Trénovací data (Training Data)

Trénovací data jsou souborem informací (texty, obrázky, tabulky, zvuky), které se používají k učení algoritmu strojového učení. Model v těchto datech hledá vzorce, vztahy a statistické souvislosti, které mu následně umožňují činit předpovědi nebo rozhodnutí u dat, která nikdy dříve neviděl.

Většina moderních systémů AI vyžaduje obrovské objemy dat (tzv. Big Data), aby dosáhla vysoké přesnosti.

1. Rozdělení dat při vývoji

Při tvorbě modelu se celkový dataset obvykle rozděluje do tří částí, aby se předešlo přetrénování (overfittingu):

- **Trénovací sada (Training set):** Největší část dat (obvykle 70–80 %), na které se model přímo učí a upravuje své vnitřní váhy.
- **Validační sada (Validation set):** Slouží k ladění parametrů modelu během učení. Pomáhá vývojáři určit, kdy je model hotov.
- **Testovací sada (Test set):** Skupina dat, která nebyla při trénování použita. Slouží k finálnímu, nestrannému ověření přesnosti modelu.

2. Strukturovaná vs. Nestrukturovaná data

Typ dat	Popis	Příklad
Strukturovaná	Data s pevným formátem, uložená v tabulkách nebo databázích.	Historie transakcí v bance, teplotní čidla.
Nestrukturovaná	Data bez jasného formátu, která tvoří většinu dnešního internetu.	Fotografie, videa, e-maily, hlasové zprávy.

3. Labeling (Anotace dat)

U **učení s učitelem** (supervised learning) musí být data „označkovaná“ (labeled). To znamená, že ke každému vstupu musí existovat správná odpověď.

- **Příklad:** Fotografie musí mít popisek „kočka“, „auto“ nebo „chodec“.
- Anotace je často nejdražší a nejpomalejší částí vývoje AI, protože ji často musí provádět lidé (anotátoři).

4. Kvalita vs. Kvantita

Pro úspěšné trénování jsou klíčové tyto vlastnosti dat:

- **Reprezentativnost:** Data musí pokrývat všechny situace, které mohou v reálném světě nastat (např. samořiditelné auto musí vidět fotky silnice ve dne, v noci, v dešti i v mlze).

- **Čistota:** Data by neměla obsahovat příliš mnoho chyb, duplicit nebo irelevantních informací (šumu).
- **Vyváženost:** Pokud bude v datech 99 % obrázků psů a 1 % koček, model se naučí každé zvíře označit jako psa.

5. Syntetická data

V poslední době se stále častěji využívají **syntetická data** – data vytvořená jinou umělou inteligencí nebo počítačovou simulací. Používají se tam, kde jsou reálná data vzácná (např. havárie letadel), příliš drahá na pořízení nebo citlivá z hlediska soukromí (lékařské záznamy).

> **Etická poznámka:** Pokud trénovací data obsahují historické předsudky (např. diskriminaci při schvalování úvěrů), model tyto předsudky převezme a bude je automaticky opakovat. Tento jev se nazývá algoritmická předpojatost.

Zpět na AI rozcestník