CS630 Lecture Notes

Lecturer: Lillian Lee
Scribes: Chris Danis (cgd3) & Brian Rogan (bcr6)

Lecture 16: 30 March 2006

1 Introduction

Today we will be covering matrix-theoretic corpus characterizations; in particular, we're building up to the singular value decomposition, or SVD.

2 Recap of last lecture

Last time we covered using clickthrough data as implicit preference information; "attribute" vector spaces, where each feature (variable) corresponds to an axis in the space, were used to store information about query-document pairs. We were given, based on clickthrough data, query-specific constraints (e.g. "d is preferred over d' for q"). Our goal was to find a vector $\vec{w}$ that represented a preferred "direction" with respect to the queries that are given. Points further along $\vec{w}$ should be scored higher.

In our diagram, $\vec{w}$ points up in y and left (negative) in x. Why is negative the preferred direction for x? What might the x-axis be in this case? It can't be term frequencies; maybe it's a binary attribute represented as 1/-1. It could also be number of Finnish terms - number of English terms. The y-axis could be a lot of things, including (for instance) $\cos(\angle(\vec{q}, \vec{d}))$.

Do corpora have preferred directions *without* being given query information? Do there exist "inherent" corpus properties? If we can find these properties, we can potentially get better application-independent representation of data.

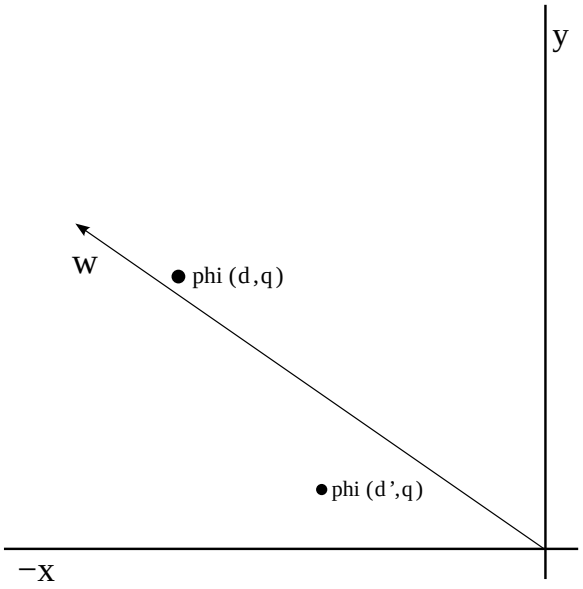

Figure 1: Illustration of $\vec{w}$ with d preferred over d' for q

3 Characterizing corpora

Assume we have some sort of feature-document matrix. In IR-speak, this would be a term-document matrix.

Assume we have n documents $d^{(1)}, \ldots, d^{(n)}$, and a m-term vocabulary.

Call D the term-document matrix; $D = \begin{pmatrix} | & & | \\ \vec{d}^{(1)} & \cdots & \vec{d}^{(n)} \\ | & & | \end{pmatrix}$. (Notational conventions: superscripts = which matrix column; subscripts = which vector component.)

What might we care about in terms of characterizing the corpus? One useful characteristic is the "spread" of document vectors – are they highly similar, and so, point in similar directions, or are the vectors spread out? This is a reasonable measure of how varied the corpus is. One idea for measuring spread: compute $\text{span}(\{\vec{d}^{(1)}, \cdots, \vec{d}^{(n)}\})$; take its dimensionality as a measure of "complexity". This quantity (the dimensionality of $\text{span}(\{\vec{d}^{(1)}, \cdots, \vec{d}^{(n)}\})$) is called $\text{rank}(D)$. An aside: weirdly, $\text{rank}(D) = \text{rank}(D^T)$. While this is easily proved for matrices in general (via the SVD), it is initially hard to understand intuitively for a term-document matrix.

3.1 Special case: $\text{rank}(D) = 1$

Suppose $\text{rank}(D) = r = 1$.

For each $\vec{d}^{(i)}$, we know $\vec{d}^{(i)} \in \text{span}(\{\vec{d}^{(1)}, \cdots, \vec{d}^{(n)}\})$.

If $r = 1$, then $\exists \vec{b} \in \mathbb{R}^n$ such that for each $\vec{d}^{(i)}$, $\vec{d}^{(i)} = \alpha_i \vec{b}$, with $\alpha_i \in \mathbb{R}$.

If we require $||\vec{b}||_2 = 1$, then $\vec{b}$ is unique up to sign.

In this case, $\vec{b} \approx$ the underlying term-distribution profile of this corpus.

3.2 General case: $rank(D) > 1$

We know that $\text{rank}(D) = r \leq \min(m, n)$ (n documents and m terms).

There exist vectors $\vec{b}^{(1)}, \cdots, \vec{b}^{(r)} \in \mathbb{R}^m$, with the $\vec{b}^{(i)}$'s orthonormal (all are mutually orthogonal and of unit length in 2-norm).

For each $\vec{d}^{(i)}$, there exist $\alpha_1^{(i)}, \cdots, \alpha_r^{(i)} \in \mathbb{R}$, such that

$$\vec{d}^{(i)} = \sum_{l=1}^{r} \alpha_l^{(i)} \vec{b}^{(l)} \tag{1}$$

We cannot say, in this general case or $r > 1$, that we can find a unique bases. There are an infinite number of such basis sets. Thus, this way of characterizing a corpus (rank) gives us some information but not as much as we'd like. What can we do to get more information in a succinct fashion?

3.3 Other approaches

The span of D is all linear combinations of the $\vec{d}^{(i)}$'s. If we restrict ourselves to only certain types of linear combinations, can we capture more information?

Linear combinations like equation 1 can, it turns out, be written as a matrix-vector multiply:

$$\sum_{i=1}^{n} \alpha_i \vec{d}^{(i)} = \begin{pmatrix} | & | & & | \\ \vec{d}^{(1)} & \vec{d}^{(2)} & \cdots & \vec{d}^{(n)} \\ | & | & & | \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$$

Here we're applying $D \in \mathbb{R}^{m \times n}$ as an operator to a coefficient vector $\vec{\alpha} \in \mathbb{R}^n$. This seems very weird conceptually (using a corpus as a mathematical operator), but it's perfectly alright algebraically.

Consider the case of restricting α_i's to be fractional assignments. All this means is that we require $\sum_{i=1}^{n} \alpha_i = 1$, and that each $\alpha_i \geq 0$. A pictorial example of this for three dimensions is shown in figure 2.

A small example: $D = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}$; three documents and two terms. We'll call our first "corpus" D_I; it has a pretty big spread. Fractional-assignment combinations of the of vectors give you the convex hull of the set of vectors. In this case, we get a triangle with vertices corresponding to each $\vec{d}^{(i)}$ (figure 3).

Our next example is D_{II}: three vectors with not a lot of spread. Applying this to α, we get a smaller triangle (figure 4).

At this point, we form a hypothesis, based on our experiences with D_I and D_{II}: the area of this triangle can measure spread. Let's try an additional case.

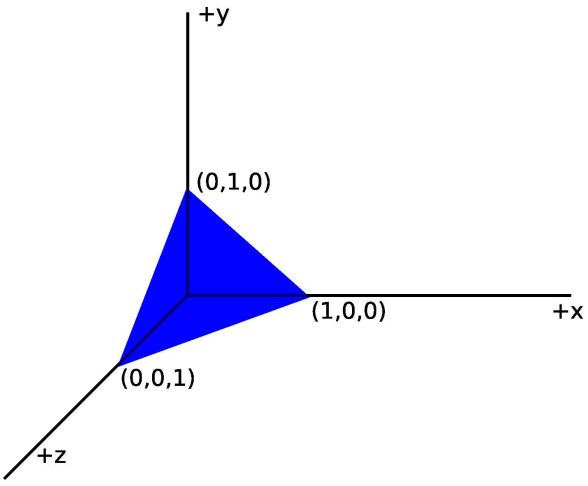

Figure 2: Fractional assignment triangle in $\mathbb{R}^3$

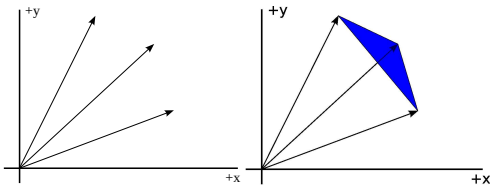

Figure 3: Corpus D_I (left) and $D_I\alpha$ (right)

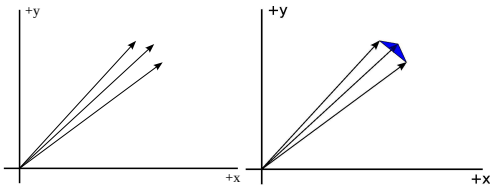

Figure 4: Corpus D_{II} (left) and $D_{II}\alpha$ (right)

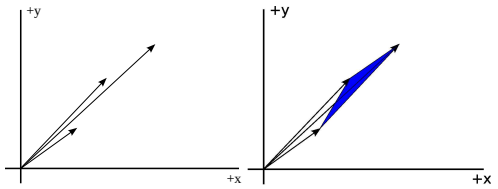

Figure 5: Corpus D_{II} (left) and $D_{II}\alpha$ (right)

D_{III} is a corpus with little spread but with large variation in length. Application to α results in a stretched-out triangle of large area (figure 5). Unfortunately, our hypothesis didn't work out.

There's *something* interesting and potentially useful going on here, but we can't quite get our fingers on it. Next lecture, we'll change the kind of linear combinations we allow (by imposing different constraints related to L_1 and L_2 normalization).

CS630 Lecture Problems

Lecturer: Lillian Lee
Scribes: Chris Danis (cgd3) & Brian Rogan (bcr6)

Lecture 16: 30 March 2006

1. Recall that in class we discussed the matrix D which was our feature-document matrix. Each column corresponded to a document, and each row to a feature.

   (a) Give an intuitive idea about what $rank(D)$ is.

   ANSWER:
   Recall that in lecture we talked about $rank(D)$ as the "spread" of the documents. Thus the *rank* gives us an idea of how different or similar the documents are from one another in feature space. For more discussion of this, see the answer to the next question.

   (b) What does it mean to say that $rank(D) < min(m, n)$?

   ANSWER:
   This means that the documents are not linearly independent. The intuition behind what this means is less than clear, but it means that fewer basis vectors are required than there are features (assuming $m \leq n$), which would indicate some sort of overlap between the documents such that they can be described as linear combinations of a few feature patterns. In the case where $n < m$, it indicates that fewer basis vectors are needed then documents, which (given that there are fewer documents then features) also implies some sort of overlap between documents. Again, its not entirely clear what the basis vectors mean (we still hadn't arrived at a definite answer to that by the end of Lecture 17[1]), but its fair to say that the documents have some overlap and are not completely different.

   (c) What does it mean for us to say that $rank(D) = 1$ where D is a term-document matrix, rather than the more general form of a feature-document matrix?[2]

[1] ED: Indeed the fact that we don't have a single unique set of basis vectors in general indicates that assigning an intuitive interpretation might be problematic.
[2] ED: This question was posed in lecture.

ANSWER:
If D is simply a term-document matrix, this implies that there exists some $b \in \Re^m$ such that for all $d^{(i)}$,$d^{(i)} = \alpha^{(i)} b$ for some $\alpha^{(i)}$. What this means in practical terms is that each document contains exactly the same terms in exactly the same proportions, but some are longer than others. One example of this would be that $d^{(2)}$ is simply $d^{(1)}$ copied over several times. Another example would be that $d^{(2)}$ simply replaces each word in $d^{(1)}$with that word repeated several times. Obviously there are an infinite number of variations on this theme, but the most important point is this: the documents are indistinguishable in term space except for length.

(d) What does it mean in general for $span\{d^{(1)} \dots d^{(n)}\} \subset \Re^m$? Provide an example of how this could happen when D is a term-document matrix. Provide an intuition in the more general case where D is a feature document matrix.

ANSWER:
When D is a term-document matrix, it may simply be that none of the documents in D contain a specific term. In this case, there is no way for any $d^{(i)}$ to provide a way to "reach into" a certain dimension of $\Re^m$. Note that in the general case this could also indicate redundant features: if two features are always perfectly linearly correlated, it will be that the rows of D are not linearly independent and it may be that the documents don't span $\Re^m$ (note that when $m = n$ this is guaranteed because D cannot be inverted). Also, the earlier observations from part b apply, because the column vectors of D will not span $\Re^m$ if $rank(D) < m$ (assuming $n \geq m$).

2. Why do we say that, in the general case, there are an infinite number of possible choices for basis vectors for D? Could we place restrictions on our choices of the basis vectors such that there would be one distinct set?

ANSWER:
There are an infinite number of choices because we have not placed any restrictions on the length of the vectors. We can always make a new set of basis vectors from the old vectors by simply multiplying the basis vectors by two and halving all of our $\vec{a}^{(i)}$. Even if we specify that the basis vectors be orthonormal, we can always rotate the basis vectors, so there is no way to make these vectors distinct.

3. Recall that in class, we looked at what happened to several document matrices D, under the computation: $D\vec{\alpha}$.

(a) We claimed that the computation $D\vec{\alpha}$ was the convex hull of the document vectors. On superficial examination, it appears that $D\vec{\alpha}$ is actually a point (it is of dimension $m \times 1$). How is that we can thus

say that this quantity is actually the convex hull?

ANSWER:
The convex hull is obtained by multiplying every $\vec{\alpha}$ such that $\sum \alpha_l = 1, \alpha^{(l)} \geq 0 \forall l$ (recall in class that we called this the set of "fractional assignment" α's) by D, not just one specific $\vec{\alpha}$. Another way of looking at this is that for every point p in the convex hull of the column vectors of D, there exists some $\vec{\alpha}$ such that $\sum_l \alpha_l = 1, \alpha_l \geq 0 \forall l$ and $p = D\vec{\alpha}$.

(b) Recall in class that we discussed how it seemed strange that we were multiplying D by $\vec{\alpha}$, because we could understand this as applying a function (D) to $\vec{\alpha}$. How does the intuition from part (a) support the idea that really D provides parameters to a function?

ANSWER:
Consider the fact that the convex hull C is simply all possible linear combinations of the $d^{(i)}$ subject to the fractional assignment restriction. We can thus see $D\vec{\alpha}$ as a function which actually acts on the columns of D, rather than than acting on $\vec{\alpha}$. Thus, we can understand that $\vec{\alpha}$ varies freely, taking D as a function parameter that specifies what $\vec{\alpha}$ can operate on.