Strojové učení – Příprava dat, chyby v datech a bias, korelace a kauzalita

Strojové učení
Strojové účení je podoblastí oboru umělé inteligence, který se zabývá vývojem algoritmů schopných učit se z dat. Občas není možné pro daný problém vytvořit algoritmus, předpovědi, rozpoznání obličejů nebo věcí atd… Na rozdíl od tradičního programování, kde explicitně definujeme pravidla, u strojového učení necháváme samotný počítač, aby si pravidla odvodil sám na základě příkladů (data - správně i špatně)

Princip
Hlavním principem strojvého učení je hledání funkce, která mapuje vstupy a výstupy. Počítač na základě trénovacích dat snaží nalézt takovou funkci, která dobře funguje na trénovacích datech a zároveň dokáže správně předpovídat i pro nová data (generilazce)

Každý algoritmus pro strojové učení má využití a každý může být lepší na něco jiného. K nejvýznamnějším algoritmům patří Lineární regrese, Neuronová síť, Rozhodovací stromy atd…

Typy strojového učení

1. Supervised Learning

Pracuje s označenými daty, kde je znám správný výstup. Algoritmus se učí mapovat vstupy na známe výstupy. Dělí se na:
Klasifikace - zařazení do kategorií (rozpoznání objektu na obrázku)
Regrese - predikce číselné hodnoty (předpověď ceny nemovitosti)
Lineární regrese, rozhodovací stromy, neuronové sítě

2. Unsupervised Learning

Pracuje s neoznačenými daty, není definován správný výstup. Algoritmus sám hledá skryté vzory, struktury v datech. Detekce anomálií, Klastrování, Redukce

3. Reinforcement Learing

Neučí se z existujících dat, ale učí se z interakcí. Algoritmus provádí akce v prostředí a získává odměny/tresty. Učí se strategii, která maximalizuje celkovou odměnu. Robotika, autonomní vozidla, hraní her

Proces strojového učení

Základní proces zahrnuje:

1. Sběr trénovacích a testovacích dat
2. Příprava dat – čištění, transformace
3. Výběr modelu – volba vhodného algoritmu
4. Trénování modelu – učení z trénovacích dat
5. Evaluace – ověření kvality modelu na testovacích datech
6. Nasazení a monitoring – použití modelu v praxi

Příprava dat

Jedna z nejdůležitějších částí při strojovém času, also zabírá taky většinu času. Kvalita dat přímo ovlivňuje kvalitu modelu. Dobře připravená data mohou výrazně zlepšit výkon i jednoduchých modelů. Většinu času jsou data ve formátu CSV. Chyby v datech jsou velmi časté a jsou součástí přípravy dat

V pythonu si data nahráváme pomocí pandas package metody, která je uloží do DataFramu (dvourozměrné pole)
data = pandas.read_csv("source", "separator")

Typy dat
Strukturovaná data - snadno dostupná a strojově čitelná, pevná struktura, CSV, Excel
Nestruktrovaná data - nemají pevně danou strukturu, obtížnější na zpracování - speciální techniky, text, obrázky, audio, video
Semi-strukturovaná data - kombinace obou, JSON, XML, email
Big Data - Extrémně velké objemy dat

Způsoby vzniku dat
Generování - generované text, videa, fotky…
Průzkumy - uživatelská hodnocení, volební preference…
Měření a snímání - audiozáznamy, kamery, senzory…
(ručně)
Crawling stránek - procházení stránek….

Chyby v datech a bias
Reprezentativita dat
Reprezentativita dat je základní předpojladem pro kvalitní model. Sledujeme, zda naše data správně odrážejí realitu, kterou se snažíme modelovat. Data nejsou reprezentativní když jich je málo, když nepokrývají všechny důležité připady, když jsou zkreslená atd…. Aby data byla správně reprezentativní je nutné mít dostatečně velký vzorek, správnou metodu sběru dat a pokrytí všech případů

Chyby v datech najdou vpodstatě vždy - žádný dataset ani algoritmus není dokonalý. Chyby rozděluje na dva hlavní typy:

Šum - představuje náhodnou chybu, pokud není příliš velký nemá výrazný vliv, může vznikat různými způsoby - nepřesnost přístrojů, náhodné překlepy, drobné odchylky atd…

Příklad šumu: Při měření výšky skupiny lidí nám váha měří s přesností ±0,5 kg kvůli drobným otřesům nebo nepřesnostem senzorů.

Bias (systematická chyba) - mnohem závažnější problém než šum, představuje systematické chyby nikoliv náhodnou, výrazně ovlivňují výsledky, na rozdíl od šumu se s rostoucím množstvím dat nezmírňuje - zhoršuje se

Příklady biasu
Výběrový - Například počítáme statistiky o mzdách v ČR pouze z dat absolventů vysokých škol žijících v Praze
Posun funkce - Například trénujeme program na rozpoznávání aut na datech z běžných autosalonů, ale pak ho používáme pro autosalon prodávající výhradně luxusní vozy
Měřicí bias - Například váha, která vždy ukazuje o 2 kg méně než skutečná hmotnost

(změny podmínek, časová dimenze)

Korelace a kauzalita
Korelace
Představuje statickou závislost nebo souvislost mezi dvěma veličinami. Zajímá nás, zda existuje vzájemný vztah, při kterém se při změně jedné veličiny mění druhá

Korelační koeficient - vyjáření míry korelace
+1 - dokolaná pozitivní (jedna velična roste s druhou)
0 - žádná korelace (změna jedné veličiny nemají souvislost s druhou)
-1 - dokonalá negativní (jedna veličina roste druhá klesá)

Příklady
Počet sebevražd a prodej zmrzliny - v lednu je vyšší počet sebevražd a nižší prodej zmrzliny, zatímco v létě je tomu naopak. Tyto jevy jsou negativně korelovány.

Úroveň vzdělání a příjem - statisticky spolu korelují, protože lidé s vyšším vzděláním mají v průměru vyšší příjmy.

Kauzalita
Jedna veličina přímo způsobuje nebo ovlivňuje druhou

Příklady
Příčina toho, že je Porsche drahé auto je to, že je to známá a kvalitní značka, která si může dovolit takovou cenu. Zde je kauzalita: kvalita a prestiž značky → vysoká cena.

Příčina zvýšeného výskytu slimáků a žížal je většinou deštivé počasí. Déšť způsobuje, že oba druhy vylézají na povrch. Zde je kauzalita: déšť → výskyt slimáků a žížal.