Introduction

David Mareček

October 1, 2024

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

Why Unsupervised:

- for tasks, for which we do not have any labelled data (set of input-output examples)
- for tasks, in which we do not know exactly, what should be the output, e.g. document clustering, text segmentation, word alignment

Why in NLP:

- NLP = Natural Language Processing
- General Machine-Learning techniques will be shown on NLP examples, but are easily applicable in other research areas.

About

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

E-credits: 3

Examination: 1/1 C

Form:

- lectures, exercises, discussions
- 3 programming homeworks
- test

Course passing requirements

Three programming assignments

- For each one you can obtain at most 10 points.
- You will always have at least three weeks to complete it.
- You will obtain only half of the points for assignments (or parts of them) delivered after the deadline.

Test

You can obtain 15 points from theoretical test (on 7th January)

You pass the course by obtaining at least **30 points**.

Related courses

Recommended (not necessary but useful for this course):

- NPFL067 – Statistical Methods in NLP I (winter)
- NPFL129 – Introduction to Machine Learning with Python (winter)
- NPFL114 – Deep Learning (summer)

Other courses (related topics):

- NPFL103 – Information Retrieval
- NPFL087 – Statistical Machine Translation
- NPFL140 – Large Language Models

Reading

Books:

- Christopher Bishop: Pattern Recognition and Machine Learning, Springer-Verlag New York, 2006
- Kevin P. Murphy: Machine Learning: A Probabilistic Perspective, The MIT Press, Cambridge, Massachusetts, 2012

Tutorials, papers:

- Kevin Knight: Bayesian Inference with Tears, 2009 (`https://sites.socsci.uci.edu/~lpearl/courses/readings/Knight2009_BayesWithTears.pdf`)

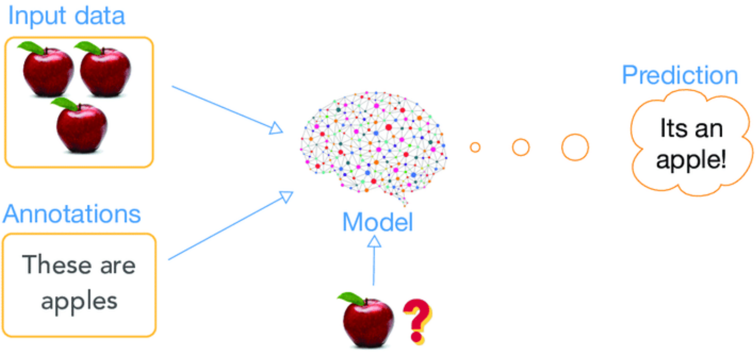
Schedule

Oct 1:	Introduction to Unsupervised ML
Oct 8:	Beta-Bernoulli probabilistic model
Oct 15:	Dirichlet-Categorical probabilistic model, Categorical Mixture Model
Oct 22:	Expectation Maximization, Bayesian Mixture Models
Nov 29:	Gibbs Sampling, Latent Dirichlet Allocation
Nov 5:	no lecture (Dean's day)
Nov 12:	Chinese Restaurant Process
Nov 19:	Pitman-Yor Process, Unsupervised Text Segmentation
Nov 26:	K-Means, Mixture of Gaussians
Dec 3:	Hierarchical Clustering, Clustering Evaluation, other clustering methods
Dec 10:	T-SNE, Principal Component Analysis, Independent Component Analysis
Dec 17:	Sparse auto-encoders and Interpretability of Neural Networks
Jan 7:	Final Test

Supervised versus Unsupervised ML

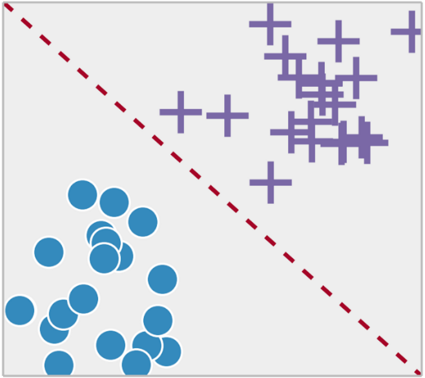
Supervised Machine Learning

- We know what the output values for our samples should be.
- We have a labelled (ground truth) data.
- Goal: to find a function that best approximates the relationship between input and output observable in the training data.

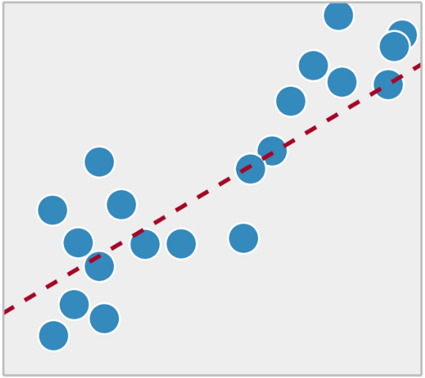


Supervised Machine Learning

Classification



Regression

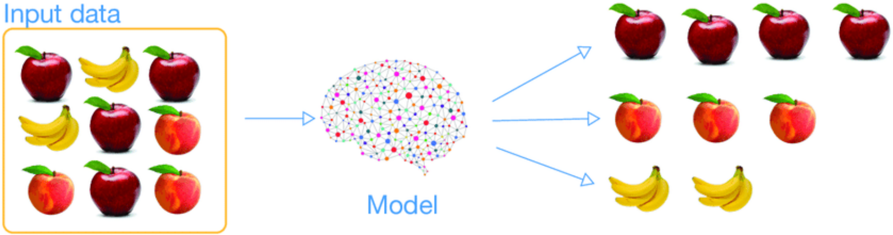


separating into multiple classes

assigning continuous real values

Unsupervised Machine Learning

- We do NOT have any labelled data.
- Goal: to discover a natural structure or previously unknown patterns present within given dataset.
- It works without the need of human intervention.



Unsupervised Machine Learning

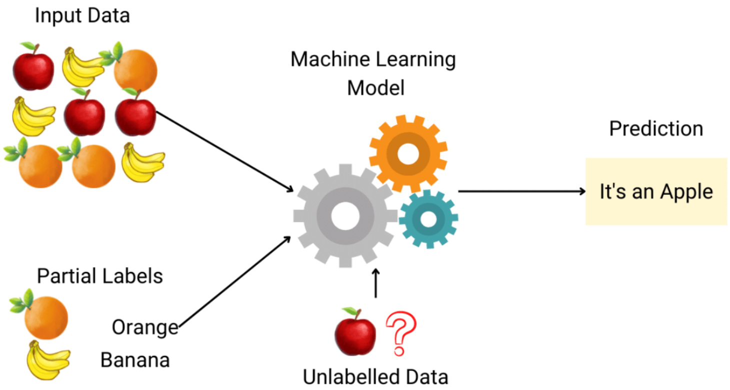
We use unsupervised learning if we want to answer following questions:

- How do you find the underlying structure of a dataset?
  $\rightarrow$ e.g. *Latent Variable Models*
- How do you summarize it and group it most usefully?
  $\rightarrow$ e.g. *Clustering*
- How do you effectively represent data in a compressed format?
  $\rightarrow$ e.g. *Dimensionality reduction*

Unsupervised machine learning techniques always start with unlabeled data.

Semi-Supervised Machine Learning

- We have only a small amount of labelled data.
- We have large unlabelled data.



Modern Machine Learning approaches

The classical division of machine learning into supervised and unsupervised is not very suitable for the modern ML methods.
Current machine learning models are very often optimized on one task but then they (or their internal representations) are used for solving another task.

Examples:

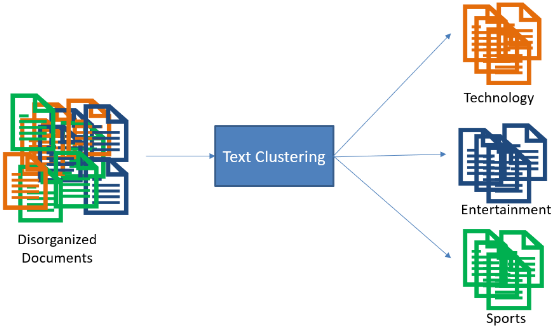
1. **Word2Vec model** (Mikolov, 2013)
    - input: large collection of texts
    - output: each word is assigned a vector of real numbers (unsupervised)
    - learning: given one word, predict another word from its context (supervised)
2. **Zero-shot in GPT3 model** (Brown, 2020)
    - input: A question or a command.
    - output: An answer (unsupervised)
    - learning: given a part of text, predict the next word (supervised)

Clustering

Clustering of documents

Assume you have a large set of documents.

You want to divide them into categories.



Clustering of documents – Supervised Setting

We have a set of documents labelled by categories (e.g. Technology, Entertainment, Sports).

We train a supervised model extracting words characterizing each of the categories.

Using this model, we can classify new documents into the given categories.

Clustering of documents – Unsupervised Setting

We do not have any labelled data.

- We do not know the labels of the clusters
- We do not even know the number of clusters we should use.

The goal here will be at least to divide the documents into groups.
The number of groups (clusters) is set manually.

Unsupervised setting may be preferred in many cases:

- We do not know what documents we have in our collections.
- We do not know, whether the topics of our labelled documents will cover all the new documents we want to classify.

Clustering of Documents – Features

What features (properties of documents) can we use?

Features:

- presence of predefined keywords - a binary vector
- TF-IDF score of a words in documents

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

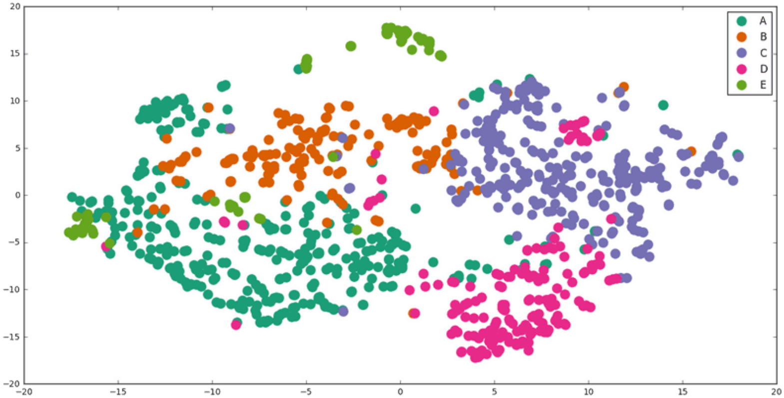
df_x = number of documents containing x

N = total number of documents

Clustering of Documents

We have a high-dimensional vector assigned to each document.

We can run a clustering algorithm.



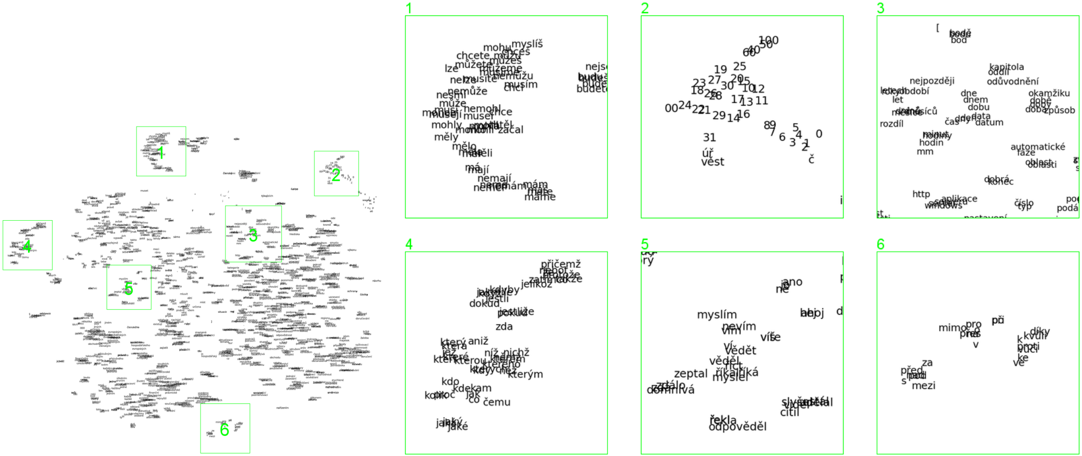
Document vectors visualized by T-SNE (picture by Sergios Karagiannakos)

What can we run clustering on?

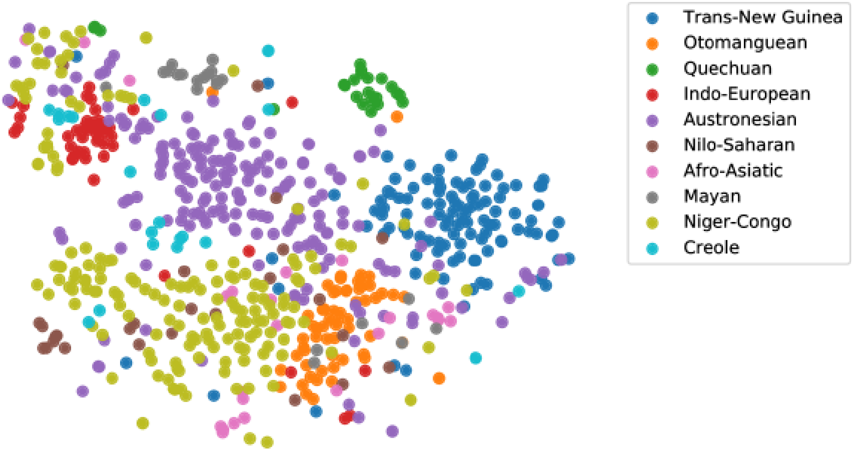
In NLP, we can cluster:

- documents
- words
- languages

Word Clustering



Language Clustering



Language vectors from multilingual MT visualized by T-SNE (picture by Jörg Tiedemann).

Clustering Algorithms

Methods:

- Hierarchical clustering
  - Aglomerative Clustering
  - Divisive Clustering
- K-means
- Gaussian Mixture Models
- ...

Properties:

- *Hard clustering:* each object belongs to a cluster or not
- *Soft clustering:* each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster)

... more about clustering in **November**

Latent Variable Models

Modelling Document Collections

In previous example, the task was to divide documents into groups.

But what if the documents cannot be easily separated into groups (many documents contains more than one topic)?

We want something more: to find an underlying structure of a given set of documents.

Previous goal: Assign a single class to each document.

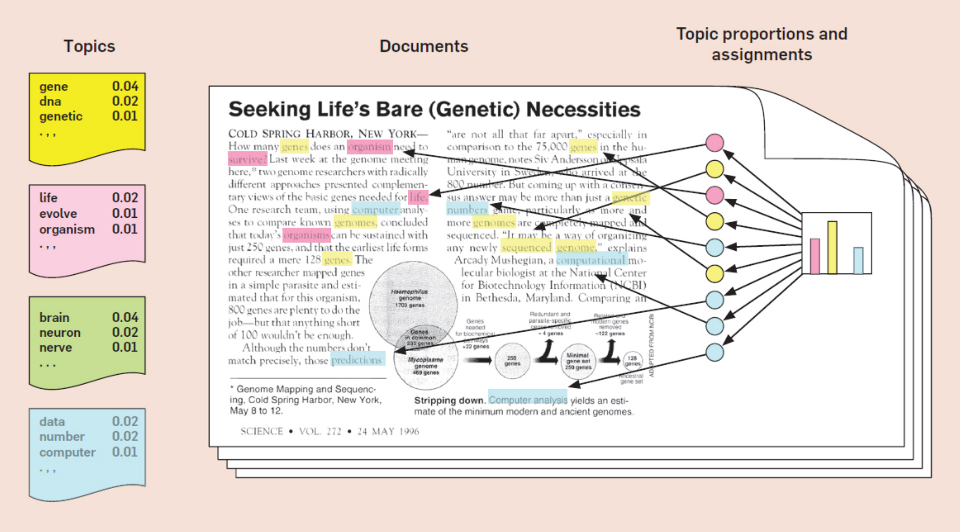
Better goal: Find a set of topics and assign several relevant topics to each document.

Latent variable model:

- 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.

Modelling Document Collections

Latent Dirichlet Allocation



Word Alignment

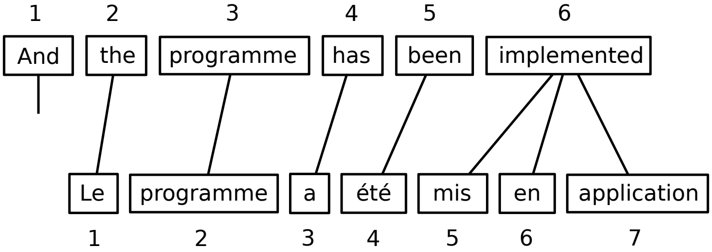
You have many parallel texts (sentences) in two languages.

- You want to align counterpart words.
- You want to extract a dictionary.

Sun's rays go to waste!	Los rayos caen en saco roto
2014 was the best year for tourism in Spain´s history, with over 65 million visitors. It doesn't take a genius to understand why this country has so much appeal to its northern neighbours. There is one **overwhelming** advantage that Spain enjoys: the sun	2014 fue el mejor año para el turismo en la historia de España, con 65 millones de visitantes. No hace falta ser un genio para entender por qué este país tiene tanto atractivo para sus vecinos del norte. Hay una ventaja **apabullante** de la que España goza: el sol.
As well as tourism, this element has **enabled** Spain to be the world's largest olive oil producer and "Europe's **vegetable garden**".	Además del turismo, este elemento ha **permitido** que España sea el mayor productor de aceite de oliva del mundo y "la **huerta** de Europa".
So far so good, and in keeping with good management of its resources. But there is a sector that seems to have **gone adrift** and is letting these rays of sunlight **go to waste**. A sector that Spain used to **back**, and as a result managed to position itself among the most advanced countries in this field.	**Hasta aquí todo bien**, y en concordancia con una buena gestión de sus recursos. Pero hay un sector que parece haberse **ido a la deriva** y está dejando **caer en saco roto** todos estos rayos de sol. Un sector por el que España **apostaba** y gracias a ello consiguió situarse entre los países más avanzados en este campo.

Word Alignment – Unsupervised setting

1. Introduce translation model with latent variables $p(en|es)$.
2. Maximize the translation probabilities over the whole data using the Expectation-Maximization (EM) algorithm.

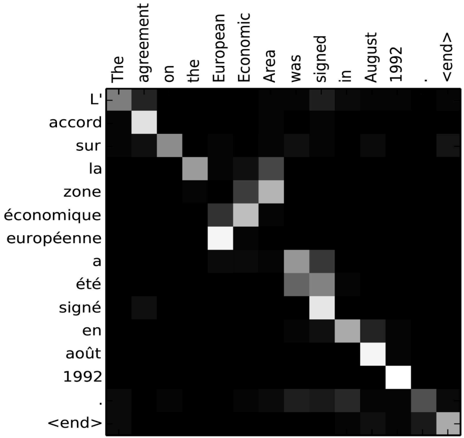


This used to be the first step of training machine translation systems (around 2010).

Word Alignment

Current best machine-translation systems use deep neural networks (since 2014).

The mapping of words between two languages may be extracted from the network as a by-product.



Text Segmentation

We need some language units for processing text.

- Paragraphs or sentences are too long, characters are too short.
- Words? There are to many of them and the size of dictionary is usually limited.
- Some languages do not have words.
- Byte-Pair Encoding, Bayessian inference

李叶的爸爸经常在外面，很少在家。李叶的妈妈是个很好看的女人，她有很多朋友，每天都和朋友一起玩。李叶的爸爸妈妈都很类，他们没有时间理他们的女儿。还有，李叶的妈妈好像一点也不喜欢李叶，她觉得李叶一点也不详她。李叶出生以后，她就告诉家里的阿姨：“如果你们想让我皇心，就不要让我看到这个孩子。”所以，李叶很少能见到她的爸爸妈妈。

Latent Variable Models

Tasks:

- Modelling Document Collections
- Text Segmentation
- Word Alignment

Methods:

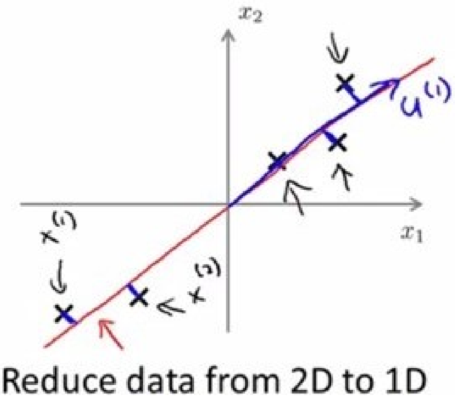
- Bayesian inference
- Dirichlet Process
- Pitman-Yor Process
- Gibbs Sampling

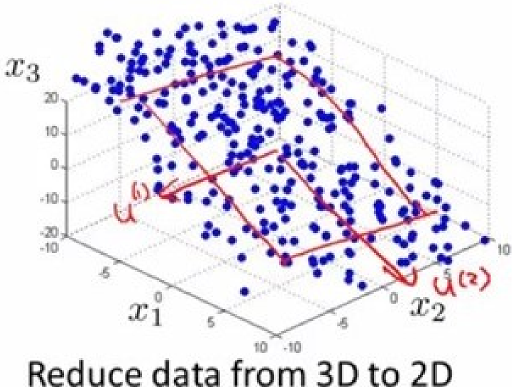
... more about Latent Variable Models in **October** lectures

Dimensionality Reduction

Principal Component Analysis

- We want to describe a highly dimensional vector space.
- E.g. 512-dimensional vector space of word embeddings.
- What are the most important features of the space?



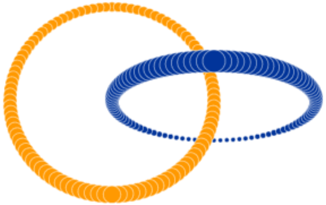


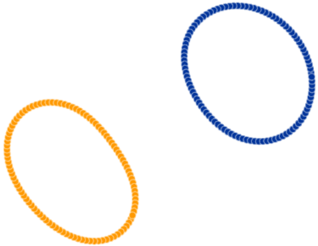
t-SNE

t-Distributed Stochastic Neighbor Embedding

A method of visualization of high-dimensional vector spaces in 2D

It keeps the very similar data points close together in lower-dimensional space





... more about dimensionality reduction in **December**

Sparse Auto-Encoders

Sparse Auto-Encoders

Dimensionality "explosion", but with sparse and interpretable features.

- preserve as much information as possible
- hidden layer representation should be sparse