UNIVERSITY OF CALIFORNIA, BERKELEY

Lecture 23: CS - 189

Oscar Ortega

July 16, 2021

1 CLUSTERING WITH MULTIPLE EIGENVALUES

For k clusters, compute first k eigenvectors $v_1 = \mathbf{1}, v2, ..., v_k$ of generalized eigensystem:

$$Lv = \lambda M v$$

Row v_i is a **spectral vector** for vertex i. Normalize each row V_i to unit length. $V \in \mathbf{R}^{n,k}$ where the first column in the matrix v is the vector of ones. K means always use a straight line separator for the components. However, Spectral clustering can come up with radial decision boundaries. The algorithm was invented by Professor, Michael Jordan, Andrew Ng, Yair Weiss.

2 RANDOM PROJECTIONS

Another dimensionality reduction technique, a cheap alternative to PCA as preprocess for clustering, classification, and regression. Approximately preserves distances between points. Pick a small ϵ, a small δ, and a random subspace $S \subseteq \mathbf{R}^d$ of dimension k, where $k =$ ceiling$\frac{2 \ln 1\delta}{\epsilon^2/2 - \epsilon^3/3}$. For any pt q, let $\hat{q}$ be the orthogonal projection of q onto S, multiplied by $\sqrt{\frac{d}{k}}$

2.1 JOHNSON - LINDENSTRUSS LEMMA(MODIFIED)

For any two pts $q, w \in \mathbb{R}^d$, $(1-\epsilon)\|q-w\|_2^2 \leq \|\hat{q}-\hat{w}\|_2^2 \leq (1+\epsilon)\|q-w\|_2^2$ with probability $\geq 1-2\delta$
Typical values: $\epsilon \in [0.02, 0.5], \delta \in [\frac{1}{n^3}, 0.05]$
The Geometry of High-Dimensional Spaces Consider a sphere within a sphere, Let the radius of the outer sphere have radius r, and let the inner sphere have a radius of $r-\epsilon$. Consider the shell between spheres of. And consider sampling points within the sphere, It turns out most of points will be in the shell between the spheres.

1. Volume of outer ball : αr^d

2. Volume of inner ball : $\alpha(r-\epsilon)^d = (1-\frac{\epsilon}{r})^d \approx \exp(-\frac{\epsilon d}{r})$ which is small for large d. So the inner sphere will have no volume compared to the larger one as $\lim d \to \infty$. The percentage of the volume of the small sphere approaches 0. So if we sample points from uniform distribution in ball: nearly all are in outer shell.

3. If we consider a gaussian, a similar situation will occur:

$$\|P\|_2^2 = \sum_{i=1}^{d} P_i^2$$

$$\mathbb{E}[\|P\|_2^2] = d\mathbb{E}[P_1^2]$$

$$\text{var}(\|P\|_2^2) = \sqrt{d}\text{var}(P_1^2)$$

4. In high dimensions the nearest neighbor and farthest neighbors don't differ by much.

5. This means that algorithms like k-means clustering and nearest neighbor classifiers are less effective.

6. marckhoury.github.io: if I want to no more about this later.

3 Latent Factor Analysis

Suppose X is a term-document matrix (bag-of-words) model: The entries of a bag of words model are very sparce.

$$X_{i,j} = \text{ number of occurences of term j in doc i}$$

In practice we often take log(1 + number of occurrences) We can weigh the entries by how common these words appear in the english language.
In this case We can compute the SVD of the bag of words model and unlike in PCA, we usually don't center X. For large singular values, the left and right singular vectors correspond to a cluster of documents. Where the left singular vectors will cluster based on the genre, and the right singular vectors will have common terms in that genre. This is like clustering, but the clusters can overlap.