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

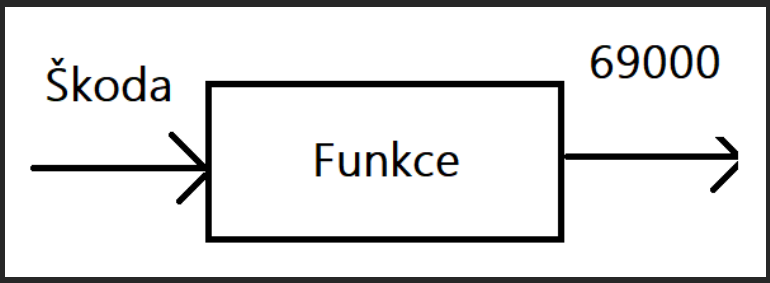
Co je to strojové učení?

Strojové učení je řešení problémů. Někdy není možné přímo vytvořit algoritmus, například pro předpověď ceny či počasí, zjišťování, co je na obrázku. Tyto algoritmy neprogramujeme přímo, proto musíme počítači ze začátku říci, co je správně a co je špatně.

Příklad: Dáme programu obrázky koček a čehokoliv jiného a on zjišťuje, na čem kočka je, a na čem není. Dopředu mu řekneme, kdy je, a kdy není.

Je důležité najít mezi daty, které už máme, vztah. Pokud je to třeba cena auta, zjišťujeme, kolik například stojí Škoda Octavia, kolik například Ferrari

Funkci je možné si v tomto případě představit jako krabičku, která má vstup a která má výstup.



Příprava dat a řešení chyb

Data jsou v drtivé většině prezentována v CSV souboru. V Pythonu si data nahráváme pomocí `data = pandas.read_csv(„zdroj“, „separátor“)` metody, která je uloží do DataFrame. DataFrame je dvojdimenzionální pole.

Chyby jsou velmi časté a vznikají například neošetřením vstupů.

Je velmi důležité v případě regrese vyřadit ty hodnoty, které nedávají smysl, nebo jsou špatně vloženy. Pro čištění dat je vhodné využít metody `data.dropna(inplace = True)`. Tato metoda vymaže všechny řádky, ve kterých je nějaká hodnota NaN (Not a Number). Pokud nevložíme `inplace = True`, tak data zkopírujeme do nové proměnné, místo toho, abychom je přímo nahradili.

V případě neplatných formátů, třeba když zadáme datum 23042004, můžeme použít různé funkce, například
`to_datetime(), to_numeric(), to_datetime()`

Je možné případně upravovat i konkrétní řádky pomocí metody loc, kterou například i pomocí cyklu můžete ošetřit celý DataFrame a postupně projet všechny data a opravit je.

V případě regrese je nutné ručně projít data a vyřadit špatná či zavádějící data, třeba že zmrzlina stojí více než 2.000 korun nebo je korunu. V případě, že bychom to neučinili, by se model začal plést.

Korelace a Kauzalita

Korelace znamená spojitost. Korelace je například i to, že je v lednu větší počet sebevražd, protože leden je depresivní měsíc a prodává se málo zmrzliny. Naopak v létě je menší počet sebevražd, protože se v létě více prodává zmrzlina. Toto spolu může korelovat.

Korelovat spolu může například zvýšená hodnota CO2 lidí a obezita. Kauzalitou bude v tomto případě to, že se jí mnohem více jídla.

Kauzalita znamená příčinnost. Například kauzalita toho, že je Porshe velmi drahé auto je to, že je Porshe velmi známá a kvalitní značka a může si dovolit takovouto cenu.

Bias

Bias je posun funkce, protože třeba jsme natrénovali program na datech z normálních autosalonů, ale využíváme tento program pro autosalon prodávající výhradně dražší auta.

Pokud nám chybí nějaká určitá data, třeba levnějších auta, protože prodáváme dražší auta, funkce bude logicky posunutá a zaujatá k tomu, aby byly všechyn auta dražší.

Bias by se také dal označit jako zkreslení výsledků kvůli vstupním datům.