Some ways to calculate similarity of distribution

**1 Abstract**

This article explores various methods for calculating the similarity of probability distributions. Four specific methods are discussed: Kullback-Leibler divergence, cosine similarity, total variation distance, and Pearson correlation coefficient.

**2 Background**

LDA (Latent Dirichlet Allocation) outputs the probability distribution of a document across different categories. When comparing the similarity between two documents, we need to compare their probability distributions. After investigation, the following algorithms are commonly used to compare probability distributions:

1. Kullback-Leibler Divergence
2. Cosine Similarity
3. Total Variation Distance
4. Pearson Correlation
5. Wasserstein Distance

**3 Introduction**

**3.1. Kullback-Leibler Divergence[1]**

It quantifies the amount of information lost when one distribution is used to approximate another. The Kullback-Leibler Divergence between two probability distributions P and Q is calculated using the following formula:

The Kullback-Leibler Divergence is not symmetric, which means is not equal to . It measures the dissimilarity between the two distributions rather than a distance metric.

A lower value of Kullback-Leibler Divergence indicates a higher similarity between the distributions. If the Kullback-Leibler Divergence is zero, it means the two distributions are identical[2].

It's important to note that Kullback-Leibler Divergence is not a true distance metric as it violates the triangle inequality. Therefore, it should be interpreted as a measure of dissimilarity rather than a distance measure.

**3.2. Cosine Similarity[3]**

Cosine Similarity is a measure of similarity between two vectors in a multi-dimensional space. It calculates the cosine of the angle between the vectors, which captures the directional similarity while disregarding the magnitude or length of the vectors.

The Cosine Similarity between two vectors A and B is calculated using the following formula:

**3.3. Total Variation Distance[4]**

It quantifies the total discrepancy or variation between the two distributions.

The Total Variation Distance between two probability distributions P and Q is calculated using the following formula:

The Total Variation Distance measures the total absolute difference between the probabilities of corresponding events in the two distributions. It provides a measure of how much the two distributions differ on average.

The Total Variation Distance ranges from 0 to 1. A value of 0 indicates that the two distributions are identical, while a value of 1 indicates maximum dissimilarity, meaning the two distributions have no common events.

**3.4. Pearson Correlation**

It quantifies the strength and direction of the linear association between the variables. The Pearson Correlation between two variables X and Y is calculated using the following formula[5]:

The Pearson Correlation coefficient, denoted by ρ, ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, meaning that as X increases, Y also increases proportionally. A value of -1 indicates a perfect negative linear relationship, where as X increases, Y decreases proportionally. A value of 0 suggests no linear relationship or correlation between the variables.

Pearson Correlation assumes that the relationship between the variables is linear and that the variables are normally distributed. It is sensitive to outliers and can be affected by non-linear relationships or the presence of influential points.

**3.5. Wasserstein Distance**

It quantifies the minimum amount of work required to transform one distribution into another.

The Wasserstein Distance between two probability distributions P and Q is calculated by finding the optimal transport plan that minimizes the cost of moving mass from one distribution to another. It considers the geometry of the underlying space and the costs associated with moving mass from one location to another[6].

It can be expressed as:

where inf denotes the infimum (greatest lower bound), and c represents the cost function that specifies the cost of moving mass from one location to another[7].

The Wasserstein Distance provides a metric that captures both the similarity in shape and the distance between distributions. It is especially useful when dealing with distributions that have different modes, heavy tails, or when comparing distributions in high-dimensional spaces.

**Reference:**

1. Joyce, J.M. (2011). Kullback-Leibler Divergence. In: Lovric, M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg.

<https://doi.org/10.1007/978-3-642-04898-2_327>

1. COUNT BAYESIE. (2017). Kullback-Leibler Divergence Explained.

<https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained>

1. Selva Prabhakaran. (2018). Cosine Similarity – Understanding the math and how it works (with python codes).

<https://www.machinelearningplus.com/nlp/cosine-similarity/?utm_content=cmp-true>

1. Chatterjee, Sourav. "Distances between probability measures" (PDF). UC Berkeley. Archived from the original (PDF) on July 8, 2008. Retrieved 21 June 2013.
2. Shaun Turney. (2022). Pearson Correlation Coefficient (r) | Guide & Examples.

<https://www.scribbr.com/statistics/pearson-correlation-coefficient/>

1. Kowshik chilamkurthy. (2020). Wasserstein Distance, Contraction Mapping, and Modern RL Theory.

<https://kowshikchilamkurthy.medium.com/wasserstein-distance-contraction-mapping-and-modern-rl-theory-93ef740ae867>

1. SciPy document.

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html>