Introduction

David Mareček

February 25, 2020

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research, Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

About

Formerly: Selected Methods in Machine Learning

Webpage: `http://ufal.mff.cuni.cz/courses/npfl097`

E-credits: 3

Examination: 1/1 C

Course passing requirements

There will be three programming assignments:

- for each, you can obtain at most 10 points
- you will have three weeks to finish it
- you will obtain only half of the points for assignments delivered after the deadline

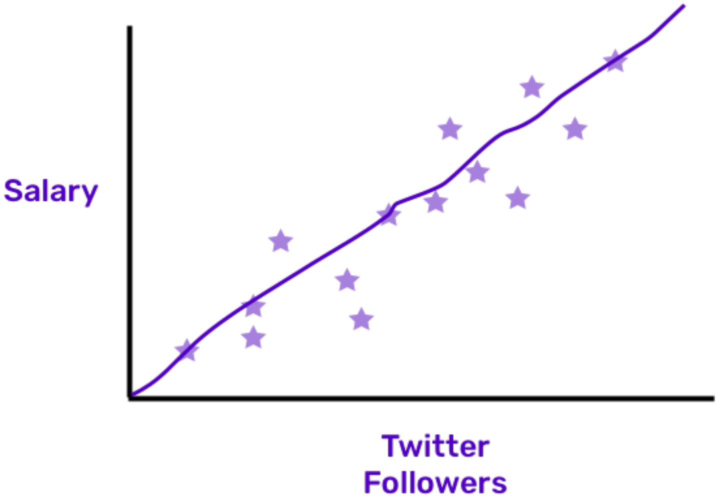
You can obtain 10 points for individual presentation:

- at least 30 minutes presentation
- selected machine learning method or task
- need to be confirmed at least one week before

You pass the course by obtaining at least **20 points**.

Supervised vs. Unsupervised Learning

Supervised learning



Supervised vs. Unsupervised Learning

Unsupervised learning

Unsupervised Machine Learning

A type of machine learning that helps find previously unknown patterns in data set without pre-existing labels.

- How do you find the underlying structure of a dataset?
- How do you summarize it and group it most usefully?
- How do you effectively represent data in a compressed format?

These are the goals of unsupervised learning, which is called "unsupervised" because you start with unlabeled data.

Problems to Solve

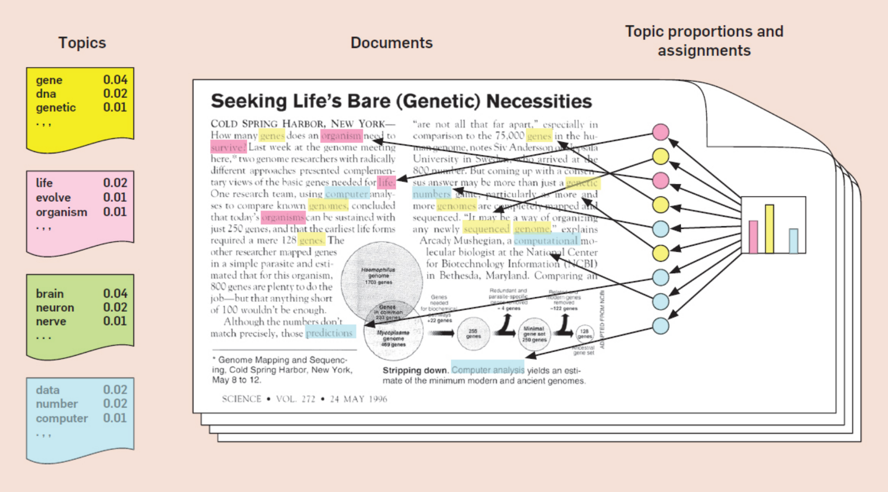
Modelling Document Collections

- We want to find an underlying structure of given documents.
- Goal: divide the documents to classes.
- Better goal: find a set of topics and assign several relevant topics to each document.
  - Each topic is represented by a distribution over words.
  - Each document has a distribution over its topics (typically 1 to 5 main topics)
  - The total amount of topics is the only constant chosen by the user.

method: Bayesian Inference – Latent Dirichlet Allocation

related course:
NPFL103 – Information Retrieval

Modelling Document Collections



Problems to Solve

Word Clustering, Language Clustering

- we can generate many features for each word or language
- Goal: categorize words (part-of-speech tags) or languages (language families)

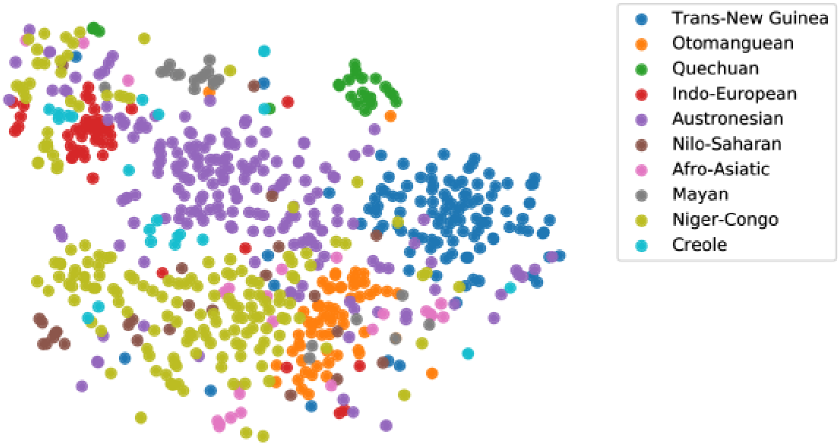
methods: K-means, Mixture of Gaussians, Hierarchical Clustering

related course:
NPFL129 – Machine learning for Greenhorns

Problems to Solve

Language Clustering

Language vectors from multilingual MT visualized by T-SNE



(picture by Jörg Tiedemann)

Problems to Solve

Word Embeddings, Contextual Embeddings, Sentence Embeddings

- Let's suppose we have huge number of texts.
- We want to find a vector of real numbers representing each word (or sentence).
- Similar words (or sentences) should be represented by similar vectors.

Are Skipgram and BERT unsupervised?

related course:
NPFL114 – Deep learning

Other Problems

Unsupervised Machine Translation

- Let's suppose we have huge number of comparable texts in two languages, but only very little or no parallel data.
- We want to infer a dictionary or a translation system.

related courses:
NPFL087 – Statistical Machine Translation
NPFL120 – Multilingual Natural Language Processing

Course overview [1]

1) Probabilistic Machine learning (Bayesian inference):
(cca 5 lectures)

- Beta-Bernoulli and Dirichlet-Categorical models
- Mixture models
- Expectation-Maximization
- Metropolis-Hastings, Gibbs sampling
- Modelling Document Collections – Latent Dirichlet Allocation
- Chinese Restaurant Process, Pitman-Yor Process
- Text Segmentation, Unsupervised Tagging, Unsupervised Parsing

2) Clustering:
(cca 2 lectures)

- K-means
- Mixture of Gaussians
- Hierarchical Clustering
- Evaluation of Unsupervised Clustering

Course overview [2]

3) Component analysis:
(cca 1 lecture)

- Principal Component Analysis
- Independent Component Analysis

4) Interpreting deep neural networks:
(cca 2 lectures)

- Word embeddings
- Contextual embeddings
- Sentence embeddings
- Probing
- Analysis of Attentions

Prerequisities and related courses

Basic probabilistic and ML concepts:

- NPFL067 – Statistical methods in NLP I
- NPFL054 – Introduction to Machine Learning
- NPFL129 – Machine learning for Greenhorns

Basic deep-learning concepts:

- NPFL114 – Deep Learning

Other related courses:

- NPFL104 – Machine Learning Methods
- NPFL087 – Statistical Machine Translation
- NPFL103 – Information Retrieval
- NPFL120 – Multilingual Natural Language Processing

Assignments

There will be three programming assignments:

1. Topic Modelling – Latent Dirichlet Allocation (LDA)
2. Unsupervised Text Segmentation
3. Clustering and Principal Component Analysis on Word Embeddings

Preferred language: Python

For each assignment there will be one programming lecture reserved for implementation, questions and discussions over preliminary results.

Basic concepts

Frequentist vs. Bayesian interpretation of probability

Frequentist probability: Probability of an event is the limit of its relative frequency in a large number of trials.

$$P(x) \approx \frac{n_x}{n}, \qquad P(x) = \lim_{x \to \infty} \frac{n_x}{n}$$

Bayesian probability: Probability of an event is interpreted as reasonable expectation representing a state of knowledge.

You toss a coin 10 times, 7 times head and 3 times tail. What is your expectation about the probability of head?

$$P(x) \approx \frac{n_x + \alpha_x}{n + \alpha}$$

Bayes theorem

Conditional probability: probability of event X given that event Y has occured

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Curse of dimensionality

Various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.

- *Sampling* - exponential increase of volume
- *Machine learning* - high-dimensional feature space need enormous number of training data (several samples of each combination of features)
- *Distances* - in highly dimensional space, the euclidean distances between different pairs of samples are very similar. Relative volume of inscribed hypersphere decreases.