COMP1046 Tutorial 3 : Vector Space Models

Anthony Bellotti

A Vector Space Model is a way to represent documents numerically. Each document is coded as a vector of n words. Each component of the vector is the frequency of a particular word in the document. Typically, connecting words such as "and" and "the" are ignored and "s" is removed from ends of words. For example, consider this document,

> *The best kinds of people are warm and kind,*
> *They are always there and they never mind.*
> *The best kinds of people smile and embrace,*
> *They support you with strength and grace.*
> copyright "Alex"

I can code it using the word list ("always", "best", "butter", "embrace", "grace", "kind", "never", "people", "smile", "strength", "support", "there", "they", "vector", "warm") as the vector

$$(1, 2, 0, 1, 1, 3, 1, 2, 1, 1, 1, 1, 1, 0, 1).$$

If my domain language has n words, I can express the space of all documents by the vector space $(\mathbb{R}^n, +, .)$. Some observations:-

- In reality, n will be large (number of words in a language) and each document vector will be *sparse*, meaning most components will be zero.
- Using real numbers $\mathbb{R}$ allows for fractions and negative numbers, but this will be useful if we want to express comparisons between documents.

Answer the following questions about this vector space.

1. Consider the word list ("algebra", "cool", "cook", "learning", "linear"), so $n = 5$, and the following three documents:
   - *We are learning linear algebra.*
   - *Linear algebra is cool.*
   - *I am learning to cook.*

   (a) Code each document as vectors $\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}$ respectively.

   (b) Compute $\mathbf{d_1} - \mathbf{d_2}$ and $\mathbf{d_3} - \mathbf{d_1}$.

(c) Are $\mathbf{d_1} - \mathbf{d_2}$ and $\mathbf{d_3} - \mathbf{d_1}$ in the vector space $(\mathbb{R}^n, +, .)$?

(d) If I introduce a $\mathbf{d_4}$ so that $\mathbf{d_3} - \mathbf{d_4} = \mathbf{o}$, what does this mean about these two documents?

(e) Give an example of a sentence, different from the ones above, that codes as $\mathbf{d_4}$ with the property $\mathbf{d_3} - \mathbf{d_4} = \mathbf{o}$.

(f) Compute $(\mathbf{d_1} - \mathbf{d_2}) \cdot (\mathbf{d_1} - \mathbf{d_2})$ and $(\mathbf{d_3} - \mathbf{d_1}) \cdot (\mathbf{d_3} - \mathbf{d_1})$. How could you interpret the values you have computed?

Answer:

(a) $\mathbf{d_1} = (1, 0, 0, 1, 1), \mathbf{d_2} = (1, 1, 0, 0, 1), \mathbf{d_3} = (0, 0, 1, 1, 0)$.

(b) $\mathbf{d_1} - \mathbf{d_2} = (0, -1, 0, 1, 0)$ and $\mathbf{d_3} - \mathbf{d_1} = (-1, 0, 1, 0, -1)$.

(c) Yes.

(d) It means that the documents have the same word count, but they may be in a different order.

(e) *The cook is learning.*

(f) $(\mathbf{d_1} - \mathbf{d_2}) \cdot (\mathbf{d_1} - \mathbf{d_2}) = 2$ and $(\mathbf{d_3} - \mathbf{d_1}) \cdot (\mathbf{d_3} - \mathbf{d_1}) = 3$. These can be interpreted as measures of dissimilarity.

2. Let U be the set of vectors representing documents that have the same proportion of words as a given document represented by the vector $\mathbf{d_1}$. The actual proportion may be different for different documents.

(a) How can the condition that a document represented by vector $\mathbf{d_2}$ has the same proportion of words as $\mathbf{d_1}$ be expressed in vector notation?

(b) Give a formal definition of U using set notation.

(c) Show whether or not U is a vector subspace of $(\mathbb{R}^n, +, .)$.

Answer:

(a) $\mathbf{d_1} = \lambda \mathbf{d_2}$ for some $\lambda \neq 0$.

(b) $U = \{\mathbf{d_2} \in \mathbb{R}^n \mid \exists \lambda \neq 0 : \mathbf{d_1} = \lambda \mathbf{d_2}\}$.

(c) Since $U \subset \mathbb{R}$, just need to prove closure of internal and external composition:

- Take arbitrary $\mathbf{x}, \mathbf{y} \in U$. Then

$$\mathbf{d_1} = \lambda_1 \mathbf{x} \Rightarrow \frac{1}{\lambda_1} \mathbf{d_1} = \mathbf{x}$$

and

$$\mathbf{d_1} = \lambda_2 \mathbf{y} \Rightarrow \frac{1}{\lambda_2} \mathbf{d_1} = \mathbf{y}.$$

Therefore, adding together,

$$\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right)\mathbf{d_1}=\mathbf{x}+\mathbf{y}\Rightarrow\mathbf{d_1}=\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}(\mathbf{x}+\mathbf{y})$$

which shows $\mathbf{x}+\mathbf{y}\in U$ demonstrating closure for internal composition.

- Take arbitrary $\mathbf{x}\in U$ and scalar $\mu\in\mathbb{R}$. Then

$$\mathbf{d_1}=\lambda\mathbf{x}\Rightarrow\mathbf{d_1}=\left(\frac{\lambda}{\mu}\right)(\mu\mathbf{x})$$

which shows $\mu\mathbf{x}\in U$ demonstrating closure for external composition.

3. Let $\mathbf{1_j}$ denote a vector of all zeroes except the jth component which is 1.
For example, with $n=4$, $\mathbf{1_3}=(0,0,1,0)$.
The vectors $\mathbf{1_1},\cdots,\mathbf{1_n}$ form a basis for $(\mathbb{R}^n,+,.)$.

   (a) What scalars on the basis vectors are needed to express an arbitrary vector $(x_1,\cdots,x_n)\in\mathbb{R}^n$?

   (b) Suppose that the 1st and 2nd words commonly appear together in documents. It is then proposed to replace $\mathbf{1_2}$ with $\mathbf{b}=\mathbf{1_1}+\mathbf{1_2}$ in the basis. Show that this new set of vectors is still a basis, and show the scalars needed to express an arbitrary vector $(x_1,\cdots,x_n)\in\mathbb{R}^n$ in terms of the new basis.

Answer:

(a) $\lambda_1=x_1,\cdots,\lambda_n=x_n$.

(b) Need to show that the set of vectors span $\mathbb{R}^n$ and is linearly independent:

- Taking an arbitrary vector $(x_1,\cdots,x_n)\in\mathbb{R}^n$ and choosing $\lambda_i=x_i$ for all $i\neq 2$ and $\lambda_2=x_2-x_1$, the linear combination of basis vectors forms $(x_1,\cdots,x_n)$ and hence spans $\mathbb{R}^n$. In particular, the 2nd component is $\lambda_1+\lambda_2=x_1+(x_2-x_1)=x_2$.
- For linear independence, check when

$$\lambda_1\mathbf{1_1}+\lambda_2\mathbf{b}+\cdots+\lambda_n\mathbf{1_n}=\mathbf{o}.$$

Rewrite as

$$\lambda_1\mathbf{1_1}+\lambda_2(\mathbf{1_1}+\mathbf{1_2})+\cdots+\lambda_n\mathbf{1_n}=\mathbf{o}$$
$$\Rightarrow(\lambda_1+\lambda_2)\mathbf{1_1}+\lambda_2\mathbf{1_2}+\cdots+\lambda_n\mathbf{1_n}=\mathbf{o}.$$

Now observe that if $\lambda_i\neq 0$ for any $i\geq 2$ then this would mean a non-zero outcome, hence $\lambda_i=0$ for all $i\geq 2$. Then,

$$\lambda_1\mathbf{1_1}=\mathbf{o}$$

which means $\lambda_1=0$ also. Therefore this set of vectors is linearly independent.

- To represent an arbitrary $(x_1, \cdots, x_n) \in \mathbb{R}^n$, write

$$\lambda_1 \mathbf{1_1} + \lambda_2 \mathbf{b} + \cdots + \lambda_n \mathbf{1_n} = (x_1, \cdots, x_n).$$

  Then $\lambda_3 = x_3, \cdots, \lambda_n = x_n$ since only $\mathbf{1_j}$ contributes to the jth element of the vector with a non-zero term, for $j > 2$.
  Also $\lambda_2 = x_2$ since only $\mathbf{b}$ contributes to the 2nd element of the vector with a non-zero term.
  Then for the first element, both $\mathbf{1_1}$ and $\mathbf{b}$ take value 1 hence
  $\lambda_1 + \lambda_2 = x_1 \Rightarrow \lambda_1 = x_1 - x_2$.