Lecture 15 - Finding Independent Features

Summary

In the last lecture we saw how the SVD could be used in the context of Latent Semantic Analysis (LSA) to decompose documents (and terms) into sets of concepts. The problem with this approach was that the concepts were not necessarily interpretable by a human. Topic modelling extends the idea of LSA, but works to ensure that the concepts (henceforth called "topics") are actually meaningful and interpretable.

In this lecture we'll look at alternate approaches to extracting topics from a corpus of documents: namely a linear-algebraic approach called Non-Negative Matrix Factorisation, and an ensemble of probabilistic topic modelling approaches, including Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation.

Key points

Topic Modelling

- When we looked at LSA, we saw that it created concepts that were linear mixtures of words
  - But, the weightings were unconstrained, and could be negative
  - Very difficult to interpret or give semantic meaning to the topic
- Would be really nice if we could determine thematic topics for a corpus of documents
  - **Topic models uncover the hidden thematic structure in document collections.**
  - Simple intuition: **documents exhibit multiple topics**
- Many different ways to achieve this:
  - Broadly fall into probabilistic approaches
    - **Probabilistic Latent Semantic Analysis (PLSA)**
    - **Latent Dirichlet Allocation (LDA)**
    - Pachinko allocation (PAM)
    - …
  - and non-probabilistic approaches:
    - **Non-Negative Matrix Factorisation ([N]NMF)**
    - …

Non-Negative Matrix Factorisation (NMF)

- NMF: an unsupervised family of algorithms that simultaneously perform dimension reduction and clustering.
  - No strong statistical justification or grounding, but has been successfully applied in a range of areas
- NMF produces a **parts-based** decomposition of the latent relationships in a data matrix
  - Like SVD/PCA can reduce the dimensionality
  - Mathematically, given a non-negative matrix $\mathbf{A}$, finds k-dimension approximation in terms of non-negative matrices $\mathbf{W}$ and $\mathbf{H}$:
    $\mathbf{A} \approx \mathbf{WH}$, $\mathbf{W} \geq 0$, $\mathbf{H} \geq 0$
    where $\mathbf{W}$ has k columns and $\mathbf{H}$ has k rows.

        - Each item (i.e. column of **A**) approximated by a linear combination of k reduced dimensions or *basis vectors* in **W**.
        - Each *basis vector* can be interpreted as a cluster. The memberships of items in these clusters is encoded by **H**.
- NMF can be solved by optimising against a suitable objective
    - Commonly Euclidean Distance (Frobenius Norm) of the residual **A-WH**
    - Many different approaches, but common one is called **multiplicative update rules**.
        - Starting with a guess for **H** and **W**, iteratively repeat the following steps until convergence:
            - Update **H** on basis of current **W**
                - Updated **H** become the current one
            - Update **W** on basis of current **H**
                - Updated **W** become the current one
- Lots of NMF variants
    - Different objectives
    - Better optimisation schemes
    - Application of constraints (**especially sparsity**)
- Topic Modelling can be achieved by applying NMF to a term-document matrix
    - Typically first apply TF-IDF weighting to occurrences and normalise document vectors to unit length
    - The **basis vectors** (columns of **H**) represent the **clusters** of terms
        - i.e. **the topics**
    - The coefficient matrix, **W** represents the membership weights of each document to the topics
- Main problems/challenges of NMF:
    - No definitive model selection strategy to choose k
    - Standard random initialisation of NMF factors can lead to *instability*

Probabilistic Topic Models

- Overall idea:
    - Each **topic** is a distribution over words
    - Each **document** is a mixture of corpus-wide topics
    - Each **word** is drawn from one of those topics
    - We can only observe the documents - the other structures are hidden variables
        - Goal of probabilistic topic models is to **infer** the hidden variables
            - compute their distribution conditioned on the documents:
              p(topics, proportions, assignments | documents)
- **Probabilistic Latent Semantic Analysis (PLSA)** was the earliest probabilistic topic model
    - Given a corpus, observations produced in the form of pairs of words and documents (w,d)
    - Each observation is associated with an unobserved latent class variable, c
    - PLSA model assumes that the probability of a co-occurrence $P(w,d)$ is a mixture of *conditionally independent multinomial distributions*
    - Limitation/critisism is that this model is generative, but only for the training documents...
- **Latent Dirichlet Allocation (LDA)** is a Bayesian extension of PLSA
    - Adds a **Dirichlet prior** on the per-document topic distribution
        - Making the model fully generative for new documents
    - Parameters must be learned using Bayesian inference (e.g. variational Bayes/Gibbs

Sampling/etc)

- In practice: better than PLSA for small datasets; with lots of data tends to perform similarly

Further Reading

- Chapter 10 of Programming Collective Intelligence talks about NMF and shows how the multiplicative update rules algorithm works.
- Modern interest in NNMF was really started by Lee and Seung's Nature paper in 1999, although the techniques had been around for longer:
  - Daniel D. Lee and H. Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". Nature 401 (6755): 788–791
- PLSA was shown by Thomas Hofmann's seminal paper at NIPS:
  - Thomas Hofmann, Learning the Similarity of Documents : an information-geometric approach to document retrieval and categorization, Advances in Neural Information Processing Systems 12, pp-914-920, MIT Press, 2000. http://papers.nips.cc/paper/1654-learning-the-similarity-of-documents-an-information-geometric-approach-to-document-retrieval-and-categorization.pdf (http://papers.nips.cc/paper/1654-learning-the-similarity-of-documents-an-information-geometric-approach-to-document-retrieval-and-categorization.pdf)
- Overview of probabilistic topic modelling and LDA by its inventor David Blei: http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf (http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf)
- The original LDA journal paper:
  - Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research 3 (4–5): pp. 993–1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf)