Proseminar Theoretische Informatik 21.11.2017

Eigenvectors and Applications of Singular Value Decomposition

Kristin Knorr *Wolfgang Mulzer*

1 Singular Vectors and Eigenvector

Definition 1. *Let matrix B be a square matrix. Vector x is called eigenvector of B if*

$$Bx = \lambda x$$

holds for x. The corresponding λ is called eigenvalue.

Let B be A^TA then the right singular vectors v_i of A are eigenvectors of B with corresponding eigenvalue σ^2 following 3.7.
Let B be AA^T then the left singular vectors u_j of A are eigenvectors of B with corresponding eigenvalue σ^2 by the same argument.

Definition 2. *A matrix B with with property*

$$x^TBx \leq 0$$

for all x is called positive semi-definite.

Every matrix of the Form A^TA is positive semi-definite.

2 Applications of Singular Value Decomposition

2.1 Centering Data

The centering of the data is essential for for some applications of SVD. The data is centered by subtracting the centroid of the data from each point. In the case of finding the best fitting affine subspace the data is centered first.

Lemma 3. *The best-fit line (minimizing the sum of perpendicular distances squared) of a set of data points must pass through the centroid of the points.*

Addition to proof:

$$\begin{aligned}
&= \sum_{i=1}^{n} \left(\sqrt[2]{\sum_{j=1}^{m} (a_{ij} - a_j)^2}^2 - (v \cdot a_i)^2 \right) \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{m} (a_{ij}^2 - 2a_{ij}a_j + a_j^2) - (v \cdot a_i)^2 \right) \\
&= \sum_{i=1}^{n} \left(\sqrt[2]{\sum_{j=1}^{m} a_{ij}^2}^2 + \sqrt[2]{\sum_{j=1}^{m} a_j^2}^2 - \sum_{j=1}^{m} (2a_{ij}a_j) - (v \cdot a_i)^2 \right)
\end{aligned}$$

An analogous statement holds for higher dimensional objects. Define an affine space as a subspace translated by vector.

$$\{v_0 + \sum_{i=1}^{k} c_i v_i | c_, c_2, ..., c_k \in \mathbb{R}\}$$

v_0 is the translation and v_1, v_2,..., v_k form an orthonormal basis for the subspace.

Lemma 4. *The k dimensional affine space which minimizes the sum of squared perpendicular distances to the data points must pass through the centroid of the points.*

Idea of proof: $\sum_{j=1}^{k} (v_j \cdot a_i)$ instead of $\sum_{j=1}^{n} (v \cdot a_i)$ where v_j are an orthonormal basis of subspace.

2.2 Principal Component Analysis

Example Movie recommendation system:

- n customers
- d movies
- $n \times d$ Matrix A, $a_{ij} \rightarrow$ customer i likes movie j

The idea is to reduce on k basic factors with $k \ll n, d$. Movie is described as k-dimensional vector e.g. ratio of genre. Customer is described as k-dimensional vector corresponding to favourite genres. The dot-product of those determines how much a customer will like a movie. Matrix A can be written as

$$A = UV$$

where U is $n \times k$ matrix describing customers and V is $k \times d$ matrix describing movies. Best rank k approximation A_k by SVD gives U and V. The problem is A being sparse but number of $\Omega(nd)$ entries is needed. This issue is solved by collaborative filtering.

2.3 Clustering a Mixture of Spherical Gaussians

This chapter describes how two solve clustering problems by SVD. The problem is solved by mixture models.

Definition 5. *A mixture is a probability density or distribution that is weighted sum of simple component probability densities. It is of the form*

$$f = w_1 p_1 + w_2 p_2 + ... + w_k p_k$$

where p_i are basic probability densities and w_i are positive reals numbers called mixture weights.

The *model fitting problem* is to fit a mixture of k basic densities to n independent, identically distributed samples, each sample drawn according to the same mixture distribution f. Only the class of basic densities is known (here: spherical Gaussians). There are two ways to procedure:

1. Pick each sample according to the density f on $\mathbb{R}_d$.
2. Pick a random i from $\{1..k\}$ where probability of picking is w_i. Then pick a sample according to the density p_i.

1. Cluster the set of samples into k clusters $C_1, ..., C_k$ where C_i is the set of samples generated according to p_i.
2. Fit a single Gaussian distribution to each cluster of sample points. (as seen before)

The problem is the distance between clusters. Remember we need at least $cd^{\frac{1}{4}}$ distance between centers. But it is shown $\Omega(1)$ standard deviations suffice if number k of Gaussians is $O(1)$.

Lemma 6. *Suppose p is a d-dimensional spherical Gaussian with center μ and standard deviation σ. The density of p projected onto a k-dimensional subspace V is a spherical Gaussian with the same standard deviation.*

As we can see a spherical Gaussian can be projected on subspace.

Definition 7. *If p is a probability density in d space, the best-fit line for p is the line $l = \{cv_1 | c \in \mathbb{R}\}$ where*

$$v_1 = arg \max_{|v|=1} \underset{x \sim p}{E}[(v^T x)^2]$$

Lemma 8. *Let the probability density p be a spherical Gaussian with center $\mu \neq 0$. The unique best fit 1-dimensional subspace is the line passing through μ and the origin. If $\mu = 0$, then any line through the origin is best-fit line.*

Remember: $Var(X) = E((X - \mu)^2)$

But yet we have just one dimension. Now it has to be shown to work on k-dimensional subspace.

Definition 9. *If p is a probability density in d-space then the best-fit k-dimensional subspace V_k is*

$$V_k = \underset{V:dim(V)=k}{argmax} \underset{x \sim p}{E}[|proj(x, V)|^2]$$

where proj(x,v) is the orthogonal projection of x onto V.

Lemma 10. *For a spherical Gaussian with center μ, a k-dimensional subspace is a best fit subspace if and only if it contains μ.*

Theorem 11. *If p is a mixture of k spherical Gaussians, then the best fit k-dimensional subspace contains the centers. In particular, if the means of the Gaussians are linearly independent, the space spanned by them is the unique best-fit k dimensional subspace.*

Therefore we can conclude the k-dimensional SVD subspace will be close to the centers.

2.4 Ranking Documents and Web Pages

Ranking documents is an important task for collections. One possibility to evaluate relevance is a documents projection onto best-fit direction for the collection, namely the left singular vector of the term-document matrix.

2.4.1 Hyperlink-Induced Topic Search(HITS) Algorithm

HITS is an algorithm to evaluate relevance of results in websearches. There are pages called authorities which are prominent on a topic(determined by incoming links). There are pages called hubs which have many links to authorities.[1]

```
identify base set G of pages p_i
G := base set
k := natural number
z := (1,1,...,1) ∈ R^n
x := z     //vector of authority weights
y := z     //vector of hub weight
for i ∈ {1,...,k} do
  for p ∈ G
    x'_p := sum_{q|(q,p)∈E} y_q
    y'_p := sum_{q|(p,q)∈E} x_q
    x := normalize(x')
    y := normalize(y')
return(x,y)
```

x converges to right singular vector with large k. y converges to left singular vector respectively.

2.5 An Application of SVD to a Discrete Optimization Problem

This chapter describes the application of SVD to approximate solutions of optimization problems.

Definition 12. *The maximum cut problem is to partition the nodes of an n-node directed graph into to subsets S and $\bar{S}$ so that the number of edges from S to $\bar{S}$ is maximized.*

Let A be the adjacency matrix of the graph and x a vector where x_i is set to 1 for $x_i \in S$ and 0 otherwise. Therefore the maximal cut problem can be written as

$$maximize \sum_{i,j} x_i(1-a_j)a_{ij} = maximize\ x^T A(1-x)$$

This problem is NP-hard. But it can be found a near optimal solution by SVD in dense graphs ($\Omega(n^2)$ edges).

1. Replace A by $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$
2. maximize $x^T A_k(1-x)$

Theorem 13. *Given a directed graph G(V,E), a cut of size at least the maximum cut minus $O(\frac{n^2}{\sqrt{k}})$ can be computed in time polynomial in n for any fixed k.*

[1] Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (September 1999), 604-632. DOI=http://dx.doi.org/10.1145/324133.324140