Machine Learning and Pervasive Computing

Stephan Sigg

Georg-August-University Goettingen, Computer Networks

06.07.2015

Overview and Structure

13.04.2015	Organisation
13.04.2015	Introduction
20.04.2015	Rule-based learning
27.04.2015	Decision Trees
04.05.2015	A simple Supervised learning algorithm
11.05.2015	–
18.05.2015	Excursion: Avoiding local optima with random search
25.05.2015	–
01.06.2015	High dimensional data
08.06.2015	Artificial Neural Networks
15.06.2015	k-Nearest Neighbour methods
22.06.2015	Probabilistic graphical models
29.06.2015	Topic models
06.07.2015	Unsupervised learning
13.07.2015	Anomaly detection, Online learning, Recom. systems

Outline

Introduction

Probabilistic Topic Models

Extraction of topics

mvote.ugoe.de/2829

Introduction

Topic models Documents are mixtures of topics

Topics Probability distribution over words

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

mvote.ugoe.de/2829

Introduction

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Assumption

The process of generating documents is to iteratively

1. choose a topic
2. draw word from that topic wrt topic's distribution

mvote.ugoe.de/2829

Introduction

Generative models

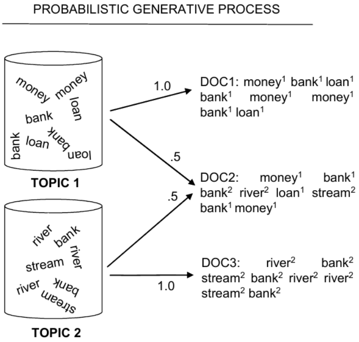
Generative models for documents describe probabilistic sampling rules which describe how words in documents might be generated based on latent (random) variables

mvote.ugoe.de/2829

Introduction

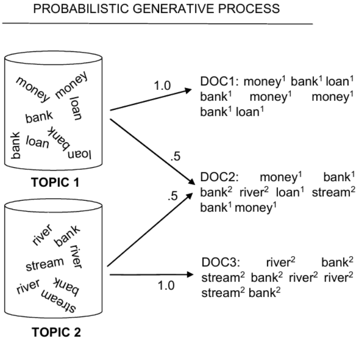
Generative models

Generative models for documents describe probabilistic sampling rules which describe how words in documents might be generated based on latent (random) variables



Introduction

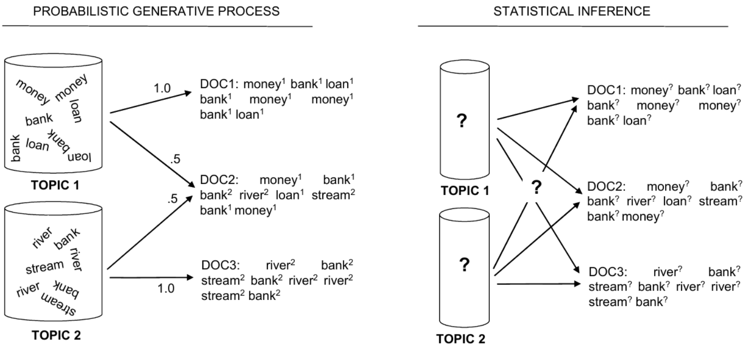
→ Fitting a generative model is the task of finding the best set of latent variables that might explain the observed words in a document



mvote.ugoe.de/2829

Introduction

→ In statistical inferences we would like to know what topic model is most likely to have generated the data



mvote.ugoe.de/2829

Bag-of-words

Generative models do not take any assumption on the order of words in documents but only on their frequency. This is called the bag-of-words assumption

[1] Griffiths, Steyvers, Blei and Tenenbaum, Integrating topics and syntax, Advances in Neural Information Processing, 2005

mvote.ugoe.de/2829

Bag-of-words

Generative models do not take any assumption on the order of words in documents but only on their frequency. This is called the bag-of-words assumption

→ Since word-order might contain important cues to the content of a document, also Topic models which are sensitive to word-order have been defined too.[1]

[1] Griffiths, Steyvers, Blei and Tenenbaum, Integrating topics and syntax, Advances in Neural Information Processing, 2005

mvote.ugoe.de/2829

Outline

Introduction

Probabilistic Topic Models

Extraction of topics

mvote.ugoe.de/2829

Probabilistic Topic Models

Notation

$\mathcal{P}[z]$ Distribution over topics z in a particular document

$\mathcal{P}[w|z]$ Probability distribution over words w given topic z

w_i i-th word in a document

$\mathcal{P}[z_i = j]$ Probability that the j-th topic was sampled for the i-th word token

$\mathcal{P}[w_i|w_i = j]$ Probability of word w_i under topic j

mvote.ugoe.de/2829

Probabilistic Topic Models

Notation

$\mathcal{P}[z]$ Distribution over topics z in a particular document

$\mathcal{P}[w|z]$ Probability distribution over words w given topic z

w_i i-th word in a document

$\mathcal{P}[z_i = j]$ Probability that the j-th topic was sampled for the i-th word token

$\mathcal{P}[w_i|w_i = j]$ Probability of word w_i under topic j

Assumption: Each w_i in a document was generated by first sampling a topic from the topic distribution and then choosing a word from the topic-word distribution.

mvote.ugoe.de/2829

Probabilistic Topic Models

Notation

$\mathcal{P}[z\ ]$ Distribution over topics z in a particular document

$\mathcal{P}[w|z\ ]$ Probability distribution over words w given topic z

w_i i-th word in a document

$\mathcal{P}[z_i = j\ ]$ Probability that the j-th topic was sampled for the i-th word token

$\mathcal{P}[w_i|w_i = j\ ]$ Probability of word w_i under topic j

Distribution over words within a document: ($T \equiv \#$ of topics)

$$\mathcal{P}[w_i] = \sum_{j=1}^{T} \mathcal{P}[w_i|z_i = j]\mathcal{P}[z_i = j]$$

mvote.ugoe.de/2829

Probabilistic Topic Models

Multinomial distribution

Generalization of the binomial distribution.

For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability (Topics in documents)

mvote.ugoe.de/2829

Probabilistic Topic Models

Multinomial distribution

Generalization of the binomial distribution.

For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability (Topics in documents)

$\rightarrow$ multinomial distribution gives probability of any particular combination of numbers of successes for the categories

mvote.ugoe.de/2829

Probabilistic Topic Models

Multinomial distribution

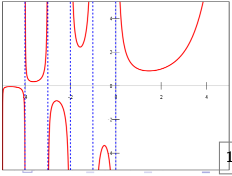
Generalization of the binomial distribution.

For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability (Topics in documents)

$\rightarrow$ multinomial distribution gives probability of any particular combination of numbers of successes for the categories

Let $p = (p_1, \ldots, p_T)$ be a multinomial distribution and Γ the gammafunction.

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \, dx.$$



mvote.ugoe.de/2829

Probabilistic Topic Models

Multinomial distribution

Generalization of the binomial distribution.

For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability (Topics in documents)

$\rightarrow$ multinomial distribution gives probability of any particular combination of numbers of successes for the categories

Let $p = (p_1, \dots, p_T)$ be a multinomial distribution and Γ the gammafunction.

$\rightarrow$ Latent Dirichlet Allocation (LDA) is then a generative model:

$$Dir(p, \alpha_1, \dots, \alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

mvote.ugoe.de/2829

Probabilistic Topic Models

$$Dir(p, \alpha_1, \ldots, \alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma\left(\alpha_j\right)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

mvote.ugoe.de/2829

Probabilistic Topic Models

$$Dir(p, \alpha_1, \dots, \alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

a_j can be interpreted as prior observation count for the number of times topic j is sampled in a document

→ Usually, choose $\alpha_1, \dots, \alpha_T = \alpha$

mvote.ugoe.de/2829

Probabilistic Topic Models

$$Dir(p, \alpha_1, \dots, \alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

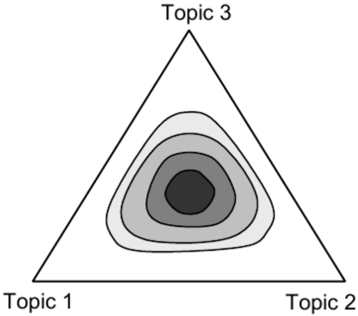
a_j can be interpreted as prior observation count for the number of times topic j is sampled in a document

$\rightarrow$ Usually, choose $\alpha_1, \dots, \alpha_T = \alpha$

Placing a Dirichlet prior on the topic distribution θ will result in a smoothed topic distribution.

Amount of smoothing determined by α

Probabilistic Topic Models



Symmetric Dirichlet distribution for three topics.

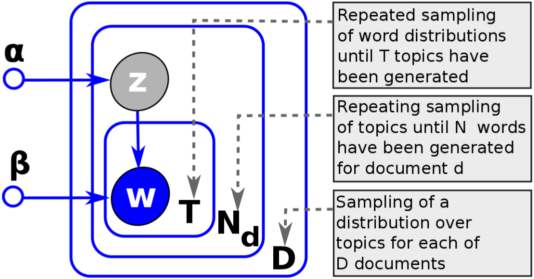
Left: $\alpha = 4$;

Right: $\alpha = 2$

→ Darker colors indicate higher probability; $\sum_j p_j = 1$

mvote.ugoe.de/2829

Probabilistic Topic Models

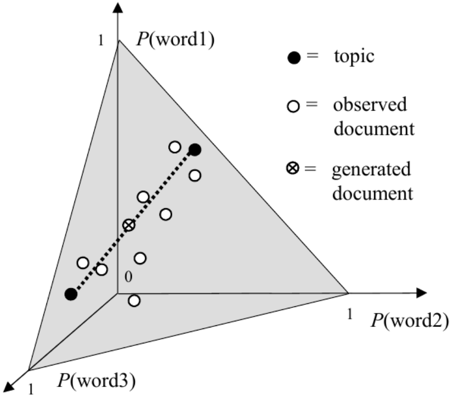


Graphical model using plate notation

α, β parameters to control the prior distribution over topics and documents

mvote.ugoe.de/2829

Probabilistic Topic Models

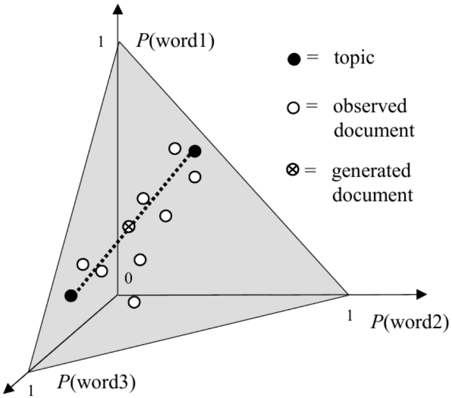


Geometric interpretation of probabilistic topic models

Probability distribution over words given by the $w - 1$ dimensional simplex (since $\sum_j p_j = 1$)

mvote.ugoe.de/2829

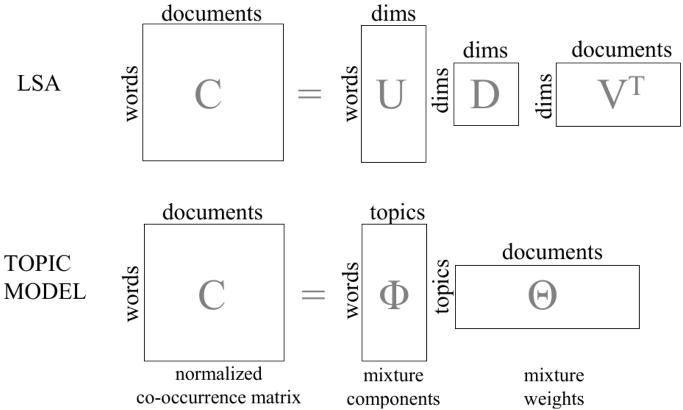
Probabilistic Topic Models



For $T << W$, the topics span a lower dimensional subsimplex. Projection of each document onto that subsimplex can be thought of as dimensionality reduction

mvote.ugoe.de/2829

Probabilistic Topic Models



Interpretation as Matrix Factorization

Probability distribution over words given by the W-1 dimensional simplex (since $\sum_j p_j = 1$)

mvote.ugoe.de/2829

Outline

Introduction

Probabilistic Topic Models

Extraction of topics

Extraction of topics

The main variables of interest for topic models are

topic-word distributions $\phi = \mathcal{P}[w|z]$

topic distributions over documents $\theta = \mathcal{P}[z]$

Extraction of topics

mvote.ugoe.de/2829

The main variables of interest for topic models are

topic-word distributions $\phi = \mathcal{P}[w|z]$

topic distributions over documents $\theta = \mathcal{P}[z]$

Approaches to estimate these distributions

Expectation Maximisation (EM)

Markov chain Monte Carlo (MCMC)

Extraction of topics – EM

mvote.ugoe.de/2829

Expectation Maximisation Algorithm

An iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models which depend on unobserved latent variables

Extraction of topics – EM

mvote.ugoe.de/2829

Expectation Maximisation Algorithm

An iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models which depend on unobserved latent variables

The EM algorithm starts with an initial estimation of parameters

mvote.ugoe.de/2829

Extraction of topics – EM

Expectation Maximisation Algorithm

An iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models which depend on unobserved latent variables

The EM algorithm starts with an initial estimation of parameters

$\rightarrow$ an interation of the EM algorithm consits of the steps

Expectation step Calculation for the expectation of the log-likelihood based on current parameter estimates (the latent variables)

Maximisation step Computing parameters which maximise the expected log-likelihood

mvote.ugoe.de/2829

Extraction of topics – EM

E and M steps for topic models:

$\rightarrow$ an interation of the EM algorithm consits of the steps

$(n(d, w)$ specifies how often w occurs in document $d)$

mvote.ugoe.de/2829

Extraction of topics – EM

E and M steps for topic models:

$\rightarrow$ an interation of the EM algorithm consits of the steps

Expectation step

$$\mathcal{P}[z|d,w] = \frac{\mathcal{P}[z]\mathcal{P}[d|z]\mathcal{P}[w|z]}{\sum_{z'\in\mathcal{Z}}\mathcal{P}[z']\mathcal{P}[d|z']\mathcal{P}[w|z']}$$

($n(d,w)$ specifies how often w occurs in document d)

mvote.ugoe.de/2829

Extraction of topics – EM

E and M steps for topic models:

→ an interation of the EM algorithm consits of the steps

Expectation step

$$\mathcal{P}[z|d,w] = \frac{\mathcal{P}[z]\mathcal{P}[d|z]\mathcal{P}[w|z]}{\sum_{z'\in\mathcal{Z}}\mathcal{P}[z']\mathcal{P}[d|z']\mathcal{P}[w|z']}$$

Maximisation step

$$\mathcal{P}[w|z] \propto \sum_{d\in\mathcal{D}} n(d,w)\mathcal{P}[z|d,w]$$

$$\mathcal{P}[d|z] \propto \sum_{w\in\mathcal{W}} n(d,w)\mathcal{P}[z|d,w]$$

$$\mathcal{P}[z] \propto \sum_{d\in\mathcal{D}}\sum_{w\in\mathcal{W}} n(d,w)\mathcal{P}[z|d,w]$$

($n(d,w)$ specifies how often w occurs in document d)

mvote.ugoe.de/2829

Outline

Introduction

Probabilistic Topic Models

Extraction of topics

mvote.ugoe.de/2829

Questions?

Stephan Sigg
stephan.sigg@cs.uni-goettingen.de

mvote.ugoe.de/2829

Literature

- C.M. Bishop: Pattern recognition and machine learning, Springer, 2007.
- R.O. Duda, P.E. Hart, D.G. Stork: Pattern Classification, Wiley, 2001.