Quantitative Text Analysis

Meeting 11

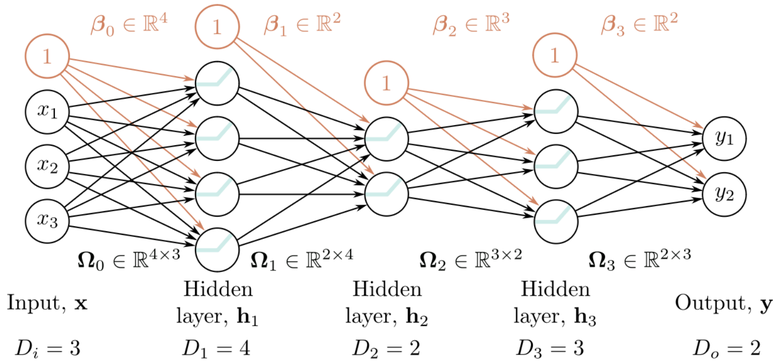
Petro Tolochko

Data Availability

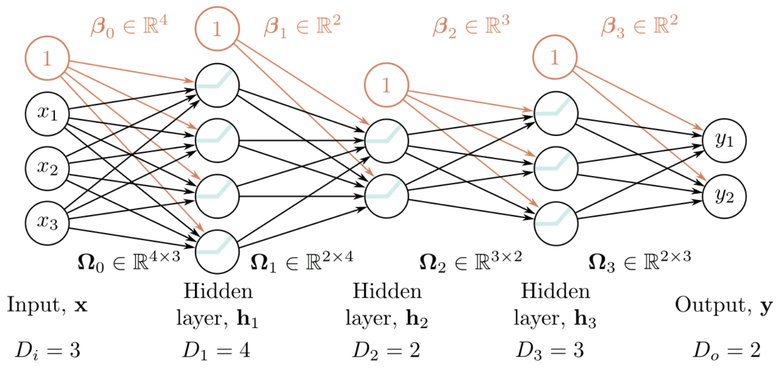
- State of the Union Addresses (e.g., SOTU package in R)
- Federalist Papers (in the Course repository)
- Project Gutenberg — free ebooks (e.g., https://cran.r-project.org/web/packages/gutenbergr/index.html)
- Manifesto Project Database — Political Parties' manifestos (e.g., manifestor in R; https://manifesto-project.wzb.eu/)
- https://github.com/niderhoff/nlp-datasets

Fine Tuning

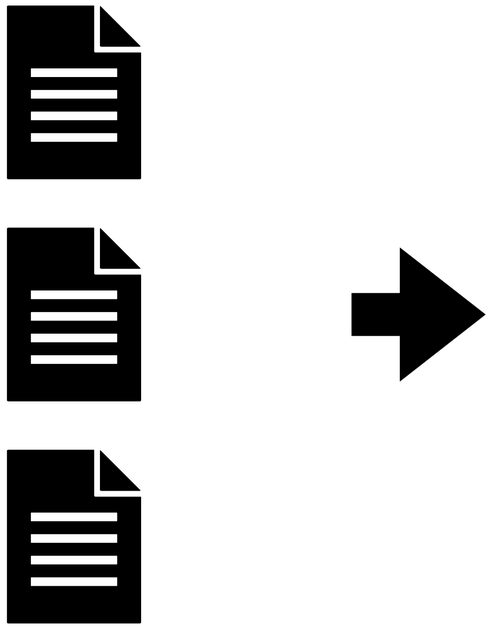
Fine Tuning


$\boldsymbol{\beta}_0 \in \mathbb{R}^4$
$\boldsymbol{\beta}_1 \in \mathbb{R}^2$
$\boldsymbol{\beta}_2 \in \mathbb{R}^3$
$\boldsymbol{\beta}_3 \in \mathbb{R}^2$
1
x_1
x_2
x_3
y_1
y_2
$\boldsymbol{\Omega}_0 \in \mathbb{R}^{4\times3}$
$\boldsymbol{\Omega}_1 \in \mathbb{R}^{2\times4}$
$\boldsymbol{\Omega}_2 \in \mathbb{R}^{3\times2}$
$\boldsymbol{\Omega}_3 \in \mathbb{R}^{2\times3}$
Input, $\mathbf{x}$
$D_i = 3$
Hidden layer, $\mathbf{h}_1$
$D_1 = 4$
Hidden layer, $\mathbf{h}_2$
$D_2 = 2$
Hidden layer, $\mathbf{h}_3$
$D_3 = 3$
Output, $\mathbf{y}$
$D_o = 2$

Fine Tuning


$\boldsymbol{\beta}_0 \in \mathbb{R}^4$
$\boldsymbol{\beta}_1 \in \mathbb{R}^2$
$\boldsymbol{\beta}_2 \in \mathbb{R}^3$
$\boldsymbol{\beta}_3 \in \mathbb{R}^2$
1
x_1
x_2
x_3
y_1
y_2
$\boldsymbol{\Omega}_0 \in \mathbb{R}^{4\times3}$
$\boldsymbol{\Omega}_1 \in \mathbb{R}^{2\times4}$
$\boldsymbol{\Omega}_2 \in \mathbb{R}^{3\times2}$
$\boldsymbol{\Omega}_3 \in \mathbb{R}^{2\times3}$
Input, $\mathbf{x}$
$D_i = 3$
Hidden layer, $\mathbf{h}_1$
$D_1 = 4$
Hidden layer, $\mathbf{h}_2$
$D_2 = 2$
Hidden layer, $\mathbf{h}_3$
$D_3 = 3$
Output, $\mathbf{y}$
$D_o = 2$

Assistants (like Chat-GPT)



Input / Output
pairs

Transformer Models for Social Science

- Reduction of labeled data, performance increase
- Pre-trained Language Models:
    - Pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are used for text classification tasks.
    - These models learn contextual word representations from large amounts of unlabeled text data and can be fine-tuned for specific classification tasks.

Transformers for “Feature Engineering”

- What features are the most important for (e.g.,) discriminating text types?
- Feed the model several example texts
- It distinguishes features that are important
- Use these features to classify/scale/etc., large quantities of data

Open vs. Closed source

- Ethics
- Performance vs. Reproducibility
- Use cases
    - Do you ***need*** it?

Open Source

- Huggingface:
    - https://huggingface.co/
    - Many pre-trained models and datasets
- spaCy:
    - https://spacy.io/
    - Ecosystem for NLP

Open Source

- Better in Python
- But:
    - spaCyR
    - https://github.com/chainsawriot/grafzahl
        - Huggingface wrapper

Course Assessment

- Participation (20%)
- Final paper: application of one or several automated text analysis methods on a topic related to the Master's thesis or a topic of free choice (80%)
- Contents: Methods + Results + Discussion section of a (pseudo)-academic paper (6-8 pages)
- Commented code
    - Python / R / Julia / …
- Upload to Moodle
- Deadline: May 15th 23:59:59

Your Thoughts?