Ответ 2

Объясните разницу между евклидовым расстоянием (Euclidean Distance) и косинусным сходством (Cosine Similarity) при сравнении векторов. В каких случаях предпочтительнее использовать каждую из этих метрик?

На основе предоставленного учебного материала, разница между евклидовым расстоянием и косинусным сходством, а также случаи их предпочтительного использования, объясняются следующим образом.

Введение: Как измерить "близость" векторов

Для того чтобы компьютер мог выполнять семантический поиск (поиск по смыслу), данные (например, текст) преобразуются в числовые векторы, или **embeddings**. Ключевая идея заключается в том, что объекты с похожим смыслом будут иметь похожие векторы. Чтобы математически определить, насколько два вектора "близки" друг к другу, используются метрики расстояния или сходства. Двумя самыми популярными метриками, описанными в материале, являются евклидово расстояние и косинусное сходство.

1. Евклидово расстояние (Euclidean Distance, L2)

Евклидово расстояние измеряет **«прямолинейное» расстояние** между двумя точками (концами векторов) в многомерном пространстве. Это интуитивно понятная метрика, аналогичная измерению расстояния линейкой на двухмерной карте.

Математическое определение

Для двух векторов $\vec{A} = (a_1, a_2, \ldots, a_n)$ и $\vec{B} = (b_1, b_2, \ldots, b_n)$ в n-мерном пространстве, евклидово расстояние $d(\vec{A}, \vec{B})$ вычисляется по следующей формуле:

$$d(\vec{A}, \vec{B}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Пошаговое объяснение формулы:

1. $(a_i - b_i)$: Для каждого измерения (от 1 до n) вычисляется разница между соответствующими компонентами векторов. Это показывает, насколько векторы далеки друг от друга по каждой отдельной оси.
2. $(a_i - b_i)^2$: Каждая разница возводится в квадрат. Это делается для того, чтобы все значения стали положительными (так как расстояние не может быть отрицательным) и чтобы придать больший "вес" большим различиям между компонентами.
3. $\sum_{i=1}^{n} \ldots$: Знак суммы (сигма) означает, что все полученные квадраты разностей складываются. В результате получается общая сумма квадратов расстояний по всем измерениям.
4. $\sqrt{\ldots}$: Из полученной суммы извлекается квадратный корень. Этот шаг возвращает нас к исходным единицам измерения и является аналогом теоремы Пифагора для многомерного пространства.

Практическое применение

Чем **меньше** значение евклидова расстояния, тем ближе векторы друг к другу, и, следовательно, тем более похожими считаются исходные данные. Если расстояние равно 0, это означает, что векторы полностью идентичны. Эта метрика чувствительна к **магнитуде (длине)** векторов.

2. Косинусное сходство (Cosine Similarity)

В отличие от евклидова расстояния, которое измеряет дистанцию, косинусное сходство измеряет **угол** между двумя векторами. Оно определяет, насколько векторы "смотрят" в одном направлении, **независимо от их длины (магнитуды)**.

Математическое определение

Косинусное сходство вычисляется как косинус угла θ между векторами $\vec{A}$ и $\vec{B}$:

$$\text{similarity}(\vec{A}, \vec{B}) = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

Пошаговое объяснение формулы:

1. $\vec{A} \cdot \vec{B}$ **(Скалярное произведение)**: Числитель дроби. Вычисляется путем перемножения соответствующих компонент векторов и сложения результатов ($\sum_{i=1}^{n} a_i b_i$). Скалярное произведение будет большим, если векторы "со-направлены".
2. $\|\vec{A}\|$ **(Магнитуда или норма вектора)**: Знаменатель дроби. Это длина вектора, которая вычисляется как $\sqrt{\sum_{i=1}^{n} a_i^2}$. По сути, это евклидово расстояние от начала координат до конца вектора. То же самое вычисляется и для вектора $\vec{B}$.
3. $\frac{}{\dots}$ **(Нормализация)**: Скалярное произведение делится на произведение магнитуд векторов. Эта операция **нормализации** устраняет влияние длины векторов, оставляя только информацию об их направлении.

Практическое применение

Результат косинусного сходства всегда находится в диапазоне от -1 до 1:

- **1**: Максимальное сходство. Векторы указывают в одном направлении (угол 0°).
- **0**: Нет сходства. Векторы ортогональны (перпендикулярны) друг другу (угол 90°).
- **-1**: Максимальная непохожесть. Векторы указывают в противоположных направлениях (угол 180°).

При поиске похожих объектов мы ищем векторы со значением косинусного сходства, близким к **1**.

Разница и предпочтительное использование

Основное различие между двумя метриками заключается в том, **что они измеряют**:

- **Евклидово расстояние** измеряет **дистанцию** и учитывает как направление, так и магнитуду (длину) векторов. Два вектора могут указывать в одном направлении, но если один из них намного длиннее другого, евклидово расстояние между их концами будет большим.
- **Косинусное сходство** измеряет **направление** и игнорирует магнитуду. Оно фокусируется исключительно на угле между векторами.

В каких случаях что предпочесть:

Согласно учебному материалу, выбор метрики зависит от задачи.

Косинусное сходство предпочтительнее в случаях, когда важна тематика или семантическое содержание, а не "объем" или длина исходных данных. В материале приводится конкретный пример:

> "Это особенно полезно для **текстовых данных**, где длина документа может сильно варьироваться, но нас интересует именно его тематика (направление)."

Например, короткое предложение «Какая погода?» и длинный документ, подробно описывающий прогноз погоды на сегодня, могут иметь векторы разной длины. Евклидово расстояние между ними может быть большим. Однако, поскольку оба текста об одном и том же (о погоде), их векторы будут указывать в схожем направлении в семантическом пространстве, и косинусное сходство между ними будет высоким (близким к 1).

Евклидово расстояние является более общей и интуитивной мерой "близости". Хотя материал не указывает конкретных сценариев, где оно предпочтительнее, его природа (измерение прямолинейного расстояния) делает его полезным, когда магнитуда вектора несет важную информацию, и разница в абсолютных значениях компонент вектора имеет значение.