CS 683: Advanced Design & Analysis of Algorithms	March 14, 2008

Lecture 24

Lecturer: John Hopcroft *Scribe: Chris Provan*

1 High Dimensional Data

This work was originally done by Gerard Salton at Cornell 20-30 years ago. The techniques are now used by Google.

1.1 Vector Space Model

Suppose we have 1 million documents that we would like to efficiently represent. How can we do this?

Compile a list of words occurring in at least one of the documents. For each document, create a frequency table:

Word	Number of Occurrences
aardvark	0
abacus	0
⋮	⋮
antitrust	42
⋮	⋮
ceo	17
⋮	⋮
microsoft	61
⋮	⋮
windows	14
⋮	⋮
zoology	0

From these tables, create a matrix $A = (a_{ij})$ where a_{ij} is the number of occurrences of word j in document i. We would like to reduce the size of A while minimizing the amount of information loss.

Let's project A onto a k-dimensional space. How do we choose the space?

(1) Choose randomly. This actually works pretty well.

(2) Choose the k-dimensional space B minimizing $|(A-B)|_F^2 = \sum_i \sum_j (A-B)_{ij}^2$, the Frobenius norm of $A - B$.

We will explore the second option.

1.2 Singular Value Decomposition

Suppose matrix C is symmetric. Then C has real-valued eigenvalues, and there exists an orthonormal matrix U such that $C = U\Sigma U^T$ where Σ is the diagonal matrix whose diagonal elements $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$ are the eigenvalues of C.

If we replace Σ with the matrix Σ_k, whose first k diagonal entries are $\sigma_1, \sigma_2, \ldots, \sigma_k$ with zeros everywhere else, then it will turn out that $B = U\Sigma_k U^T$ will minimize $|(C - B)|_F^2$ over all k-dimensional spaces B.

So take AA^T, whose ij-th element is the dot product of the rows corresponding to documents i and j. This is called the "matrix of similarities" since a larger value for a given element implies more words in common between two papers. Normalizing A so that the diagonal elements of AA^T are one would give relative similarities.

AA^T is symmetric and positive definite (i.e. $x^T Ax > 0$ for all non-zero vectors x), so the eigenvalues of AA^T are real and strictly positive. Thus we can find orthonormal U and diagonal Σ^2 and Σ such that $AA^T = U\Sigma^2 U^T = (U\Sigma)(U\Sigma)^T$, where the diagonal elements of Σ^2 are the eigenvalues of AA^T and the diagonal elements of Σ are their positive square roots.

Before we continue our analysis, let's review a few linear algebra results.

1.3 Linear Algebra Review

Let A be an $n \times n$ real matrix. If there exists a non-zero vector x and scalar λ such that $Ax = \lambda x$ then λ is an eigenvalue of A and x is a corresponding eigenvector.

For a given λ and $n \times n$ identity matrix I, $(A - \lambda I)x = 0$ gives a set of homogeneous equations. The set of equations has a non-trivial solution (and thus λ is an eigenvalue) if and only if $\det(A - \lambda I) = 0$.

$\det(A - \lambda I)$ is a degree n polynomial in λ, so it will have n not necessarily distinct roots. These roots are the eigenvalues of A. If a root is of order k

then there exists a vector space of dimension k of eigenvectors corresponding to this root. Our convention will be to normalize a basis of one of these spaces to a unit basis.

Definition 1. *Matrices A and B are* ***similar*** *if there exists an invertible P such that $A = PBP^{-1}$.*

Theorem 2. *If A and B are similar then they share the same eigenvalues.*

Proof:

$$\begin{aligned}\det(A - \lambda I) = \det(PBP^{-1} - \lambda PIP^{-1}) &= \det\left[P(B - \lambda I)P^{-1}\right] \\ &= \det P \cdot \det(B - \lambda I) \cdot \det(P^{-1}) \\ &= \det(B - \lambda I) \cdot \frac{\det P}{\det P} \\ &= \det(B - \lambda I).\end{aligned}$$

□

Definition 3. *A is* ***diagonalizable*** *if it is similar to a diagonal matrix.*

Theorem 4. *A is diagonalizable if and only if there exist n linearly independent eigenvectors of A.*

Proof: We will just prove in one direction.

Suppose A is diagonalizable. Then $A = PDP^{-1}$, and thus $AP = PD$, for some diagonal D. Let d_i be the i-th diagonal element of D and p_i be the i-th column vector of P. Then

$$[Ap_1 \quad Ap_2 \quad \cdots \quad Ap_n] = AP = PD = [d_1p_1 \quad d_2p_2 \quad \cdots \quad d_np_n],$$

where the Ap_i and d_ip_i are column vectors. So for each i, $Ap_i = d_ip_i$. Since P is invertible, its column vectors must be linearly independent and non-zero, so $p_1, p_2, \ldots, p_n$ are linearly independent eigenvectors of A. □

Note also that $\lambda_1 = \max_x x^T Ax$ is the largest eigenvalue of A and $|A|_F^2 = \sum_{i=1}^n \lambda_i^2$ where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A.