Introduction

David Mareček

October 5, 2020
Webpage: http://ufal.mff.cuni.cz/courses/npfl097
E-credits: 3
Examination: 1/1 C
Form:
• 8 lectures
• 3 programming sessions
• discussions
Schedule

- **Oct 5**: Introduction to Unsupervised ML
- **Oct 12**: Beta-Bernoulli probabilistic model
- **Oct 19**: Dirichlet-Multinomial probabilistic model, Modeling document collections
- **Oct 26**: Categorical Mixture Models, Expectation-Maximization
- **Nov 2**: Gibbs Sampling, Latent Dirichlet Allocation
- **Nov 9**: Assignment 1 - Latent Dirichlet Allocation
- **Nov 16**: Chinese Restaurant Process, Text Segmentation, Tagging, Parsing
- **Nov 23**: Assignment 2 - Text Segmentation
- **Nov 30**: Clustering Evaluation, Mixture of Gaussians
- **Dec 7**: Principal Component Analysis, Independent Component Analysis
- **Dec 14**: Assignment 3 - Component Analysis
- **Dec 21**: Unsupervised Interpretation of Neural Networks
- **Jan 4**: TBA
Course passing requirements

Three programming assignments:
• for each one, 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**.
Prerequisites and related courses

Basic probabilistic and ML concepts:
- NPFL067 – Statistical Methods in NLP I
- NPFL129 – Introduction to Machine Learning with Python

Basic deep-learning concepts:
- NPFL114 – Deep Learning

Other related courses:
- NPFL087 – Statistical Machine Translation
- NPFL103 – Information Retrieval
- NPFL120 – Multilingual Natural Language Processing
Books:
• Christopher Bishop: Pattern Recognition and Machine Learning, Springer-Verlag New York, 2006

Tutorials, papers:
• Kevin Knight: Bayesian Inference with Tears, 2009
  (https://www.isi.edu/natural-language/people/bayes-with-tears.pdf)
All the lectures are going to be recorded.

- The recordings will be available in SIS.
- Only the students enrolled in this course will have access to the recordings.
- In case you do not want to be recorded and have a question, you can use the Zoom chat.
Unsupervised Machine Learning
Supervised Machine Learning

![Graph showing the relationship between Twitter followers and salary.](image-url)
Unsupervised Machine Learning

Unsupervised ML Problems

Basic concepts
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.

The machine learning methods which are optimized on one task but then their internal learned representations are taken as outputs may also be called *unsupervised*. 
Unsupervised ML Problems
We want to find an underlying structure of a given set of 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.

*related course:*
NPFL103 – Information Retrieval
Modelling Document Collections

Latent Dirichlet Allocation

**Topics**
- gene: 0.04
- dna: 0.02
- genetic: 0.01
- life: 0.02
- evolve: 0.01
- organism: 0.01
- brain: 0.04
- neuron: 0.02
- nerve: 0.01
- data: 0.02
- number: 0.02
- computer: 0.01

**Documents**

**Topic proportions and assignments**

---

**Seeking Life’s Bare (Genetic) Necessities**

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known sequences, concluded that today’s bacteria can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a single parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 120 wouldn’t be enough. Although the numbers don’t match precisely, those predictions are not all that far apart, especially in comparison to the 75,000 genes in the human genome, notes Stu Andronica, a biochemist at the University of Colorado, who published last week an analysis of 800 microbes. But coming up with a consensus answer may be more than just a genetic problem. More and more scientists are realizing that some genes are sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arca Mbuta, a computational molecular biologist at the National Center for Biotechnology Information in Bethesda, Maryland. Computing an...
Language Clustering

- What is the underlying structure of all world’s languages?
- Do they fit their linguistic categorization to language families?

*Language vectors from multilingual MT visualized by T-SNE*
Text Segmentation

We need some language units for processing text.

- Paragraphs or sentences are too long, characters are too short.
- Words? 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美，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：“如果你们想让我空心，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。
Embeddings of Linguistic Units

• Word Embeddings
• Contextual Embeddings
• Sentence Embeddings

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 the methods like Skipgram, ELMo, and BERT unsupervised?

related course:
NPFL114 – Deep learning
Unsupervised Machine Translation

- Let’s suppose we have a 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
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?
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
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.

\[ P(x) \approx \frac{n_x}{n}, \quad P(x) = \lim_{x \to \infty} \frac{n_x}{n} \]

**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?

\[ P(x) \approx \frac{n_x + \alpha_x}{n + \alpha} \]
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.