

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

درس : یادگیری ماشین

استاد : خانم دکتر زرین بال ماسوله

موضوع : Similarity Measures in ML

دانشجو : فرزاد محسنی

مقطع : کارشناسی ارشد

رشته تحصیلی : مهندسی آیتی

کد دانشجویی :

دانشگاه : صنعتی امیرکبیر

| Method ----- | Page |
|--|-------------|
| Cosine Similarity ----- | 3 |
| Euclidean Distance ----- | 5 |
| Pearson Correlation Coefficient ----- | 7 |
| Jaccard Similarity ----- | 9 |
| Mahalanobis Distance ----- | 11 |
| Manhattan Distance ----- | 13 |
| Sørensen–Dice Coefficient ----- | 15 |
| Hamming Distance ----- | 17 |
| Minkowski Distance ----- | 19 |
| Spearman Correlation ----- | 21 |
| Chebyshev Distance ----- | 23 |
| Bray-Curtis Similarity / Distance ----- | 25 |
| Canberra Distance ----- | 27 |
| Jensen-Shannon Distance ----- | 29 |
| Kullback–Leibler Divergence (KL Divergence) ----- | 31 |

Cosine Similarity

تعریف: معیاری برای سنجش زاویه بین دو بردار در فضای n بعدی است، بدون توجه به اندازه بردارها.

به عبارت ساده:

- اگر دو بردار هم جهت باشند، شباهت آنها ۱ خواهد بود (کاملاً مشابه)
- اگر زاویه ۹۰ درجه باشد، شباهت صفر است (کاملاً نامرتبط)
- اگر برخلاف جهت باشند، مقدار منفی خواهد شد (کاملاً مخالف)

کاربردها :

- پردازش زبان طبیعی (NLP)
 - اندازه گیری شباهت بین اسناد یا جملات
 - مقایسه بردارهای TF-IDF
 - بررسی شباهت بین کلمات (word embeddings)
- سیستم‌های توصیه‌گر
 - پیدا کردن کاربران یا آیتم‌های مشابه
 - مقایسه رفتار کاربران بر اساس الگوهای رفتاری
- خوشه‌بندی (Clustering)
 - برای متون یا داده‌های برداری در فضاهای برداری نرمال شده
 - تشخیص کپی یا سرقت ادبی
 - مقایسه اسناد یا متن‌ها از لحاظ محتوا

ویژگی‌ها :

- مستقل از بزرگی بردار هاست
- سریع و محاسباتی سبک
- مناسب برای داده‌های با بعد بالا (مثل متن‌ها)

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|}$$

• ضرب داخلی دو بردار $\vec{A} \cdot \vec{B}$

• طول یا نُرم بردار ها $\|\vec{B}\|$ و $\|\vec{A}\|$

مراحل محاسبه :

فرض کن دو بردار A و B داریم :

• ضرب داخلی (dot product) بین A و B رو حساب کن.

• نُرم (طول) هر کدام از بردار ها رو محاسبه کن.

• مقدار ضرب داخلی رو تقسیم بر حاصل ضرب طول ها کن.

Example:

$$A = (1, 2, 3)$$

$$B = (4, 5, 6)$$

$$A \cdot B = 1 \times 4 + 2 \times 5 + 3 \times 6 = 32$$

$$|A| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}$$

$$|B| = \sqrt{4^2 + 5^2 + 6^2} = \sqrt{77}$$

$$\text{Cosine Similarity} = \frac{32}{\sqrt{14} \cdot \sqrt{77}} \approx 0.9746$$

يعنى بردار ها بسيار شبیه هستند

Euclidean Distance

- فاصله اقلیدسی فاصله مستقیم (خط راست) بین دو نقطه در فضای n بعدی است.
- این متریک ساده ترین و رایج ترین روش برای اندازه گیری فاصله بین دو نقطه است، دقیقاً مثل اندازه گیری با خط کش در هندسه.

کاربرد ها :

- الگوریتم KNN (نزدیک ترین همسایه ها)
- خوشه بندی (مانند K-Means)
- مقایسه تصاویر یا ویژگی ها
- تشخیص ناهنجاری (Anomaly Detection)
- طبقه بندی داده های عددی در فضای برداری

ویژگی ها :

- ساده و قابل فهم
- مناسب برای داده های پیوسته (عددی)
- به مقیاس ویژگی ها حساسه - بهتره داده ها نرمال سازی یا استاندارد سازی بشن

برای دو نقطه :

$$\mathbf{A} = (A_1, A_2, \dots, A_n)$$

$$\mathbf{B} = (B_1, B_2, \dots, B_n)$$

$$Euclidean\ Distance = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \dots + (A_n - B_n)^2}$$

یا بصورت خلاصه :

$$d(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^n (A_i - B_i)^2}$$

مراحل محاسبه :

- از هر مؤلفه‌ی متناظر دو بردار، تفریق بگیر
- تفریق‌ها را به توان ۲ برسان
- همه را با هم جمع کن
- از مجموع، جذر بگیر

Example:

$$\mathbf{A} = (1, 2)$$

$$\mathbf{B} = (4, 6)$$

$$Euclidean\ Distance = \sqrt{(1 - 4)^2 + (2 - 6)^2} = \sqrt{9 + 16} = \sqrt{25} = 5$$

Pearson Correlation Coefficient

تعريف: ضریب همبستگی پیرسون (Pearson) میزان همبستگی خطی بین دو متغیر را اندازه‌گیری می‌کند.

مقدار آن بین -1 تا 1 است :

- +1 → رابطه کاملاً خطی و مستقیم
- 0 → هیچ رابطه خطی
- -1 → رابطه کاملاً خطی و معکوس

کاربرد ها :

- تحلیل همبستگی در داده های آماری
- سیستم های توصیه‌گر (بررسی شباهت کاربران)
- یادگیری ماشین برای بررسی رابطه بین ویژگی ها
- بررسی اثربخشی دارو یا مداخله در علوم پزشکی
- فیلترسازی مشارکتی (Collaborative Filtering)

ویژگی ها :

- مناسب برای بررسی روابط خطی
- فقط همبستگی خطی را بررسی می‌کند، نه غیرخطی
- به مقیاس و داده های پرت حساس است

برای دو مجموعه داده :

$$A = (A_1, A_2, \dots, A_n)$$

$$B = (B_1, B_2, \dots, B_n)$$

$$r = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^n (A_i - \bar{A})^2} \cdot \sqrt{\sum_{i=1}^n (B_i - \bar{B})^2}}$$

که در آن :

- \bar{A} و \bar{B} میانگین های مجموعه های A و B هستند.
- صورت کسر، کواریانس بین A و B است.
- مخرج کسر، حاصل ضرب انحراف معیار دو مجموعه است.

مراحل محاسبه :

- محاسبه میانگین هر مجموعه
- محاسبه تفاضل هر عضو از میانگین خود
- محاسبه حاصل ضرب تفاضل ها
- تقسیم مجموع آن بر حاصل ضرب انحراف معیار دو مجموعه

Example :

$$A = (1, 2, 3) \quad , \quad B = (1, 5, 7)$$

$$\bar{A} = 2 \quad , \quad \bar{B} \approx 4.33$$

$$r = \frac{(1-2)(1-4.33) + (2-2)(5-4.33) + (3-2)(7-4.33)}{\sqrt{(1-2)^2 + (2-2)^2 + (3-2)^2} \cdot \sqrt{(1-4.33)^2 + (5-4.33)^2 + (7-4.33)^2}}$$

$$r \approx \frac{3.33 + 0 + 2.67}{\sqrt{2} \cdot \sqrt{20.67}} \approx 0.981$$

Jaccard Similarity

تعریف: شباهت جکارد معیاری برای اندازه گیری میزان اشتراک بین دو مجموعه است. این معیار اندازه اشتراک نسبت به اندازه اجتماع دو مجموعه را نشان می دهد.

مقدار دهی :

- اگر هیچ اشتراکی وجود نداشته باشد \rightarrow شباهت = ۰
- اگر هر دو مجموعه یکسان باشند \rightarrow شباهت = ۱
- دامنه این معیار از ۰ تا ۱ است

کاربرد ها :

- مقایسه اسناد (Document Similarity)
- سیستم های پیشنهاددهنده (Recommender Systems)
- خوشه بندی داده های متنی یا دسته ای
- مقایسه پروفایل های کاربران در شبکه های اجتماعی
- پردازش تصویر (برای تطبیق ناحیه ها)

ویژگی ها :

- بسیار مناسب برای داده های مجموعه ای یا باینتری
- نادیده گرفتن فروانی یا وزن عناصر
- نسبت به اندازه مجموعه ها حساس است

برای دو مجموعه A و B ، فرمول شbahت جکارد به صورت زیر است:

$$\text{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|}$$

مراحل محاسبه :

- اشتراک دو مجموعه را پیدا کن: عناصر مشترک
- اجتماع دو مجموعه را پیدا کن: همه عناصر غیرتکراری
- تعداد عناصر اشتراک را بر تعداد عناصر اجتماع تقسیم کن

Example :

$$A = \{1, 2, 3\}, \quad B = \{2, 3, 4, 5\}$$

$$A \cap B = \{2, 3\} \Rightarrow |A \cap B| = 2$$

$$A \cup B = \{1, 2, 3, 4, 5\} \Rightarrow |A \cup B| = 5$$

$$\text{Jaccard Similarity} = \frac{2}{5} = 0.4$$

Mahalanobis Distance

تعريف:

- فاصله ماهالانوبیس یک معیار فاصله است که تفاوت بین دو نقطه را با در نظر گرفتن همبستگی بین ویژگی ها (کوواریانس-کوواریانس) اندازه‌گیری می‌کند.
- برخلاف فاصله اقلیدسی که همه ویژگی ها را مستقل و با وزن برابر در نظر می‌گیرد، فاصله ماهالانوبیس به ساختار آماری داده ها (پراکندگی و همبستگی بین ویژگی ها) توجه می‌کند.

کاربرد ها:

- تشخیص ناهنجاری (Anomaly Detection)
- نقاطی که فاصله زیادی از میانگین دارند اما در نظر گرفتن ساختار داده اهمیت دارد.
- دسته بندی (Classification)
- بهخصوص در الگوریتم هایی مثل Linear Discriminant Analysis (LDA)
- تحلیل چند متغیره
- بررسی شباهت در داده هایی که ویژگی ها با هم همبستگی دارند

ویژگی ها:

- مقیاس ناپذیر نیست: به کمک ماتریس کوواریانس، مقیاس ویژگی ها را خنثی می‌کند
- دقیق‌تر در داده های با همبستگی بالا
- محاسبه‌اش نیاز به معکوس ماتریس کوواریانس دارد (ممکنه سنگین باشه برای داده های بزرگ)

| فاصله ماهالانوبیس | فاصله اقلیدسی | ویژگی |
|-----------------------------|------------------------|----------------------|
| نه (همبستگی را لحاظ می‌کند) | بله | نادیده گرفتن همبستگی |
| استانداردسازی شده درون خودش | حساس | مقیاس ویژگی ها |
| بله حتی بدون نرمال سازی | بله اگر نرمال سازی شود | مناسب برای داده خام |

$$D_M(\vec{x}, \vec{\mu}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

که در آن:

• \vec{x} : بردار داده (نقاط‌ای که فاصله‌اش را می‌سنجیم)

• $\vec{\mu}$: میانگین داده‌ها

• S^{-1} : معکوس ماتریس کوواریانس داده‌ها

• T : عملگر ترانهاده (Transpose)

Example:

نقاط داده:

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad x = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

ماتریس کوواریانس:

$$S = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

محاسبه تفاضل:

$$x - \mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

محاسبه معکوس ماتریس کوواریانس:

$$S^{-1} = \frac{1}{1 - (0.5)^2} \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} = \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{bmatrix}$$

محاسبه عبارت کامل:

$$D_M(\vec{x}, \vec{\mu}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})} = \sqrt{\begin{bmatrix} 3 & 2 \end{bmatrix} \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix}} = \sqrt{\begin{bmatrix} 3 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 0.333 \end{bmatrix}}$$

$$= \sqrt{3 \cdot 2 + 2 \cdot 0.333} = \sqrt{6.666} \approx 2.58$$

نتیجه: فاصله ماهالانوبیس بین نقطه $(3, 2) = x$ و میانگین $(0, 0) = \mu$ با در نظر گرفتن ساختار کوواریانس، تقریباً برابر است:

$D_M \approx 2.58$

Manhattan Distance

تعريف:

- فاصله منهتن مجموع قدر مطلق تفاوت‌ها بین مؤلفه‌های متناظر دو بردار است.
- بهش می‌گن «منهتن» چون مثل حرکت در خیابان‌های شبکه‌ای هست؛ فقط می‌توانی در راستای محورهای X و Y حرکت کنی، نه مورب.

کاربرد‌ها:

- الگوریتم‌های ساده مثل KNN (وقتی داده‌ها گسسته باشند)
- تشخیص ناهنجاری
- خوشه‌بندی (Clustering)
- مسائل مسیریابی (Pathfinding) در هوش مصنوعی یا گراف‌ها
- تحلیل داده‌هایی که فقط در راستای محورها معنی دار هستند

ویژگی‌ها:

- ساده و سریع در محاسبه
- مناسب برای داده‌های گسسته یا شبکه‌ای
- کمتر تحت تأثیر داده‌های پرت نسبت به اقلیدسی
- برای داده‌های با مقیاس متفاوت، نیاز به نرمال‌سازی دارد

| فاصله منهتن | فاصله اقلیدسی | ویژگی |
|-------------|---------------|-----------------------|
| نه | بله | با ریشه‌گیری |
| کمتر | بله | حساس به پرت‌ها |
| بله | نه | فقط در راستای محور‌ها |

برای دو نقطه:

$$\mathbf{A} = (A_1, A_2, \dots, A_n) , \quad \mathbf{B} = (B_1, B_2, \dots, B_n)$$

فرمول فاصله منهتن به صورت زیر است:

$$\text{Manhattan Distance} = \sum_{i=1}^n |A_i - B_i|$$

Example:

$$\mathbf{A} = (1, 2, 3) , \quad \mathbf{B} = (4, 0, 6)$$

$$\text{Manhattan Distance} = |1 - 4| + |2 - 0| + |3 - 6| = 3 + 2 + 3 = 8$$

Sørensen–Dice Coefficient

تعریف:

- ضریب سورنسن-دایس یک معیار برای سنجش شباهت بین دو مجموعه است.
- خیلی شبیه شباهت جکارد هست، با این تفاوت که در فرمولش به اشتراک وزن بیشتری داده می‌شود (ضریب ۲ در صورت کسر).

کاربردها:

- شباهت متون و استناد
- مقایسه فایل‌ها یا رشته‌ها (string similarity)
- تشخیص کپی / سرقت ادبی
- پردازش تصویر (مقایسه نواحی)
- زیست‌اطلاعاتی (Bioinformatics)

دامنه مقدار:

- بین ۰ و ۱
- ۰ → هیچ اشتراکی ندارند
- ۱ → کاملاً یکسان هستند

مراحل محاسبه:

• اشتراک دو مجموعه را پیدا کن: $|A \cap B|$

• اندازه هر مجموعه را جداگانه محاسبه کن: $|A|$ و $|B|$

• فرمول را اعمال کن:

$$\text{Dice}(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

Example:

$$A = \{1, 2, 3\} , B = \{2, 3, 4, 5\}$$

$$|A \cap B| = 2 , |A| = 3 , |B| = 4$$

$$\text{Dice} = \frac{2 \times 2}{3 + 4} = \frac{4}{7} \approx 0.571$$

Hamming Distance

تعریف:

فاصله همینگ تعداد موقعیت هایی (index ها) را می شمارد که در آن ها دو رشته یا بردار با طول برابر، مقادیر متفاوت دارند.

یعنی:

فاصله همینگ = چند جای رشته اول با رشته دوم فرق دارد؟

کاربرد ها:

- بررسی خطا در داده های باینری (Error detection/correction)
- شبکه های کامپیوتروی و مخابرات
- مقایسه رشته های DNA در بیوانفورماتیک
- الگوریتم های جستجو و تطبیق رشته
- یادگیری ماشین با داده های باینری یا دسته ای

ویژگی ها:

- بسیار ساده و سریع
- فقط با داده های هم طول کار می کند
- مخصوص مقایسه داده های گسسته یا باینری

فرمول:

برای دو رشته یا بردار با طول مساوی:

$$\text{Hamming Distance} = \sum_{i=1}^n [A_i \neq B_i]$$

که در آن:

اگر $[A_i \neq B_i]$ باشد، ۱ در جمع لحاظ می‌شود •

اگر برابر باشند، ۰ لحاظ می‌شود •

شرط مهم:

طول دو رشته یا بردار باید برابر باشد. در غیر این صورت، فاصله همینگ تعریف نمی‌شود. •

Example:

$$A = 1011101 , \quad B = 1001001$$

مقایسه رقم به رقم:

| Index | A | B | Different? |
|-------|---|---|------------|
| 1 | 1 | 1 | ✗ No |
| 2 | 0 | 0 | ✗ No |
| 3 | 1 | 0 | ✓ Yes |
| 4 | 1 | 1 | ✗ No |
| 5 | 1 | 0 | ✓ Yes |
| 6 | 0 | 0 | ✗ No |
| 7 | 1 | 1 | ✗ No |

Differences at positions 3 and 5 only.

Hamming Distance = 2

Minkowski Distance

تعريف:

- فاصله مینکوفسکی یک فرمول کلی و تعمیم یافته برای اندازه‌گیری فاصله بین دو نقطه در فضای n بعدی است.
- این متریک، یک پارامتر p دارد که بسته به مقدار آن، می‌تواند به فاصله اقلیدسی یا فاصله منهتن یا سایر حالتها تبدیل شود.

کاربردها:

- الگوریتم KNN با فاصله قابل تنظیم
- خوشه‌بندی داده‌ها
- مقایسه نقاط در فضای ویژگی‌ها با انعطاف بالا
- داده‌هایی که ویژگی‌ها نقش متفاوتی دارند

ویژگی‌ها:

- تعمیم یافته و قابل تنظیم
- قابل استفاده در فضاهای با ابعاد بالا
- انتخاب مقدار مناسب p اهمیت زیادی دارد
- نیاز به نرمال‌سازی ویژگی‌ها در صورت تفاوت مقیاس

| نوع فاصله‌ای که به دست می‌آید | مقدار p |
|------------------------------------|------------------------|
| (Manhattan Distance) فاصله منهتن | $P = 1$ |
| (Euclidean Distance) فاصله اقلیدسی | $P = 2$ |
| (Chebyshev Distance) فاصله چبیشف | $P \rightarrow \infty$ |

برای دو نقطه:

$$\mathbf{A} = (A_1, A_2, \dots, A_n) , \quad \mathbf{B} = (B_1, B_2, \dots, B_n)$$

فاصله مینکوفسکی برابر است با:

$$\text{Minkowski Distance} = \left(\sum_{i=1}^n |A_i - B_i|^p \right)^{\frac{1}{p}}$$

Example:

$$\mathbf{A} = (1, 2) , \quad \mathbf{B} = (4, 6)$$

If $p = 1$: $|1 - 4| + |2 - 6| = 3 + 4 = 7 \Rightarrow \text{Manhattan Distance}$

If $p = 2$: $\sqrt{(1 - 4)^2 + (2 - 6)^2} = \sqrt{9 + 16} = \sqrt{25} = 5 \Rightarrow \text{Euclidean Distance}$

If $p = 3$: $(|1 - 4|^3 + |2 - 6|^3)^{1/3} = (27 + 64)^{1/3} = 91^{1/3} \approx 4.481$

Spearman Correlation

تعريف:

- Spearman Correlation Coefficient (ضریب همبستگی رتبه‌ای اسپیرمن) یک معیار غیرخطی برای اندازه‌گیری همبستگی بین دو متغیر است که بر اساس رتبه‌ی مقادیر عمل می‌کند، نه مقدار عددی خودشان.
- برخلاف Pearson که رابطه‌ی خطی بین دو متغیر را بررسی می‌کند، Spearman بررسی می‌کند که آیا ترتیب (رتبه) داده‌ها با هم هماهنگ است یا نه.

ویژگی‌ها:

- مقدار آن بین $-1+$ و $1-$ است
- $1+ \rightarrow$ رابطه کاملاً صعودی (ترتیب‌ها یکسان)
- $-1- \rightarrow$ رابطه کاملاً نزولی (ترتیب‌ها برعکس)
- $0 \rightarrow$ بدون رابطه یکنواخت

کاربرد‌ها:

- داده‌هایی که رابطه‌شان غیرخطی ولی ترتیبی است
- علوم اجتماعی و روانشناسی
- یادگیری ماشین (برای ارزیابی رتبه‌گذاری)
- رتبه‌بندی جستجوها یا نتایج

اگر دو متغیر X و Y دارای n داده باشند و d_i تفاوت رتبه‌های هر جفت باشد:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- ضریب همبستگی اسپیرمن

- $d_i = \text{rank}(X_i) - \text{rank}(Y_i)$

مراحل محاسبه:

- مرتب‌سازی هر ستون و تعیین رتبه‌ها
- محاسبهٔ تفاضل رتبه‌ها برای هر جفت داده
- به توان ۲ رساندن و جمع کردن
- قرار دادن در فرمول بالا

Example:

فرض کنیم دو لیست داریم:

$$X = [10, 20, 30], \quad Y = [100, 200, 300]$$

رتبه‌های هر متغیر:

$$\text{rank}(X) = [1, 2, 3] \quad , \quad \text{rank}(Y) = [1, 2, 3]$$

$$d_i = 0 \text{ for all, } \rho = 1$$

یعنی رابطهٔ صعودی کامل

Chebyshev Distance

تعريف:

- فاصله چبیشف (Chebyshev) یک معیار اندازه‌گیری فاصله بین دو نقطه است که به جای مجموع یا جذر مجموع، بیشترین اختلاف مطلق بین مؤلفه‌های متناظر دو بردار را در نظر می‌گیرد.
- این فاصله زمانی به دست می‌آید که پارامتر p در فرمول فاصله مینکوفسکی به سمت بی‌نهایت میل کند.

ویژگی ها:

- فقط به بزرگ‌ترین اختلاف اهمیت می‌دهد
- مستقل از جمع کل اختلاف‌ها
- ساده و سریع در محاسبه
- مناسب فقط برای کاربردهایی که بیشترین تفاوت مهم است

کاربرد ها:

- مسائل شطرنج یا مسیریابی در شبکه‌های گردیدی
- الگوریتم‌های KNN یا خوشبندی در داده‌های خاص
- مقایسه رشته‌هایی که تغییرات شدید در یک بعد اهمیت دارند
- تشخیص ناهنجاری در داده‌هایی که یک ویژگی ممکن است ناگهان تغییر زیادی کند

برای دو نقطه:

$$\mathbf{A} = (A_1, A_2, \dots, A_n) , \quad \mathbf{B} = (B_1, B_2, \dots, B_n)$$

$$\text{Chebyshev Distance} = \max_i(|A_i - B_i|)$$

یعنی فقط بیشترین مقدار اختلاف بین مؤلفه‌های دو بردار را در نظر می‌گیریم.

Example:

$$\mathbf{A} = (1, 5, 3) , \quad \mathbf{B} = (4, 2, 8)$$

$$|1 - 4| = 3 , \quad |5 - 2| = 3 , \quad |3 - 8| = 5$$

$$\text{Chebyshev Distance} = \max(3, 3, 5) = 5$$

Bray-Curtis Similarity / Distance

تعریف:

- یک معیار برای سنجش تشابه (یا فاصله) بین دو بردار غیرمنفی است (معمولًاً در اکولوژی یا داده های شمارشی).
- این معیار بررسی می کنند که چقدر دو بردار از نظر ترکیب نسبی عناصرشون شبیه یا متفاوت هستند.
- اگر دو بردار کاملاً یکسان باشند \rightarrow شباهت = 1 و فاصله = 0
- اگر هیچ اشتراکی نداشته باشند \rightarrow شباهت = 0 و فاصله = 1

ویژگی ها:

- فقط از مقادیر مثبت یا صفر پشتیبانی می کند
- به مقادیر نسبی (نه صرفاً قدر مطلق) حساس است
- مقدار نرمال شده بین 0 و 1
- برای داده های منفی مناسب نیست

کاربرد ها:

- تحلیل تنوع زیستی (ecology & biology)
- مقایسه توزیع یا فراوانی گونه ها
- پردازش متن و داده های شمارشی
- خوشه بندی داده های ترکیبی

برای دو بردار غیرمنفی:

$$A = (A_1, A_2, \dots, A_n) , \quad B = (B_1, B_2, \dots, B_n)$$

$$\text{Bray-Curtis Distance} = \frac{\sum_{i=1}^n |A_i - B_i|}{\sum_{i=1}^n (A_i + B_i)}$$

Similarity = 1 – Distance

Example:

$$A = (1, 2, 3), \quad B = (2, 2, 4)$$

$$|1 - 2| = 1, \quad |2 - 2| = 0, \quad |3 - 4| = 1 \Rightarrow \sum |A_i - B_i| = 2$$

$$1 + 2 + 2 + 2 + 3 + 4 = 14 \Rightarrow \sum (A_i + B_i) = 14$$

$$\text{Distance} = \frac{2}{14} = 0.143 , \quad \text{Similarity} = 1 - 0.143 = 0.857$$

Canberra Distance

تعريف:

- فاصله کانبرا یک معیار فاصله است که مقدار فاصله بین دو بردار را با در نظر گرفتن نسبت تفاوت به مجموع قدر مطلق مقادیر برای هر مؤلفه محاسبه می کند.
- این فاصله بیشتر برای داده هایی با مقادیر کوچک یا شامل صفر مفید است، چون هر مؤلفه به صورت نرمال شده در خودش تأثیر دارد.

ویژگی ها:

- نرمال سازی شده: برای داده هایی با مقیاس های مختلف بهتر عمل می کند
- حساس به تغییرات کوچک در مقادیر نزدیک به صفر
- به داده های شامل صفر یا نزدیک به صفر حساس است
- نسبت به نویز و خطاهای اندازه گیری در مقادیر کوچک حساس تر است

کاربرد ها:

- مقایسه بردارهای ویژگی با مقیاس های متفاوت
- داده های نرمال نشده یا شامل صفر
- داده هایی که تفاوت نسبی مهم تر از مطلق است
- متن کاوی، تشخیص ناهنجاری، و تحلیل های عددی خاص

برای دو بردار:

$$\mathbf{A} = (A_1, A_2, \dots, A_n), \quad \mathbf{B} = (B_1, B_2, \dots, B_n)$$

فرمول فاصله کانبرا:

$$\text{Canberra Distance} = \sum_{i=1}^n \frac{|A_i - B_i|}{|A_i| + |B_i|}$$

نکته: اگر $B_i = A_i = 0$ آن مؤلفه را در مجموع در نظر نمی‌گیریم (برای جلوگیری از تقسیم بر صفر).

Example:

$$\mathbf{A} = (3, 0, 4) \quad , \quad \mathbf{B} = (6, 0, 2)$$

$$\frac{|3 - 6|}{|3| + |6|} = \frac{3}{9} = 0.333$$

$$\frac{|0 - 0|}{|0| + |0|} = \text{undefined} \Rightarrow \text{ignored}$$

$$\frac{|4 - 2|}{|4| + |2|} = \frac{2}{6} = 0.333$$

$$\text{Canberra Distance} = 0.333 + 0.333 = 0.666$$

Jensen-Shannon Distance

تعریف:

- فاصله جنسن-شانون (JSD) یا (JS) یک معیار فاصله بین دو توزیع احتمال است که نسخه‌ی متقارن و همواره قابل تعریف از KL Divergence به حساب می‌آید.
- برخلاف KL Divergence که ممکن است نامحدود یا تعریف‌نشده باشد، JSD همیشه مقدار محدود، متقارن و دارای ریشه دوم پذیر دارد.

ویژگی ها:

- متقارن: $JSD(P||Q) = JSD(Q||P)$
- همیشه تعریف شده (برخلاف KL)
- مقدار آن بین ۰ تا ۱ قرار دارد (اگر از $\log \text{base } 2$ استفاده شود)
- می‌توان ریشه آن را گرفت تا یک معیار فاصله واقعی (metric) باشد

کاربرد ها:

- مقایسه بین دو توزیع احتمال (مثل توزیع کلمات در متون)
- طبقه‌بندی اسناد در NLP
- مقایسه مدل‌های زبانی
- خوشبندی داده‌های آماری
- در یادگیری ماشین برای اندازه‌گیری شباهت بین پیش‌بینی مدل‌ها

برای دو توزیع احتمال P و Q (که مجموع مؤلفه هایشان برابر ۱ است)، ابتدا توزیع میانگین M را تعریف می کنیم:

$$M = \frac{1}{2}(P + Q)$$

سپس فاصله جنسن-شانون به صورت زیر تعریف می شود:

$$JSD(P || Q) = \sqrt{\frac{1}{2}D_{KL}(P || M) + \frac{1}{2}D_{KL}(Q || M)}$$

که در آن **KL Divergence** همان D_{KL} است.

Example:

$$P = (0.5, 0.4, 0.1) \quad , \quad Q = (0.2, 0.2, 0.6)$$

1. محاسبه میانگین

$$M = \frac{1}{2}(P + Q) = (0.35, 0.3, 0.35)$$

2. محاسبه $KL(Q||M)$ و $KL(P||M)$

3. ترکیب آنها با میانگین \rightarrow سپس جذر بگیریم

(محاسبات دقیق تر معمولاً با نرم افزار انجام می شود)

Kullback–Leibler Divergence (KL Divergence)

تعریف:

- KL Divergence یک معیار آماری است که تفاوت بین دو توزیع احتمال را اندازه‌گیری می‌کند.
- در واقع، می‌سنجد اگر داده‌ها واقعاً از توزیع P آمده باشند، چقدر هزینه دارد که فرض کنیم آن‌ها از توزیع Q آمده‌اند.

کاربردها:

- یادگیری ماشین — برای سنجش تفاوت بین توزیع‌های واقعی و پیش‌بینی شده
- پردازش زبان طبیعی — مقایسه مدل‌های زبانی
- فشرده‌سازی اطلاعات
- نظریه اطلاعات
- شبکه‌های عصبی — برای توابع هزینه (loss functions) مثل cross-entropy

برای دو توزیع احتمال گسسته P و Q روی فضای مشترک:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

نکات مهم:

- $D_{KL}(P \parallel Q) \geq 0$ همیشه نا منفی است: KL Divergence
- فقط وقتی صفر است که $P = Q$: $D_{KL}(P \parallel Q) = 0$ only if $P = Q$
- نامتقارن است: $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$
- اگر در نقطه 0 ، مقدار تعریف نشده می‌شود (تقسیم بر صفر)

تفسیر ساده:

- «چقدر اطلاعات از دست می‌دهیم اگر به جای توزیع واقعی P ، توزیع Q را در نظر بگیریم؟»

Example:

$$P = (0.5, 0.4, 0.1), \quad Q = (0.4, 0.3, 0.3)$$

$$D_{KL}(P || Q) = 0.5 \log\left(\frac{0.5}{0.4}\right) + 0.4 \log\left(\frac{0.4}{0.3}\right) + 0.1 \log\left(\frac{0.1}{0.3}\right)$$

با محاسبه (در پایه ۲ یا ۱۰)، عددی بین ۰ و ۱ به دست می‌آید.