sergey@cs.stanford.edu
Oct 15 1998

CS 349

Singular Value Decomposition Revisited and Clustering

Sergey Brin

October 15, 1998

Outline

A few more words about SVD.

Clustering Applications

Clustering Techniques

Projecting to Linear Subspaces

Data Points in a High Dimensional Space

A Low Dimensional Projection

sergey@cs.stanford.edu

Choosing Which Low Dimensional Projection

Intuitive Answer

The Mathematical Answer

Covariance Matrix

Eigenvectors

- The dominant eigenvector.

An Example

Dataset: (-1.1,-1.0), (0.1, 0.0), (1.0,1.0)

Covariances:

- Note: E(X) = E(Y) = 0
- X,X = Var(X) = E(X^2) = (1.21 + 0.01 + 1.0)/3 = 2.22/3 = 0.74
- Y,Y = Var(Y) = E(Y^2) = (1.0 + 0.0 + 1.0)/3 = 2/3 = 0.67
- X,Y = E(XY) = (1.1 + 0.0 + 1.0)/3 = 2.1/3 = 0.7

Covariance Matrix:

- 0.74, 0.7
- 0.7, 0.67

Example Continued:

eigenvalues = 1.406, 0.004

- **dom eigenvector= (1.0, 0.95)**

sergey@cs.stanford.edu

Clustering Applications

Information Retrieval

- more of the same

Visualization

Image Segmentation

Clustering in a metric space.

We are given:

- **data points: x1,x2,x3,...**
- **distances: D(x1,x2),...**

A cluster has:

- **diameter: max D(x,y)**
- **centroid: center of mass (assuming Euclidean space)**

Goal of clustering:

- **reasonable number of clusters**
- **small diameter/small mean square distance to centroid**
- **complete coverage**
- **disjoint (???)**

Minimum Spanning Tree Clustering

1. Construct MST of data

2. Remove k-1 heaviest edges.

Problems:

- Sausage shaped clusters

sergey@cs.stanford.edu

K-means Clusterings

1. Pick k random points

2. Assign each point to closest of k points.

3. Compute centroid of each cluster to yield k new points.

4. Goto 2.

Guaranteed to give locally optimum answer.

Most commonly used.

Other clustering techniques

Agglomerative Clustering

Decomposition Clustering