1. Linear Algebra Background

Note Title 1/20/2009

Lecture Outline

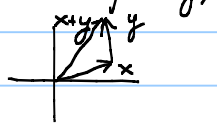
- Vectors (norms, inner products, orthogonality, linear combinations, linear independence, subspaces, basis, orthogonal basis)
- Example of mining documents (assembling document vectors into a word-document matrix)
- Matrices (norms, SVD, eigenvalues, best approximation theorem)

Vectors

A vector x has a direction and "magnitude". The latter is also called a vector's norm.

Properties of a vector norm (denoted by $\|\cdot\|$)

1. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if (iff) $x = 0$ (non-negativity)
2. $\|\alpha x\| = |\alpha| \cdot \|x\|$ (homogeneity)
3. $\|x+y\| \leq \|x\| + \|y\|$ (triangle inequality)



Most popular example of a norm:

$$\|x\|_2 = \left\| \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \ldots\ldots + |x_n|^2} = \sqrt{x^T x}$$

Other useful norms: $\|x\|_1 = |x_1| + |x_2| + \ldots + |x_n|$

$$\|x\|_\infty = \max_{i=1,2,\ldots,n} |x_i|$$

General p-norms: $\|x\|_p = \sqrt[p]{|x_1|^p + |x_2|^p + \ldots\ldots + |x_n|^p}$

Distances can be measured by norms:

For example $d(x,y) = \sqrt{(x-y)^T(x-y)} = \|x-y\|_2$ such

Triangle inequality for distances (metrics) follow from triangle inequality for norms:

$$\|x-z\| = \|(x-y)+(y-z)\| \leq \|x-y\| + \|y-z\|$$

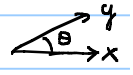
Later, we will discuss "generalized" distances given by:

$$\sqrt{(x-y)^T A (x-y)}$$

where A is a so-called positive definite matrix (discussed later, but basically it means that $z^T A z > 0$ for all $z \neq 0$.

Inner products — The most common inner product between vectors x & y is $x^T y = \sum_{i=1}^{n} x_i y_i$

Let the angle between vectors x and y be denoted by θ.



Then $x^T y = \|x\|_2 \|y\|_2 \cos\theta$.

The quantity $\cos\theta = \left(\frac{x}{\|x\|_2}\right)^T \left(\frac{y}{\|y\|_2}\right)$ is often called cosine similarity (between vectors x and y) in data mining.

Vectors x & y are orthogonal to each other if $x^T y = 0$, i.e., $\theta = 90°$.

Given two vectors x & y, $\alpha x + \beta y$ is their linear combination.

Given a collection of n-vectors $x_1, x_2, x_3, \ldots, x_n$ these vectors are said to be linearly independent if

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n = 0 \Rightarrow x_i = 0 \quad \forall i$$

A linear subspace is a set of vectors that is closed under vector addition and scalar multiplication, i.e., if x_1, x_2 belong to the subspace, then $\alpha_1 x_1 + \alpha_2 x_2$ belong to the subspace

A basis of the subspace is the maximal set of vectors in the subspace that are linearly independent of each other. An orthogonal basis is a basis where all the basis vectors are orthogonal to each other.

Question: How/where do vectors arise in data mining?

Answer: They arise whenever you can express data as a numerical combination of a fixed number of features.

For example, we can express each person superficially as a 2-d vector of "height" and "weight".

More sophisticated Example

Suppose one has a collection of text documents. A popular model for mining such text documents is the so-called vector-space model (also known as "bag of words" model). (By "mining" we mean searching through documents or trying to group them by content)

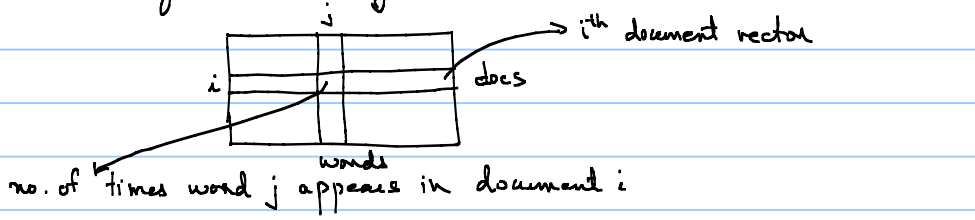
Vector-space model

1. Parse each document to extract individual words
2. Express each document as a vector, where each component of the vector corresponds to a particular word and the value given to that component is the no. of times the corresponding word appears in the document

NOTE 1: If there are a lot of documents, then the total no. of components (total no. of words in ALL the documents) will be quite large, so each individual document vector will only contain a few non-zero entries corresponding to the words that occur in that document (the zero entries correspond to the words that are absent from that document, but these words are present in other documents)

NOTE 2: Typically non-content bearing words are removed (such as "and", "the", "a", etc. – these are called "stop words")

Suppose there is a total of n words and m documents. These might be arranged as an $m \times n$ document-word matrix



EXAMPLE of text documents with words underlined (that are dimensions in the vector-space model) and the resulting matrix. Thus, matrices arise frequently in data mining (another example would be a customer-item matrix where the (i,j)-th element is the amount of the j-th item purchased by customer i.

Matrices

The "magnitude" of a matrix can be measured by a matrix norm. Properties:

1. $\|X\| \geq 0$ & $\|X\| = 0$ iff $X = 0$ - non-negativity
2. $\|\alpha X\| = |\alpha| \cdot \|X\|$ - homogeneity
3. $\|X + Y\| \leq \|X\| + \|Y\|$ - triangle inequality

Some matrix norms have an additional submultiplicative property:

4. $\|XY\| \leq \|X\| \cdot \|Y\|$

An $m \times n$ matrix can be treated as a vector in mn-dimensional space, and the corresponding vector norms result in matrix-norms that satisfy Properties 1-3 above. The most common such norm is the Frobenius norm: $\|X\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}|^2}$.

Another important class of matrix norms is obtained by viewing a matrix as an operator or linear transformation, yielding the so-called operator norm or induced norm. Suppose $X \in \mathbb{R}^{m \times n}$, then $X : \mathbb{R}^n \to \mathbb{R}^m$ maps $y \in \mathbb{R}^n$ to $Xy \in \mathbb{R}^m$.

$\|X\| = \sup_{\substack{\|y\| \leq 1 \\ y \neq 0}} \frac{\|Xy\|}{\|y\|}$, where $\|y\|$ & $\|Xy\|$ are vector norms in the appropriate spaces.

The most common such norm is the induced 2-norm:

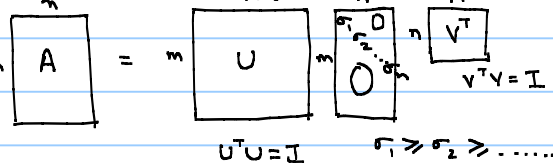
$$\|X\|_2 = \max_{\substack{\|y\|_2 \leq 1 \\ y \neq 0}} \frac{\|Xy\|_2}{\|y\|_2}$$

SVD of a matrix

Every matrix has a so-called singular value decomposition (SVD). This remarkable decomposition yields information about the four fundamental subspaces associated with a matrix A:

- column or range space of A denoted by $\mathcal{R}(A)$
- row space of A denoted by $\mathcal{R}(A^T)$
- null space of A denoted by $\mathcal{N}(A)$
- null space of A^T denoted by $\mathcal{N}(A^T)$

SVD: Every matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed as $A = U \Sigma V^T$, (with non-negative numbers on the diagonal) where $U^T U = I$, Σ is diagonal, and $V^T V = I$



If $\text{rank}(A) = r$, then $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ & $\sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_n = 0$

SVD may be written as $AV = U\Sigma$

or $A[v_1 v_2 \dots v_n] = [u_1 u_2 \dots u_n | u_{n+1} \dots u_m]\ \Sigma$

or $Av_i = u_i \sigma_i$ for $i = 1, 2, \dots, r$

Also, $A^T = V\Sigma^T U^T \Rightarrow A^T U = V\Sigma^T$ & $Av_i = 0$ for $i = r+1, \dots, n$

or $A^T [u_1 u_2 \dots u_n | u_{n+1} \dots u_m] = [v_1 v_2 \dots v_n]\ \Sigma^T$

or $A^T u_i = v_i \sigma_i$, for $i = 1, 2, \dots r$ & $A^T u_i = 0$ for $i = r+1, \dots, m$.

$A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$

$v_1 \longmapsto u_1\sigma_1$
$v_2 \longmapsto u_2\sigma_2$
$\vdots$
$v_r \longmapsto u_r\sigma_r$
$v_{r+1} \longmapsto 0$
$\vdots$
$v_n \longmapsto 0$

$A^T : \mathbb{R}^m \longrightarrow \mathbb{R}^n$

$u_1 \longmapsto v_1\sigma_1$
$u_2 \longmapsto v_2\sigma_2$
$\vdots$
$u_r \longmapsto v_r\sigma_r$
$u_{r+1} \longmapsto 0$
$\vdots$
$u_m \longmapsto 0$

$R(A) = \langle u_1, u_2, \dots, u_r \rangle$

$R(A^T) = \langle v_1, v_2, \dots, v_r \rangle$

$N(A) = \langle v_{r+1}, \dots, v_n \rangle$

$N(A^T) = \langle u_{r+1}, \dots, u_m \rangle$

$\mathbb{R}^n = R(A^T) \oplus N(A)$, and $\mathbb{R}^m = R(A) + N(A^T)$

Best Approximation Property

Given a matrix A, its rank-k truncated SVD A_k has the following important best approximation property:

The rank-k truncated SVD A_k minimizes the reconstruction error among all rank-k matrices

$$\min_{B \text{ of rank } = k} \| A - B \|_2$$ has solution: $B = A_k$

The theorem holds when the 2-norm above is replaced by the Frobenius norm.