COMP5046: Latent Feature Spaces

Joel Nothman

joel.nothman@sydney.edu.au

School of Information Technologies
University of Sydney

2018-04-24

Vectors for representing meaning

- A bag of words represents a document's meaning
- The contexts a word appears in represents a word's meaning
- Engineered features represent the syntactic context for POS tagging

- These features provide very specific information
- sparsely
- in high dimension

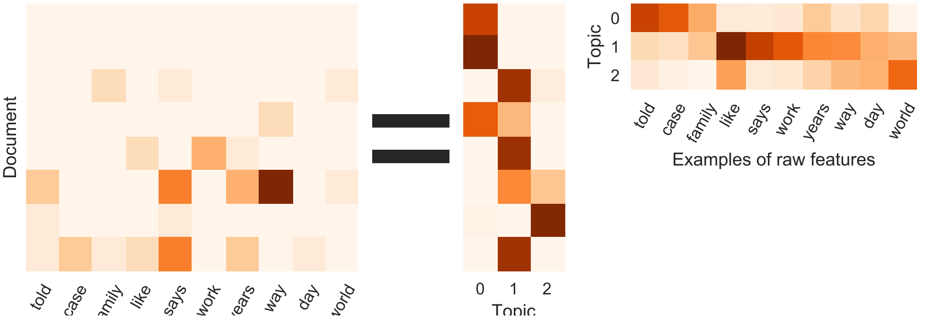
Compressing a feature vector

- Assume there is a **latent** lower-dimensional space
- that preserves the locality of the high dimensional space
- i.e. similar low-dimensional vectors represent content with similar meanings
  compressing high dimensional space while preserving locality implies that similar vectors in high dimension will have the same meaning as vectors in low dimension - smoothing and clustering the HD feature space to a lower dimension
- If we can **compress** the high-dimensional space into, say, 150 features
- it must be keeping the most important aspects of meaning
- and effectively **smoothing and clustering** the high-dim. knowledge
  - low-dim. features combine similar high-dim. features
- (NB the compression is *lossy*: you can't perfectly recover the original representation)

Principal component analysis

- first component accounts for as much variability as possible
- second accounts for some remaining variability, etc...
- take first m' components
- for term-document matrix, known as Latent Semantic Analysis

Factorising or decomposing a term-document matrix



Topic modelling with latent dirichlet allocation (LDA)

- a compression of the term-document matrix
  - like probabilistic latent semantic analysis
  - integer vectors only
- an alternative notion of language model
- topic is probability distribution over words
- document is a bag of words drawn from a mixture of topics
- top words in 20-topic representation of 10k Wikipedia people:
  - football coach played american college norwegian national head
  - born german family studied became moved later von
  - assembly scored court bar van goals smith seat
  - championships silver american world first team los won
  - member served elected politician party united president general
  - king one known son fc name also chinese
  - born english john british educated son william london

disadvantages of LDA:
- exchanging words with topics, hence loosing information
- topics arn't always meaningful or interpretable
- results are not consistent - no distinct/optimum topic assignment
- assumptions made are false

Word embeddings

unsupervised

- Lexical meaning in $\mathbb{R}^{150}$
- `word2vec` (Mikolov et al., 2013)
  - Initially reported as deep learning:
    learn a word representation that can predict the contexts it appears in
  - = matrix factorization of PMI weights (Levy and Goldberg, 2014)
  - popular variants such as GloVe (Pennington et al., 2014)
- A word's nearest neighbors are related
  - One visualisation or another to play with at home
  - frog ≈ frogs ≈ toad ≈ litoria ≈ leptodactylidae
- regularities
  - queen − king ≈ woman − man
  - 33186 − Miami ≈ 95823 − Sacramento
  - dark − darker ≈ strong − stronger
  - darker − darkest ≈ stronger − strongest

Exploiting linguistic knowledge from large corpora

- We can learn vector embeddings from large unlabelled corpora
- These become a lexical resource (like WordNet, etc.), or can be retrained

- Build a classifier of events $\rightarrow \{\text{live music}, \text{other}\}$
- Manually label a sample of events
- We expect to see guitar lots, and theorbo not so much
- Bag of words is unlikely to learn that theorbo is a musical instrument
- but hopefully they are embedded similarly

Pros and cons of latent feature spaces

- Can exploit linguistic knowledge from large corpora
- Fewer parameters to learn when classifying
- Efficient spatial data structures (e.g. kd-tree) when searching/clustering
- Can still be extended with engineered features

- May be difficult to learn a good, general-purpose embedding
- Very hard to interpret what any particular feature means
- Specific aspect of language you need to model may not be represented
- May not be simple to combine representations meaningfully

Take away

- Features do not need to be understandable to be effective
  - we already saw that with, e.g. arbitrary weights on 1-char prefix features in POS tagging
- Learning word embeddings allows us to transfer knowledge from unlabelled, in-domain text
- LDA represents document meaning in low-dimensional (topic) space
- word2vec represents word meaning in low-dimensional space