Abstract

Practitioners of Machine Learning and related fields commonly seek out embeddings of object collections into some Euclidean space. These embeddings are useful for dimensionality reduction, for data visualization, as concrete representations of abstract notions of similarity for similarity search, or as features for some downstream learning task such as web search or sentiment analysis. A wide array of such techniques exist, ranging from traditional (PCA, MDS) to trendy (word2vec, deep learning).

While most existing techniques rely on preserving some type of *exact numeric* data (feature values, or estimates of various statistics), I propose to develop and apply large-scale techniques for embedding and similarity search using purely *ordinal* data (e.g. "object a is more similar to b than to c"). Recent theoretical advances show that ordinal data does not inherently lose information, in the sense that, when carefully applied to an appropriate dataset, there is an embedding satisfying ordinality which is unique up to similarity transforms (scaling, translation, reflection, and rotation). Further, ordinality is often a more natural way to represent the common goal of finding an embedding which preserves some notion of similarity without taking noisy statistical estimates too literally.

The work I propose focuses on three tasks: selecting the minimal ordinal data needed to produce a high-quality embedding, embedding large-scale datasets of high dimensionality, and developing ordinal embeddings that depend on contextual features for, e.g., recommender systems.