Metriky vzdálenosti v Machine Learningu

Metriky vzdálenosti jsou matematické funkce, které definují „blízkost“ mezi dvěma datovými body v n-rozměrném prostoru. Volba správné metriky zásadně ovlivňuje výkon algoritmů strojového učení, zejména u učení bez učitele.

1. Euklidovská vzdálenost (Euclidean Distance)

Nejpoužívanější metrika, známá z geometrie jako „vzdálenost vzdušnou čarou“. Je to délka úsečky spojující dva body.

Vzorec: $$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Vhodná pro:** Spojitá numerická data (např. souřadnice, teplota).
- **Nevýhoda:** Je velmi citlivá na měřítko (jednotky) a na odlehlé hodnoty (outliers). Vyžaduje předchozí **normalizaci dat**.

2. Manhattanská vzdálenost (Manhattan / Taxicab Distance)

Měří vzdálenost jako součet absolutních rozdílů jejich souřadnic. Název je odvozen od mřížovitého půdorysu ulic v Manhattanu, kde se nelze pohybovat šikmo skrz bloky domů.

Vzorec: $$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Vhodná pro:** Diskrétní data nebo v situacích, kdy máme vysoký počet dimenzí. Je méně citlivá na odlehlé hodnoty než Euklidovská vzdálenost.

3. Kosinová podobnost (Cosine Similarity)

Měří úhel mezi dvěma vektory. Nezajímá ji velikost (délka) vektorů, ale pouze jejich směr.

Vzorec: $$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

- **Vhodná pro:** Textovou analýzu (NLP). Například dva dokumenty mohou mít různou délku (velikost vektoru), ale pokud používají podobná slova, jejich směr (úhel) bude podobný.

4. Hammingova vzdálenost (Hamming Distance)

Používá se pro porovnání dvou řetězců stejné délky. Počítá počet pozic, na kterých se odpovídající symboly liší.

- **Příklad:** Vzdálenost mezi „101**1**101“ a „101**0**101“ je 1.
- **Vhodná pro:** Kategorické proměnné (po One-Hot kódování), analýzu genů nebo detekci chyb v

přenosu dat.

5. Minkowského vzdálenost (Minkowski Distance)

Zobecněná forma Euklidovské a Manhattanské vzdálenosti.

Vzorec: $$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

- Pokud **$p=1$**, jde o Manhattanskou vzdálenost.
- Pokud **$p=2$**, jde o Euklidovskou vzdálenost.

Srovnávací tabulka

Metrika	Typ dat	Citlivost na outliers	Hlavní využití
Euklidovská	Numerická	Vysoká	Obecné ML, shlukování
Manhattanská	Numerická / Celočíselná	Nízká	k-NN, vysokodimenzionální data
Kosinová	Text (Vektory)	Nízká	NLP, doporučovací systémy
Hammingova	Kategorická / Binární	Nulová	Porovnávání řetězců, genetika

---

Důležité upozornění: Normalizace

Většina metrik vzdálenosti vyžaduje, aby data byla ve stejném měřítku. Pokud má jedna vlastnost rozsah 0-1 (např. pravděpodobnost) a druhá 0-1000 (např. cena), bude vlastnost s větším rozsahem dominovat výpočtu vzdálenosti. Před výpočtem vždy použijte **Min-Max Scaling** nebo **Standardizaci**.

Tagy: ml matematika statistika distance_metrics clustering