Introduction

David Mareček

October 4, 2018
Course overview

Bayesian inference in unsupervised learning tasks: (first 2/3 of the seminar)

- Beta-Bernoulli and Dirichlet-Categorical models
- Categorical mixture models
- Expectation-Maximization
- Metropolis-Hastings, Gibbs sampling
- Chinese restaurant process, Pitman-Yor process

Other topics:

- Interpretation of deep neural networks
- Sentence structures in Machine Translation
- ... your topics
Course prerequisites

Basic probabilistic and ML concepts
- NPFL067 - Statistical methods in NLP I
- NPFL054 - Introduction to Machine Learning (in NLP)

Basic deep-learning concepts
- NPFL114 - Deep Learning

Other related courses
- NPFL104 - Machine Learning Methods
- NPFL087 - Statistical Machine Translation
Assignments

There will be three programming assignments:

1. Topic modeling - Latent Dirichlet Allocation (LDA)
2. Unsupervised text segmentation
3. Finding Motifs in DNA

Preferred language: Python

For each homework there will be a time-slot reserved for questions and discussions over preliminary results.
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**.
Preliminary course schedule

- 4/10 - Introduction, Warm-Up test
- 11/10 - Beta-Bernouli and Dirichlet-Categorial models
- 18/10 - Categorial mixture models, Expectation-Maximization
- 25/10 - Monte Carlo Markov Chain, Gibbs Sampling
- 1/11 - cancelled (conference)
- 8/11 - Latent Dirichlet Allocation
- 15/11 - Chinese restaurant process, Pitman-Yor process
- 22/11 - cancelled (DOD)
- 29/11 - Unsupervised text segmentation
- 6/12 - Finding Motifs in DNA
- 13/12 - Interpretation of Deep Neural Networks
- 20/12 - Sentence structures in Machine Translation
- 3/1 - TBA
- 10/1 - TBA
Latent Dirichlet Allocation

Unsupervised learning method for finding topics in unsorted collections of text documents.

You obtain a topic model, in which:

- Each topic is represented by a distribution over words.
- Each document has a distribution over its topics (typically 1 to 5 main topics)
- Total amount of topics is the only constant chosen by the user
Latent Dirichlet Allocation [2]

Introduction

Outline

Unsupervised problems

Test

Basic concepts

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 genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other research team, mapping genes in a simple parasite and estimating that this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions are not all that far apart,
大颠瓦法，为佛教密宗无上瑜伽部的甚深密法。在藏传佛教各大教派中最著名的仍是直贡大颠
瓦法。有许多信徒不远万里来到直贡受教颇瓦法或希望从直贡噶举派上师处得到此法。享
誉海内外的直贡颇瓦法，分为口传颇瓦法和伏藏颇瓦法两类。口传颇瓦法，由初期佛持金刚
传来，又分为语旨觉受传承和证语意义传承。语旨觉受传承是四大语旨教授由金刚持依次传
授因陀罗，龙树，玛当吉，而至底洛巴。其法有父续密集及圆满次第五道直教和颇瓦法。证悟意
义传承是底洛巴亲自从金刚持处聆听，依次传授给那若巴，玛尔巴，米拉日巴，塔布拉吉，
帕木珠巴，觉巴大师，止至今天其耳传甚深密法，传承从未间断。此颇瓦法称之为彩虹颇
瓦法，因得此颇瓦法临终出现彩虹迷漫,故此得名。

伏藏颇瓦法，由阿弥陀佛依次传授给莲花生，赤颂的大巨尼玛。尼玛将其藏在塔拉岗布山后
的黑曼陀罗湖中。后来，由大巨尼玛的转世牧羊人尼达桑杰发掘并依次传授贤士。囊卡坚赞
，多旦。更堆桑布，帕果。智美洛珠，法王桑杰坚赞，门才。坚赞教授，直贡羊日岗。堪布洛。平
措朗杰。此后，由直贡噶举做为甚深密法发扬光大。此颇瓦法，称之为入草颇瓦法，因
其颇瓦法成功后，在头盖上能插入茅草，故此得名。颇瓦，意为迁识，使亡者不经过中阴，
将灵魂往生净土之法。据说，接受此甚深密法颇瓦法，从人的头盖骨里出黄水或能插入茅
草等现象。接受直贡颇瓦法，使人立刻感觉头疼，头顶渗出水珠，晕倒，打通梵净穴，流鼻
血等现象，故此灵验而闻名于世
(欲探详情，请参阅[捷径颇瓦法彩虹修行次第概要。自显自显虹光]，[捷径颇瓦法红哄教授。
利众自焰]等口传颇瓦法典，以及[不修成佛之法甚深密法。入草颇瓦法]等法典)。
There are languages not separating individual words by spaces.

Some of them does not even have any formal definitions for words.

However, we need some units for further processing of such texts (translation, search, analysis).

We will show a fully unsupervised solution for cases where no other resources (e.g. dictionaries) are available.
Warm-Up Test
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?

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