Autoencoding Variational Bayes for Latent Dirichlet Allocation

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Zach Wolpe
Supervisor: Dr De Waal

Natural Language Processing

NLP encompasses a series of techniques to process and understand text data. By nature both a science and an art; as copious contrasting interpretations can be drawn from the same body of text.

The first part of the process is covering the raw text to numerical values for computation without ballooning the dimensionality of the data but whilst still retaining maximum information. As the goal is to infer the semantic meaning of the text and not the exact diction utilised, similar words are grouped to reduce data-complexity.

A corpus (a collection of documents) is converted to a Bag-of-Words BOW matrix containing the frequency of word appearances over the documents. Thus the likelihood of a word's appearance conditional on other words in a document can be used to discover the structure of the corpus in categorising documents.

Latent Dirichlet Allocation

LDA is a clustering algorithm of text data. *Latent* variables (document topics) are learnt over a Dirichlet simplex with $(K-1)$ dimensions. Since variational parameters are learnt a distribution over each topic is approximated and thus $2 \times (K-1)$ parameters are of interest. LDA is a soft-clustering algorithm thus allocating document in a probabilistic manner.

Autoencoder Loss Function

The autoencoder's loss function elegantly describes what the network is trying to minimise:

$$\mathscr{L}(\theta, \phi) = -\mathbb{E}_{z \sim q_\phi(z|x)}(\log p_\theta(x|z)) + D_{KL}[q_\phi(z|x)||p_\theta(z)]$$

The first term maximises the likelihood of the data over the topics whilst the second term - a KL divergence term - ensures the learnt distribution is sufficiently close to the Dirichlet prior.

Mission

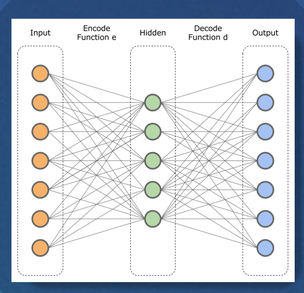
Big data has become the recipient of exceeding global interest as a consequence of it's exponentially growing market-value; however although topical it is not without challenges. Many complex functions are *intractable* and need be approximated: a notable challenge when datasets grow to gargantuan proportions. Intractable function evaluation is often resolved through Montecarlo methods (sampling techniques) however this proves unviable under the current global *big data* paradigm. Another frequently used solution is variational Bayes, which transforms an inference problem into an optimisation problem by parameterising a far simpler approximate-substitute distribution.

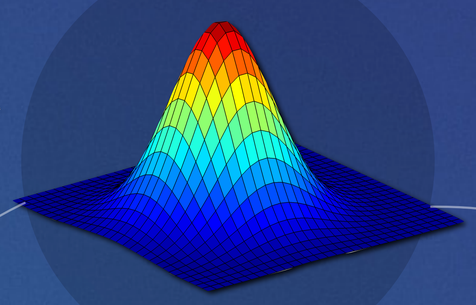
Alternatively utilising an autoencoder allows one to - theoretically - model intractable functions with deep learning in an efficient, Bayesian, manner. A practical example of these complex intractable functions arises in modelling text data with LDA (Latent Dirichlet Allocation). My dissertation investigates the viability of this alternative.

Research Questions:

- Does the autoencoder LDA provide comparable results - and predictive capability - to the well establish variational inference - online learning - LDA?
- Is it possible to scale the autoencoder to large datasets?
- Does the autoencoder offer significant processing-efficiency advantages as datasets scale?

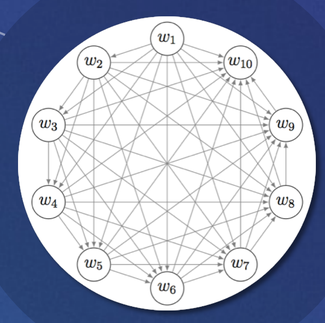
Autoencoders differ from conventional neural networks in that by definition the output replicates the input, thus the learnt parameters represent the same *information* in a reduced space.

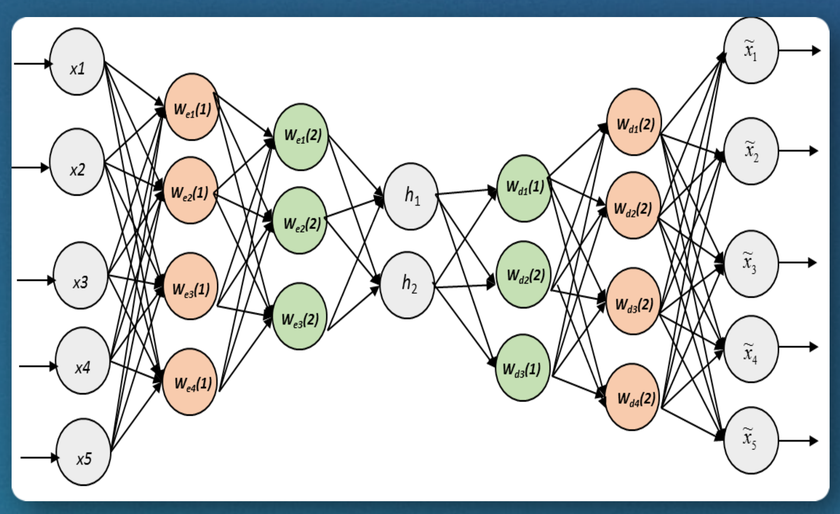




Mean-field approximation is used to learn parameter distributions over latent variables.

To account for sampling variation, random samples are draw from the corpus, the model is fitted and the coherence is calculated repeatedly - learning the sampling distribution of coherence scores.





Autoencoder Deep Learning

The Autoencoder architecture consists of ***128*** input nodes thereafter connected to ***W*** nodes in the first layer where ***W*** is size of learnt vocabulary. A simplex structure is used to learn the K topics and as such $(K-1)$ parameters are requires, however since variational approximation is performed a probabilistic distribution is learnt - as apposed to static values - and as a consequent a mean and variance are estimated: thus requiring $2 \times (K-1)$ parameters - this parameterising the second hidden layer structure.

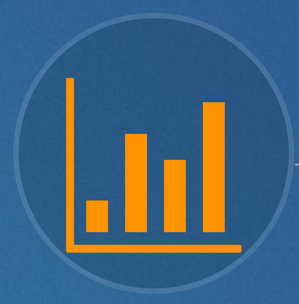
Both hidden layers are repeated 100 times, yielding a deep neural network with dimensions:

$$128 \times (W * 100) \times (2 * (K-1) * 100) \times (W * 100) \times 128$$

Natural Language Processing Pipeline

Process text data & process BOW sparse matrix

Learn distributions over the K topics based on the conditional likelihood of each work $\{w_i, w_j\}$ pair



Soft cluster documents into K topics based on probabilistic distributions

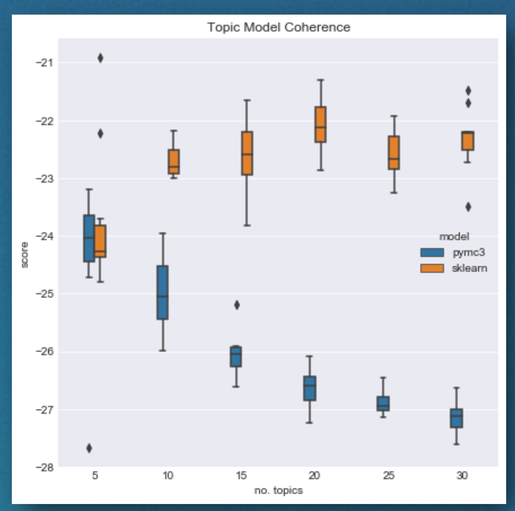
Topic model Coherence

Model Coherence was computed as a comparison metric for both LDA methods. Topic coherence is computed on a word $\{w_i, w_j\}$ pairwise basis & as such the computational requirements balloon as data complexity increases, thus only the most frequently occurring word pairs are considered.

The analysis unambiguously concludes the autoencoder fails to match the variational inference LDA's performance as models gain complexity.

Coherence is a function of the conditional likelihood between word pairs:

$$score^k_{UMass}(w_i, w_j|K) = \log \frac{D(w_i, w_j) + \epsilon}{D(w_i)}$$



Conclusion

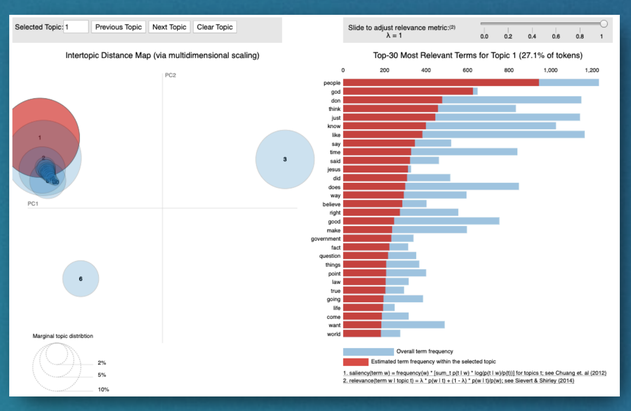
In comparing the variational inference LDA with the autoencoder LDA the results definitively concur the autoencoder fails to deliver adequate performance as model complexity scales. Model performance is comparable under simple conditions - as coherence is matched in models of $K = 5$ topics however as K increases the performance steadily declines. Interestingly the online variational Bayes method reaches optimal performance under the known true structure of the dataset.

When visualising the topic distributions learnt it is clear that the autoencoder clusters topics in an inefficient nested method - producing lackadaisical results as K grows.

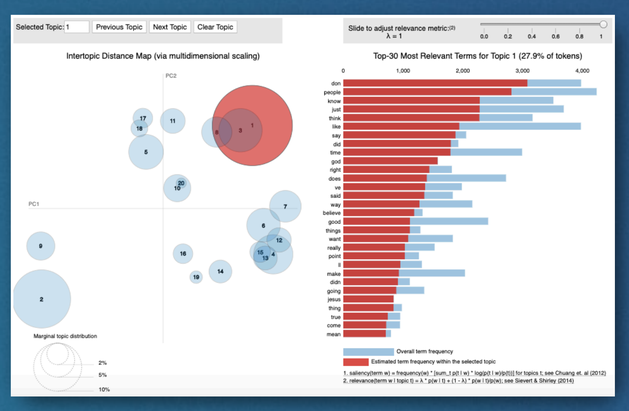
After investigating the sampling procedure of the autoencoder we found the model - although initialised with a uniform-unbiased-Dirichlet prior - fails to properly explore the domain space; vastly biasing initial random conditions.

As a consequence all documents are clustered between the first $K \approx 5$ - failing to capture the true semantic variation among documents.

Autoencoder LDA Topic Distributions



Variational Bayes LDA Topic Distributions



References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.
Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. arXiv:1312.6114 [cs, stat], December 2013.