Distributed representations

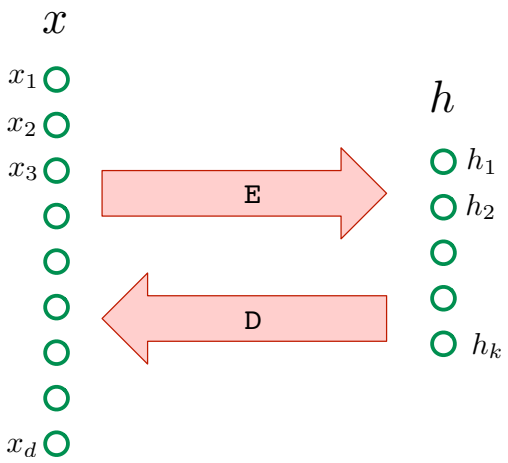
Sanjoy Dasgupta

University of California, San Diego

Topics we'll cover

1. One-hot versus distributed encodings
2. Word embeddings

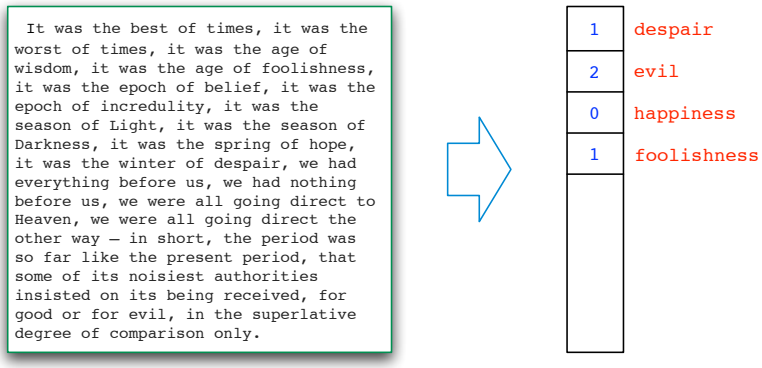
One-hot versus distributed representations



- k-means: **one-hot** encoding
- PCA: **distributed** encoding

The bag-of-words representation

One-hot encoding of words:



- Fix V = some vocabulary.
- Treat each sentence (or document) as a vector of length $|V|$:

$$x = (x_1, x_2, \ldots, x_{|V|}),$$

where x_i = # of times the ith word appears in the sentence.

Word co-occurrences

You shall know a word by the company it keeps. (J.R. Firth, 1957)

- Much of the meaning of a word w is captured by the words it co-occurs with:

$$w_1 \quad w_2 \quad w_3 \quad w \quad w_4 \quad w_5 \quad w_6$$

- Find an embedding of words based on these co-occurrences.

A simple approach to word embedding

Fix a vocabulary V. Then, using a corpus of text:

1. Look at each word w and its surrounding *context*: $w_1 \quad w_2 \quad w_3 \quad w \quad w_4 \quad w_5 \quad w_6$
   - $n(w, c) = \#$ times word c occurs in the context of word w
   - Yields a probability distribution $\Pr(c|w)$.

2. Positive pointwise mutual information:

$$\Phi_c(w) = \max\left(0, \log \frac{\Pr(c|w)}{\Pr(c)}\right)$$

   This is a $|V|$-dimensional representation of word w.

3. Reduce dimension using PCA.

The embedding

- Which word's vector is closest to that of `Africa`?
  `Asia`

- Solving analogy problems: `king` is to `queen` as `man` is to ?
  - $\text{vec}(\texttt{king}) - \text{vec}(\texttt{queen}) = \text{vec}(\texttt{man}) - \text{vec}(?)$
  - $\text{vec}(?) = \text{vec}(\texttt{man}) - \text{vec}(\texttt{king}) + \text{vec}(\texttt{queen})$
  - Nearest neighbor of this vector is $\text{vec}(\texttt{woman})$.