Bachelor's Thesis Proposal

Title

Diffusion Distance and Manifold Learning for Clustering Non-Linear Data

1. Motivation and Background

Clustering is a fundamental task in data analysis and machine learning. Classical clustering methods such as k-means rely on Euclidean distance and linear separability assumptions. However, many real-world datasets—such as images, text embeddings, biological data, or sensor measurements—lie on **non-linear manifolds** embedded in high-dimensional spaces.

In such cases, Euclidean distance fails to reflect the true geometric relationships between data points. **Manifold learning**techniques aim to uncover the intrinsic geometry of data by modeling local neighborhood relationships. One prominent approach is based on **diffusion processes on graphs**, which lead to the concept of **diffusion distance**.

Diffusion distance measures similarity between data points by comparing the behavior of random walks starting from them, thereby approximating distances along the manifold rather than through the ambient space. This thesis investigates how diffusion distance improves clustering performance on non-linear manifolds compared to standard distance measures.

2. Problem Statement

Standard clustering algorithms often perform poorly when applied directly to data lying on non-linear manifolds. The core problem addressed in this thesis is:

How can diffusion distance and manifold learning techniques be used to improve clustering quality for non-linearly structured data?

3. Objectives

The main objectives of this thesis are:

1. To explain the theoretical foundations of diffusion distance and manifold learning.

2. To implement diffusion-based embeddings (e.g. diffusion maps) for real and synthetic datasets.

3. To compare clustering results using:
    - Euclidean distance

    - Diffusion distance (via diffusion maps)

4. To evaluate clustering performance using standard metrics.

5. To analyze strengths and limitations of diffusion-based clustering methods.

4. Expected Results

It is expected that:

- Diffusion-based representations will better preserve manifold structure

- Clustering performance will improve on non-linear datasets

- Diffusion distance will be more robust to noise than Euclidean distance

6. Scope and Limitations

- Focus on **unsupervised clustering**

- Emphasis on interpretability and geometry rather than computational optimization