Quantitative Text Analysis

Meeting 8

Petro Tolochko

Unsupervised Machine Learning

Machine Learning

- Supervised
    - An outcome variable is defined
    - Focus is on prediction
- Unsupervised
    - No outcome variable has been defined
    - Focus is on patterns

Machine Learning

- Supervised
    - An outcome variable is defined
    - Focus is on prediction
- **Unsupervised**
    - **No outcome variable has been defined**
    - **Focus is on patterns**

How to use the *supervised* methods?

- Easy
- At least conceptually
- ***Clear objective function***

How to use the supervised methods?

$$Y = (y_1, y_2, ..., y_n)$$
$$X = (x_1, x_2, ..., x_n)$$
$$(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$$

Task to predict $\hat{y}$ as close to y

How to use the supervised methods?

$$L(y, \hat{y}) = (y - \hat{y})^2$$

$$\hat{y} = \underset{\theta}{\operatorname{argmin}} \; E\left[L(model(\mathbf{x}, \theta), y)\right]$$

How to use *unsupervised* learning

- Objective function?

How to use *unsupervised* learning

- Objective function?
- Quantity of interest?

How to use *unsupervised* learning

- Objective function?
- Quantity of interest?
- Objective function = your quantity of interest


LIGHTYEAR

How to use *unsupervised* learning

- Objective function?
- Quantity of interest?
- Objective function = your quantity of interest
- ***This is difficult***

Measurement

- A collection of quantitative or numerical data that describes a property of an object or event

Measurement

- A collection of quantitative or numerical data that describes a property of an object or event
- What is the ***object***?

Measurement

In Social Science

- Operationalisation ➔ Data Collection ➔ Analyses ➔ Measurement
- The quantity of interest is ***extrinsic*** to the model

Measurement

In Computer Science

- Model Building ➔ Measurement of Performance
- The quantity of interest is ***intrinsic*** to the model

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

Main quantity of
Interest for
computer science

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

Means to an end
for social science

Main quantity of
Interest for
computer science

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

Means to an end
for social science

Main quantity of
Interest for
computer science

$$\hat{\theta} \approx \theta$$

What social
science wants

Prediction vs. Inference

- Computer scientists often emphasise *prediction*
- Social scientists are often more interested in *inference*
- Vast, multidimensional parameter space = not suitable for inference
    - Good for prediction
- E.g., Turing test
    - Machine passes
    - *Why* does it pass/not pass

Problems

- Translation of Social Science concepts
- Connecting Methods to Theory
- Difficult to understand what is being *measured*

Unsupervised Learning Example

- Clustering algorithms are great tools
- Not well suitable for the “standard” social science paradigm
- Needs external validation, but there is no “best” method
- “Validation” based on “theory” or “expectation” leads to biases

The paradigm

- Approximating a data-generating process

The paradigm

- Approximating a data-generating process
- *Assumption*: there is one (and only one) “true” data-generating process

The paradigm

- Approximating a data-generating process
- *Assumption*: there is one (and only one) “true” data-generating process
- It ***is*** the reality

Sticking to the paradigm

- The “normal” paradigm works only if we assume that there is one “correct” classification
- Need to adapt to different methods
-

Sticking to the paradigm

- The “normal” paradigm works only if we assume that there is one “correct” classification
- Need to adapt to different methods
- Unsupervised methods are ***meaningless*** in conjunction with the “true” data-generating process assumption

Focus on Discovery

Objectives

- Descriptive analysis/Discriminating words:
  - What are the characteristics of a corpus? How do some documents compare to each other
  - Collocation analysis, readability scores, Cosine/Jaccard similarity
- Clustering and scaling:
  - What groups of documents are in the corpus? Can the documents be placed on a dimension?
  - Cluster analysis, principal component analysis, wordfish..
- Topic modeling:
  - What are the main themes in a corpus?
  - LDA, STM

K-Means Clustering

- Simple(ish) algorithmic method
- Partitions the data into K non-overlapping clusters

Setup

$$C_1, C_2, ..., C_k$$

$$C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$$

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

Assumption and task

- Optimal clustering solution is the one where ***within-cluster variation is as small as possible***

$$W(C_k)$$

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Within cluster variation

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Algorithm

- Randomly assign cluster numbers (1 through K) for each observation
    - Iterate until no further changes to the cluster assignment:
    - For each cluster determine the centroid (average of all observations in the cluster)
    - Re-assign observations to a cluster with the closest centroid (calculated with a distance metric).

Guarantees convergence at a local optimum

- Cannot guarantee the best solution
- But are rather good one
- Sensitive to random assignment at the start

Cluster Algorithms Validation

- Data assumptions (think data generation)
- Internal validity (best results for the data)
- External validity (matches with pre-existing understanding of data)
- Cross-validity (similar results across similar datasets)
- You are the validation method

Questions?

Topic Modelling

- A model to discover latent topics
- Not synonymous with LDA
- LDA is one of topic models

- Latent semantic analysis
- Singular value decomposition
- Even clustering methods (like the one we just discussed)

Latent Dirichlet Allocation

- Bayesian generative hierarchical model
- First introduced as a way to simultaneously model traits and genes (Pritchard, 2000)
- Adjusted for text analysis ML applications (Blei et al., 2003)

Latent Dirichlet Allocation

- Estimates a distribution of ***words*** across ***documents*** across ***latent topics***

Latent Dirichlet Allocation

- By modelling distributions of topics over words and words over documents, topic models identify the most discriminatory groups of documents automatically.
- Assumption: if a document is about a certain topic, one would expect words that are related to that topic to appear in the document more often than in documents that deal with other topics.

Mixture Models

$$P(x) = \Sigma_{k=1}^{K} \pi_k \times P(x \,|\, \theta_k)$$

Hierarchical Models

Hierarchical Models

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$$

Hierarchical Models

$$
\begin{aligned}
y &\sim Normal(\mu, \sigma) \\
\mu &= \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \\
\beta &\sim Normal(0,5) \\
\sigma &\sim Exponential(1)
\end{aligned}
$$

Hierarchical Models

Likelihood Function

$$y \sim Normal(\mu, \sigma)$$
$$\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$
$$\beta \sim Normal(0,5)$$
$$\sigma \sim Exponential(1)$$

Prior Distribution of Parameters

Hierarchical Models

$$
\begin{aligned}
y &\sim Normal(\mu, \sigma) \\
\mu &= \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \\
\beta &\sim Normal(0{,}5) \\
\sigma &\sim Exponential(1)
\end{aligned}
$$

Hyperparameters

LDA

$$\theta_k \sim Dirichlet(\alpha)$$ for each topic k

$$\eta_d \sim Dirichlet(\beta)$$ for each document d

$$z_{d,w} \sim Multinomial(\eta_d)$$ for each word w in document d

$$x_{d,w} \sim Multinomial(\theta_{z_{d,w}})$$ for each word w in document d

LDA

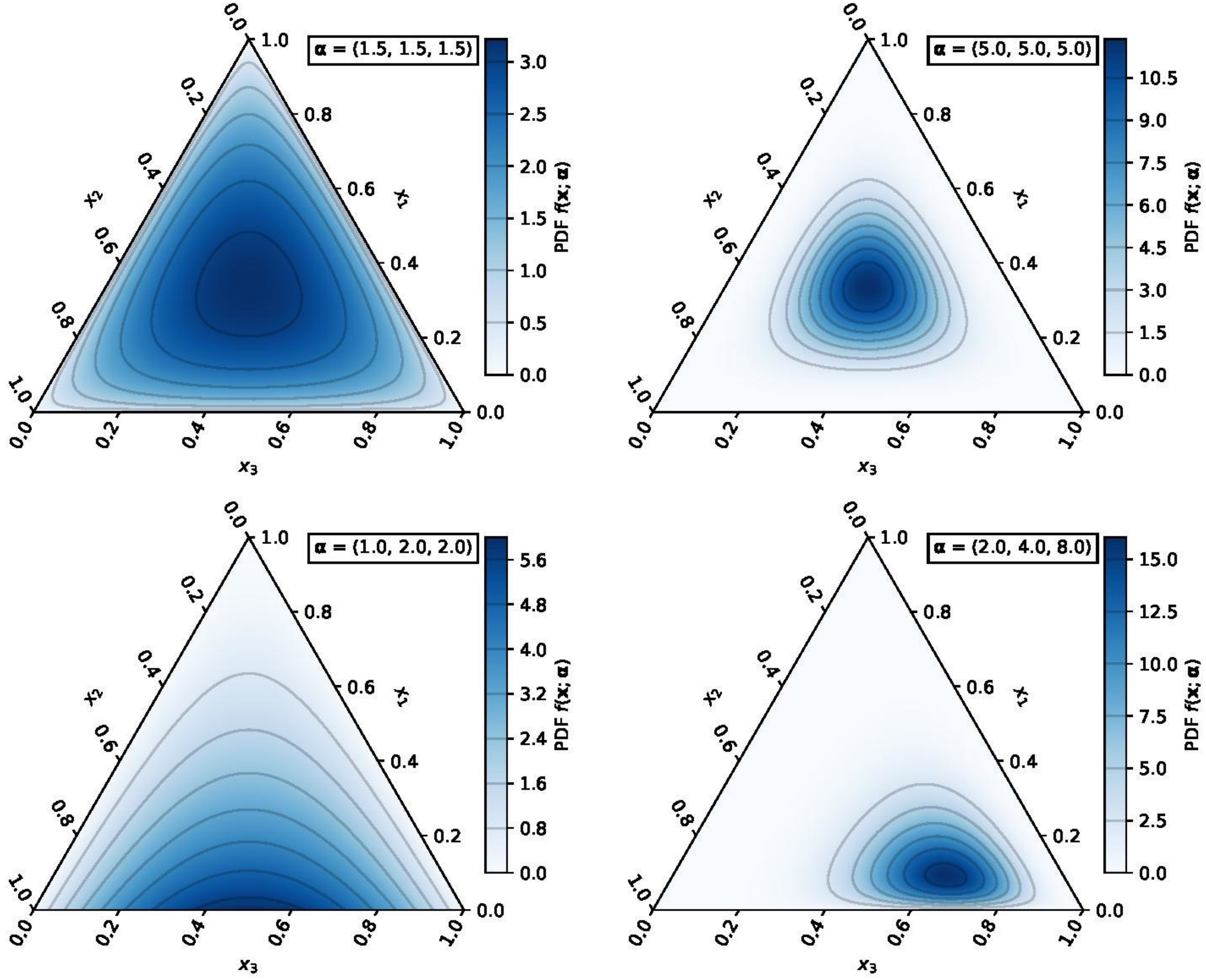
- θ_k is the topic-word distribution for topic k, representing the probability of each word given the topic.
- η_d is the document-topic distribution for document d, representing the probability of each topic in the document.
- $z_{d,w}$ is the topic assignment for word w in document d, indicating which topic generated the word.
- $x_{d,w}$ is the observed word in document d.

LDA

- For each topic $k \in \{1, ..., K\}$:
- Draw a distribution over words $\theta_k \sim \text{Dirichlet}(\alpha)$, where α is a hyperparameter representing the topic-word prior.
- For each document $d \in \{1, ..., D\}$:
- Draw a distribution over topics $\eta_d \sim \text{Dirichlet}(\beta)$, where β is a hyperparameter representing the document-topic prior.
- For each word w in document d:
- Draw a topic assignment $z_{d,w} \sim \text{Multinomial}(\eta_d)$, indicating which topic generated the word.
- Draw a word $x_{d,w} \sim \text{Multinomial}(\theta_{z_{d,w}})$, indicating the specific word generated by the chosen topic.

LDA

- The goal of LDA is to infer the posterior distributions of the latent variables θ and η given the observed documents.
- Once the posterior distributions are estimated, LDA can be used to assign topics to new documents or extract the most probable words for each topic.



α = (1.5, 1.5, 1.5)
α = (5.0, 5.0, 5.0)
α = (1.0, 2.0, 2.0)
α = (2.0, 4.0, 8.0)
PDF f(x; α)
x1
x2
x3

Tricky to estimate

Structural Topic Models (Roberts et al., 2014)

- Add additional distribution for additional structure in the text:

- Allows to add additional covariates to the estimation

Validating topic models

- Tests of:
  - topic semantic validity: assess the extent to which the keywords within each topic have a coherent underlying meaning, and how these meanings behave across topics (e.g., Quinn et al., 2010, p. 210)
  - convergent validity: topic probabilities per document are compared with an external trusted variable like manual coding for the same documents (e.g., Guo et al., 2016)
- A possible validation workflow proposed by Maier et al. (2018): combi of quantitative topic summary metrics (e.g., NPMI; Lau et al., 2014) and human expert evaluations.

Text scaling methods

- Attempts to fit documents into a unidimensional space
- Documents are “scaled” based on the frequency of used terms
- Assume “discriminating” words have a Poisson distribution

- “Ideological” successor to log odds ratio we’ve seen

Wordfish (Slapin and Proksch 2008)

$$w_{ik} \sim \mathrm{Poisson}(\lambda_{ik})$$

$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

α_i Text size from type i

ψ_k Frequency of word k

β_k Discrimination power of word k

θ_i Ideological position of type i

Wordfish (Slapin and Proksch 2008)

$$w_{ik} \sim \text{Poisson}(\lambda_{ik})$$

$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

α_i Text size from type i

ψ_k Frequency of word k

β_k Discrimination power of word k

θ_i Ideological position of type i

Wordfish (Slapin and Proksch 2008)

$$w_{ik} \sim \text{Poisson}(\lambda_{ik})$$

$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

α_i Text size from type i

ψ_k Frequency of word k

β_k Discrimination power of word k

θ_i Ideological position of type i

Questions?