řeči nebo i pro řízení virtuálních aut. Cílem tohoto projektu bude nastíněné základní problematiky a zevrubné seznámení s jednotlivými metodami.

V projektu se budeme zabývat pouze učením s učitelem to znamená, že programy umelé inteligence budou „vědět“, čeho mají dosáhnout. Např.: závodní řidič ví k čemu jsou v autě jednotlivé pedály a potřebuje se naučit, co nejrychleji dojet na určité místo. Na druhou stranu v popisu učení bez učitele by řidič teprve zjišťoval, k čemu které pedály slouží.

Základní metody strojového učení:

1. Metody hledání nejbližšího souseda - stroj se snaží najít podobné situace a chovat se v nich stejně
2. Rozhodovací stromy a lesy - stroj se znaží zachovat podle naučených podmínek
3. Genetické algoritmy - stroj vytvoří několik jedinců, kteří se vyvíjí podobně jako živočichové v přírodě
4. Neuronové sítě - stroj pracuje na podobném principu jako mozek živočichů
5. Bayesovské statistické metody - stroj se snaží daný problém klasifikovat pomocí pravděpodobnostních výpočtů (těmito se nebudeme hlouběji zabývat)

2 Popis základních algoritmů

Nejbližší soused

Metoda nejbližšího souseda se užívá při řešení klasifikačních problémů, to jest zařazení do tříd podle vlastností = atributů. Atributy hohou být dvojího typu: numerické – většinou nabývají číselných hodnot, nebo kategoriální – nabývají hodnot z konečné množiny.

Nyní si ukážeme tuto metodu, její výhody a nevýhody:

Jak to funguje:

$X=(x_1, x_2,\dots , x_n)$

X ... prvek, který chceme oklasifikovat (např. pacient)

$x_1 – x_n$ ... hodnoty atributů prvku X (např. teplota, tlak)

Abychom mohli tento prvek zařadit do třídy, musíme tuto metodu naučit jak ho zařadit.

Při učení s učitelem máme k dispozici učebmí množinu příkladů s výsledky, ta může vypadat následovně:

$L = \{(l_1,c_1), \dots, (l_m, c_m)\}$

$l_1 – l_m$ ... zařazované prvky

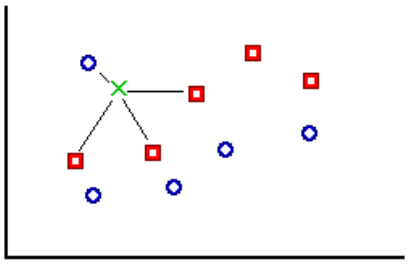
$c_1 – c_m$ ... třídy do nihž jsou prvky $l_1 – l_m$ zařazovány

L ... učební množina

Klasifikace neznámého případu probíhá tak, že k tomuto případu najdeme nejbližší známé případy z učící množiny a zařadíme ho na základě většiny.

Problémem může být otázka kolik nejbližších sousedů máme brát v úvahu, pokud jich zvolíme málo, může dojít k chybě vzhledem k možným nepřesnostem v učící množině, pokud jich naopak zvolíme příliš mnoho je prvku přiřazena třída, která je zastoupena nejhojněji.

Při správném zvolení počtu sousedů a dobré učící množině dosaheje tato metoda velice dobrých výsledků.



čtverečky	- prvky třídy 1
kolečka	- prvky třídy 2
křížek	- prvek který zařazujeme

Obrázek ukazuje zařazení do třídy podle nětšiny nejbližšíh sosedů – teď by se prvek zařadil do třídy 1

Rozhodovací stromy a lesy

Rozhodovací strom je algoritmus strojového učení, který se v praxi používá hlavně z důvodu přehlednosti a srozumitelnosti.

Rozhodovací strom se skládá z tzv. uzlů (obr. 1 – 1,2,3,4,5). Ty se dělí na rozhodovací uzly (obr. 1 – 1,2) a listy (konečný výsledek, obr. 1 – 3,4,5). Při klasifikaci dat se stromem prochází od vrchu (od tzv. kořene, obr. 1 – 1) a postupuje se směrem dolů, na každém rozhodovacím uzlu se vyhodnotí podmínka, a pokud je splněna pokračuje se vlevo jinak vpravo, dokud se nedojde k některému z listů.

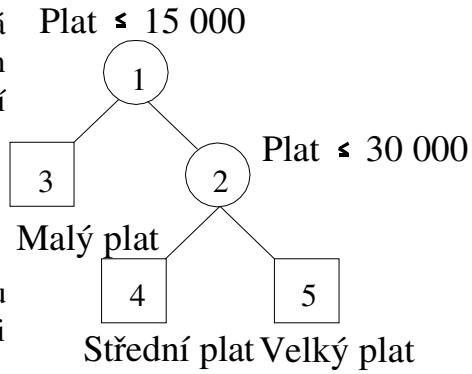
Algoritmus učení stromů je založen na rekurzivním rozdělování učící množiny:

1) Vezmi učící množinu
2) Vyber nejlepší test na základě kritéria, které minimalizuje neurčitst (Entropie, Gini index)

3) Podle testu rozděl na množiny S_1 a S_2. Pro každou S_i, ve které nejsou pouze případy, ze stejné třídy opakuj od kroku 1 a použij S_i jako učící množinu

Rozhodovací lesy

Pro zlepšení klasifikace se často používá stromů více = les. jsou to množiny různých rozhodovacích stromů, které zároveň musí obsahovat pravidlo na kombinaci výsledků.



Obr. 1 - Příklad rozhodovacího stromu

Genetické algoritmy

Genetické algoritmy řeší problémy simulovanou evolucí dle Darwinova pravidla - přežívají jen ti silnější jedinci (jen ta nejlepší řešení problému...)

Jedinci (u vážujme řešení jako jedince) mohou

- křížit se (kombinace zpravidla dvou rodičovských genů za vzniku dvou potomků)
- mutovat (samozměna)
- rekombinovat (křížení i mutace)

Pro určení, který gen je silnější a který slabší, je definován pojem „fitness“.

Záruku, že fitness bude stále hodnotnější, poskytuje selekční tlak, který nutí populaci k řešení dostačujícímu (určeno v podmínce ukončení) optimálnímu (nejlepší) maximálnímu (lokálnímu; hodnota fitness dlouhodobě stagnuje)

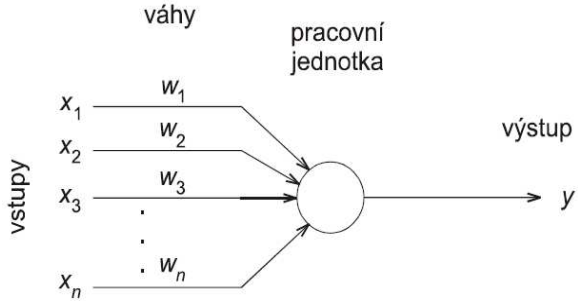
Genetické algoritmy se dají nejlépe použít v případech, kdy

1. neznáme správné vztahy (vzorce…)
2. nejsme schopni realizovat řešení v rozumném čase
3. potřebujeme řešit ve více rozměrech (samostatně kódované rozměry spojené do jednoho chromozómu)
4. stavíme neuronovou síť

Naopak selhávají v případech, kdy používáme prostory s velkým počtem lokálních maxim

Neuronové sítě

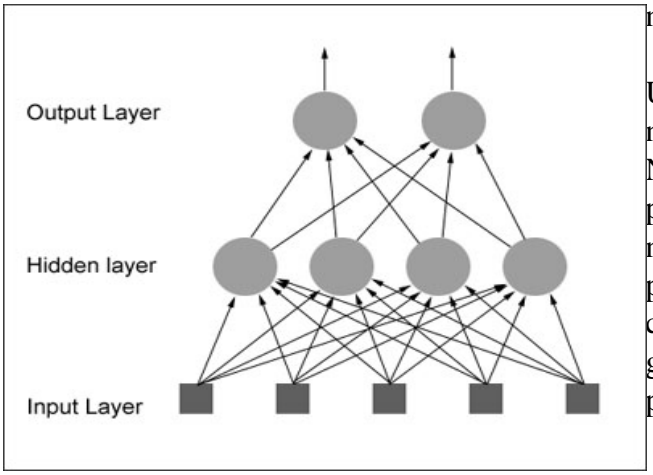
Inspirace pro tyto algoritmy pochází z živočišné nervové soustavy. Neuronová síť je soubor vzájemně provázaných malých jednotek zvaných neurony.



Každý neuron má n vstupů a jeden výstup. Každý vstup je ovážen (tzn. číslo, které příjde na vstup x je vynásobeno koeficientem w). Všechny takto vypočítané hodnoty se sečtou a předají se jako parametr signoidální funkci, jejíž výsledek putuje na výstup y. Můžeme tedy psát:

$$y = \text{sigmoid}(w_0 + x_1 * w_1 + x_2 * w_2 + ... x_n * w_n)$$

Více neuronů propojíme do neuronové sítě. Nejběžnějším typem jsou dopředné sítě, které mají jednu řadu vstupních neuronů, libovolný počet skrytých neuronů a jednu řadu výstupních neuronů.



Učení neuronové sítě probíhá tak, že nastavujeme všechna w ve všech neuronech. Na to existuje několik metod např.: back propagation (kdy postupujeme od zvrchu a nastavujeme jednotlivé váhy) a nebo pomocí genetických algoritmů (tehdy máme celou generaci neuronových sítí a do další generace vybereme ty sítě, které svoji úlohu plnili nejlépe).

3 Shrnutí

V několika odstavcích vám byly předstveny základní principy a mechanismy strojového učení. Snad se vám podařilo udělat si alespoň základní obraz o těchto komplikovaných systémech. Ať už budeme brát v úvahu jednoduché principy hledání nejbližšího souseda nebo složité neuronové sítě na rozpoznání libovolných znaků, musíme brát v úvahu, že algoritmy pro samovzdělávání počítačů jsou zde a ukrývají nezměrné možnosti.

Poděkování

Děkujeme všem organizátorům FyzTydu, zvláště pak supervizorovi Emilu Kotrčovi.

Reference:

[1] T.M. Mitchell: Machine learning

[2] M.I. Schlesinger: Deset přednášek z teorie statistického a strukturního rozpoznávání

[3] Rozhodovací stromy: http://lisp.vse.cz/%7Eberka/docs/SL-IDT.PDF

[4] Strojové učení: http://lisp.vse.cz/%7Eberka/docs/SL-ML.PDF

[5] Mat Bucklet: Neural network tutorial, http://www.ai-junkie.com/ann/evolved/nnt1.html