Vector Embeddings

We have been using vectors to represent inputs and outputs.

eg. "2" $= [0\ 0\ 1\ 0\ \cdots] \in \{0,1\}^{10}$

"e" $= [0\ 0\ 0\ 0\ 1\ 0\ \cdots] \in \{0,1\}^{26}$

What about words?

Consider the set of all words encountered in a dataset. We will call this our vocabulary.

Let's order and index our vocabulary and represent words using one-hot vectors, like above. Let $word_i$ = i^{th} word in vocab.

ie. "cat" $\sim v \in \mathbb{W}$

$\mathbb{W} \subset \{0,1\}^{N_v} \subset \mathbb{R}^{N_v}$ where N_v is the # of words in our vocab (eg. 70 000)

Then $v_i = \begin{cases} 0 & \text{if } word_i \neq \text{"cat"} \\ 1 & \text{if } word_i = \text{"cat"} \end{cases}$

This is nice, but when we are doing Natural Language Processing (NLP), how do we handle the common situation in which different words can be used to form a similar meaning?

Example:

"CS 489 is interesting"

"CS 489 is fascinating"

We could form synonym groups, but where do we draw the line when words have similar, but not identical, meanings?

eg. content, happy, elated, ecstatic

These issues reflect the semantic relationships between words. We would like to find a different representation for each word, but one that also incorporates their semantics.

Predicting Word Pairs

We can get a lot of information from the simple fact that some words often occur together (or nearby) in sentences.

Example:

"Trump returned to Washington Sunday night, though his wife Melania Trump stayed behind in Florida."

From <http://www.cbc.ca/news/world/stormy-daniels-trump-threat-1.4594060>

"Human activity is degrading the landscape, driving species to extinction and worsening the effects of climate change"

From <http://www.cbc.ca/news/thenational/national-today-newsletter-russia-diplomats-biodiversity-1.4592950>

For the purposes of this topic, we will consider "nearby" to be within words.

Example: $d=2$

"Trump returned to (Washington Sunday $\boxed{\text{night,}}$ though his) wife Melania Trump stayed behind in Florida."

This gives us the word pairings:

(night, Washington), (night, Sunday), (night, though), (night, his)

Example:

Source Text	Training Samples
The quick brown fox jumps over the lazy dog. →	(the, quick) (the, brown)
The **quick** brown fox jumps over the lazy dog. →	(quick, the) (quick, brown) (quick, fox)
The quick **brown** fox jumps over the lazy dog. →	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
The quick brown **fox** jumps over the lazy dog. →	(fox, quick) (fox, brown) (fox, jumps) (fox, over)

Our approach is to try to predict these word co-occurrences using a 3-layer neural network.

- its input is a one-hot word vector, and
- its output is the probability of each word's co-occurrence.

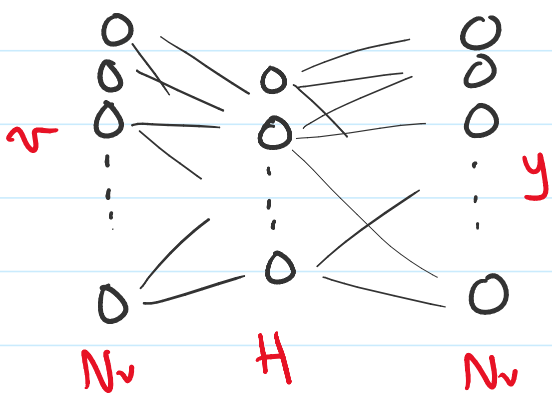
Our neural network performs

$$y = f(v, \theta) \quad \text{where } v \in \mathbb{W}$$

and $$y = P^{N_v} = \{ p \in \mathbb{R}^{N_v} \mid p \text{ is a probability vector} \}$$

$$\text{i.e. } \sum_i p_i = 1, \quad p_i \geq 0 \ \forall i$$

Then, y_i equals the probability that word$_i$ is nearby v.



output layer uses

The hidden layer is much smaller.

This hidden-layer squeezing forces a compressed representation, requiring similar words to take on similar representations.
This is called an embedding.

word2vec

Word2vec is a popular embedding strategy for words (or phrases, or sentences). It uses additional tricks to speed up the learning.

1) Treats common phrases as new words. eg. "New York" is one word
2) Randomly ignores very common words
   eg. "the car hit the post on the curb"

Of the 56 possible word pairs, only 20 don't involve "the"

3) Negative Sampling

Backprops only some of the negative cases

The embedding space is a relatively low-dimensional space where similar words are mapped to similar locations.

Where have we seen this before?

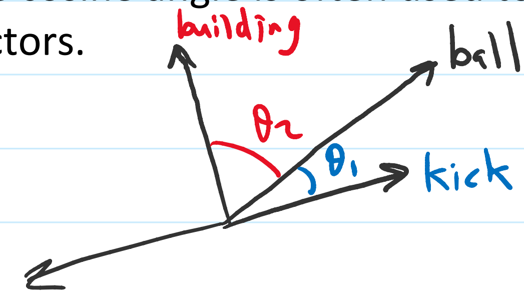
SOM

Why does this work?

Words with similar meaning likely co-occur with the same set of words, so the network should produce similar outputs.

$\therefore$ similar hidden-layer activation

The cosine angle is often used to measure the "distance" between two vectors.



$\cos\theta_1 = 0.81$

$\cos\theta_2 = -0.05$

To some extent, you can do a sort of vector addition on these representations.

eg. king − man + woman = queen