INTRODUCTION TO DATA SCIENCE

This lecture is based on course by E. Fox and C. Guestrin, Univ of Washington

5/12/2017

WFAiS UJ, Informatyka Stosowana II stopień studiów

What we've learned so far

Nearest neighbor search

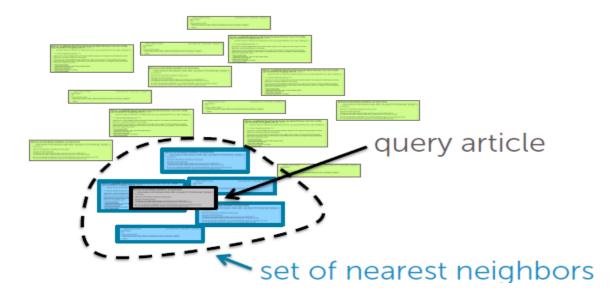
1-NN search

Space of all articles, organized by similarity of text



k-NN search

Space of all articles, organized by similarity of text

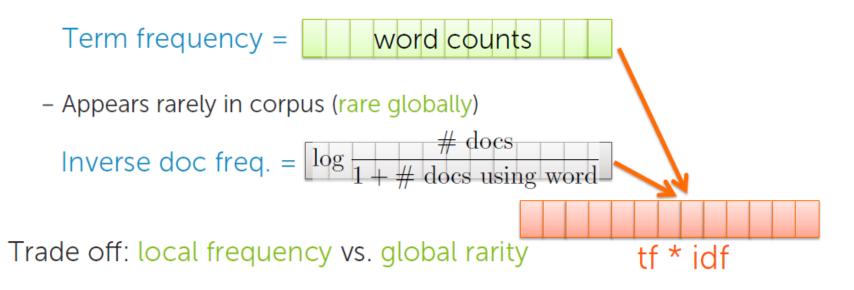


Э

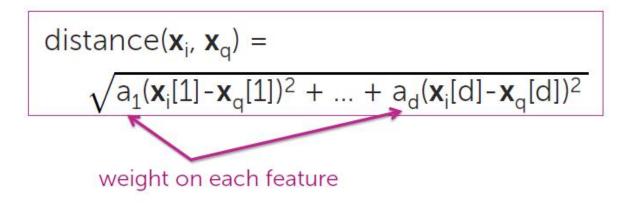
TF-IDF document representation

Emphasizes important words

Appears frequently in document (common locally)



Scaled Euclidean distance

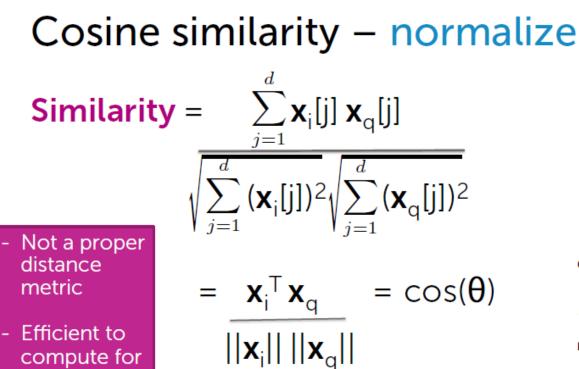


Advant & Assessed	Real Transport Artifician
at hit second of the last in proceeding the other	and the second se
printing provides of case have been been as only in the	Name of all large \$2 to an other of the other factor of the
first of the Dankerson of the spatial data business over	or the surgery of Family and Loss Langels. They below a
diding disandrong	Annalized party or mother of a doring providence
I colored the day, pel pe day over	family in first final to on or owned that restance to the the
denotes for a first a first of the last	with annuality in marks of scheduly for some
printed for the based of the last of the based	and all the B is the figure of the and the first of
	Two are reprint to be want of (W3, do-test
the factory could assert part using the risk could show on pri- tice working particular or could assert colors are bands. Not	passes of the burghts brush out which if you have it does
	and the set of the backs of the set of the s
and an information of any other stands of the formula	the I was based the one had I and not prove a fill
	the 1 year bags due one had 1 and we prove a latter automatic has an inferential respect to estimation in
and then to have degree another her build upon the control to	
	Transmitted as programmed at the billing of
Payled density at a breshall a loss of an at feat	terms in the delated are produce finghed in some
moving sile to all out and to difference of a sile	the its Bank sight on its could of Sping a compared
of Sector Spiner & Sector Son & Arms print	Sunan it was have been access to the Hall had
professor promote of vessel figures for non-tag latter	Equipart Privates Research end-dear and heights (.)
Tail works property of the bay style pulsar	And other has minimum. Hence the Care Series And and the in
and the second se	New York and Washington In: the total is particular
Press was notinger press as the presigned of one even	pright to provide other an inspiring theories are write one of it.
Breeds MARL to had black a new beards with	But Repty that Billion as Desph Louis A constant in
ar responsible work). Though I have long store response	apper Prodite Manager & manager
the angles one Third more gesteed to storand	Total factor that the confidence for used 2 month rate
mak, Wi-Appended spin on another Weissmool Red and	here they have adden as it and a subger further of
mate and implements are assisted We beyond these and	and the second se
singlet one analys. No worklaged with one pitches. No	The Banda responses of an Iban James of the Safara
for all new standary. We want manually have a star to starte	the own radiant did to story yields. If we belief you don-
1.7.10	1.000

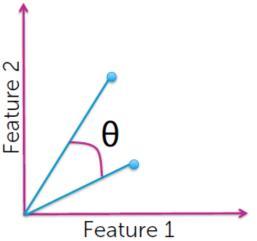
title abstract main body conclusion

Aduat if Apropa	Real Departy & Algorith
a for successing the first succession from the set	and a second state of the second state of the
printing a start at all cases that there is the start of the line	the local logs for manipul, "Bus lots from the in-
the star backward by solar frie before the	or the persons of franch and from Largeth. This holds, or
have described by by he for Subscreen	hardware of bard of the branch. He wanted
I address the star of a second	hand in free back or a press for character in the
interaction, Spring stratemy in the last and half strength	and a survey of a stand of a local to the survey of the su
strend of and have and here of a subsectionary	word first a to have all out in the second
to taken which provide a risk the hot with the error of the	I service just off it has ease of you, the last
to an Ha and the transfer on the last the	menter of shy beacher problems' effect. It say here: if you
our as inferiously suggestive security of the Property-	and the ofpice had all over both long time highest
the party party in magness of Fall and the film.	the I am Rear the sorbal I and an avera in 199
ad data in loss degre modif at held taken ins country's	and sound sharing balanched magnetic in printer depthy-
adius, even with open hours in parts out descents.	· Constrainty on programming the little program
The physical metric game in a single and in the data from	lacena in their other and people in Regime to rearry
service second and a contribution of a second	the in Book age as in the solution of the second
a lattice lighter a linear size is solve prove	Saturday in a contribution of the test in the Part and
photosent all require of laters Tagened The next my latitud.	Replace Peoples Research and the analysis (1)
Barll souths parameters of the lost of the colling	the effected and the New Archard Installation of
the state of the s	- this first and Walksprei in analying a ground rate
from my company proto to the presigner of the school	angle in the effort of basising freedom onto the time 12
Reserved. Maders, on local And are a result for staff which	And Report Best Adherance Analy Enclosed in-
a surprise do critici Norgh (have tog dess respond	Sugar Buddy Assessed protection
for prophets yors. I facts arrest generated in restored	Then have due to tableads in the Local Property rate
and the Appendix ages, see presiden the spectral Real and	Invariant separate on the Cope and they invold it walk, need
and and heplics will us suffice Works and they and	A THE CONTRACT OF A DESCRIPTION OF A DES
apple interpreters for the property of the state of the	Children and the first and the second second
and parameters Witness sample land man a second	10. and other olds to on ty , shift if an inder own dis-
0.238	114500





sparse vecs



To normalize or not?

	And Address of Marcol
on halves verification had to prove the Neural to a provide an end on the state of the Neural Top of t	medit compares, office from the order of the second
And an advertising I continued these things and on the same Annual the galaxies is to be first additional	teer back many or profess of the first procedure band on the back or proceed the common the te- rate generating to describe the back. The comm
anters of a suffrant and has the other settler of blics and entiger device an estimate a si- tle prefer policy in collector with an ends. An	in all Pharins in Acad 20th and a third and increasing operatively accessed third and annual in the basilow out any shart a patient of an
the second descent of any second seco	en den objeten het die en bei het het het begebet. Im 1 en bege die er beken is en progen biet
allian company all control to an Upper all Resonant They have been reported as the second sec	These enterthis recoming an encoder of the first sector of the sector of the Market sector of the Market System of the Antibert Sector of the sector of Spherey recomposed and the Sector of Sect
And the figure of the set of the	Company to cault have been according to the fact that depleted if backets benefit with forecasts remain to the day day for the same benefit (age have beeneded in the day of the best of the same beam of the fact of the off the same beam of the same beam of the fact of the off the beam of the same beam of the fact of the off the beam of the same beam of the fact of the same beam of the same beam of the fact of the same beam of the fact of the fact of the fact of the fact of the fact of the fact of the fact of the fact of the fact of the fact of the fact of the fact of the
Non-to-compage of the compact of the com- tempted with the second sector of the compact of the compact of the second sector of the sector of the second sector of the sect	See Test and Takington in an excising a generation de- angle scientific allocation (Applications) and the same for the Application of the second science of the same for the application of the second science of the same form
An anything street, I have street particular in monoser- and and compared with the street of the street in- man will have not an another We benefiting and	And farm the in testing, is unit read on bootstance and tags within south will will
ingle on ander Te serliged of an ander Te best-an ander We en analy heit on the or	To half, any singular line from it is follow in our entire allowing with Raylolar words:

short tweet

Normalizing can make dissimilar objects appear more similar



Common compromise: Just cap maximum word counts

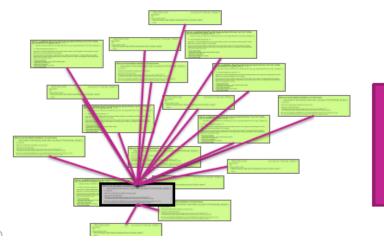
8

9

Complexity of brute-force search

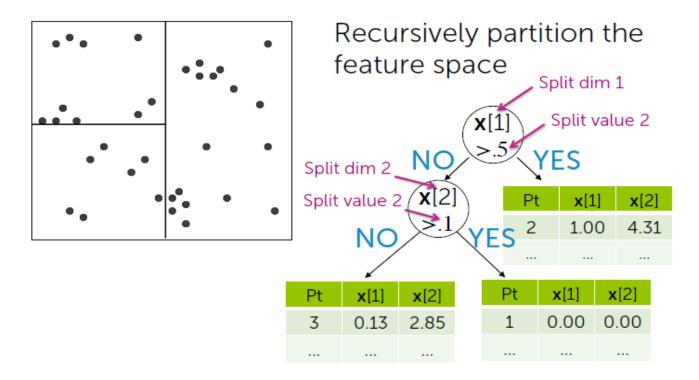
Given a query point, scan through each point

- O(N) distance computations per 1-NN query!
- O(Nlogk) per k-NN query!



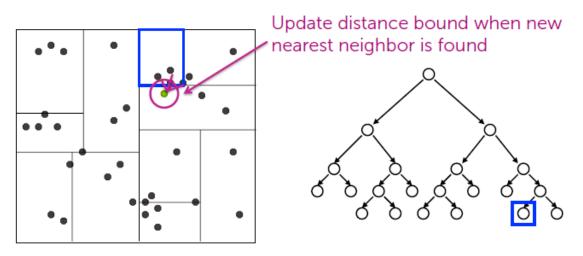
What if *N* is huge??? (and many queries)

KD-trees



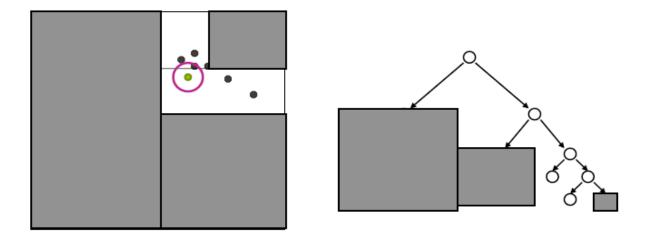
11

Nearest neighbor with KD-trees



- 1. Start by exploring leaf node containing query point
- 2. Compute distance to each other point at leaf node
- 3. Backtrack and try other branch at each node visited

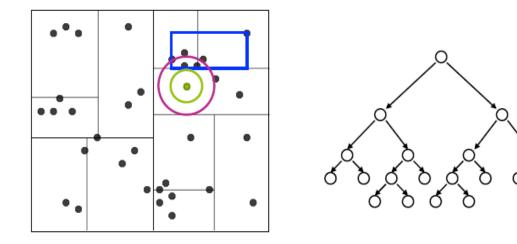
Nearest neighbor with KD-trees



Use distance bound and bounding box of each node to prune parts of tree that cannot include nearest neighbor

12

Approximate k-NN with KD-trees

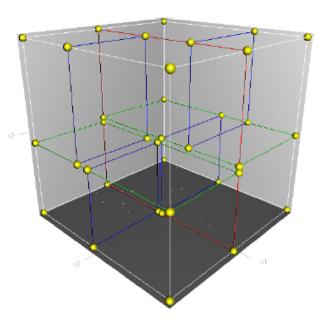


Before: Prune when distance to bounding box > r **Now:** Prune when distance to bounding box > r/α

Saves lots of search time at little cost in quality of NN!

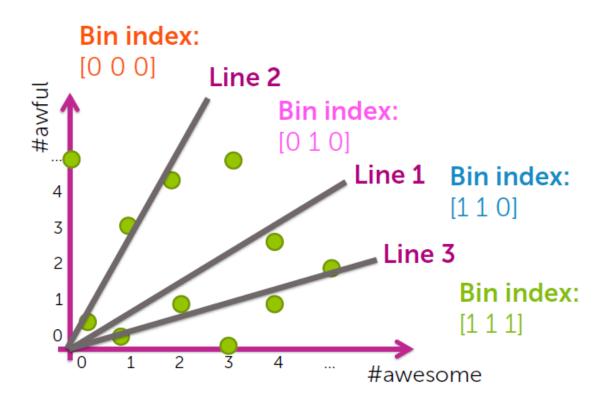
Limitations of KD-trees

- Difficult to implement
- Don't tend to perform well in high dimensions
 - Under some conditions, visit at least 2^d nodes



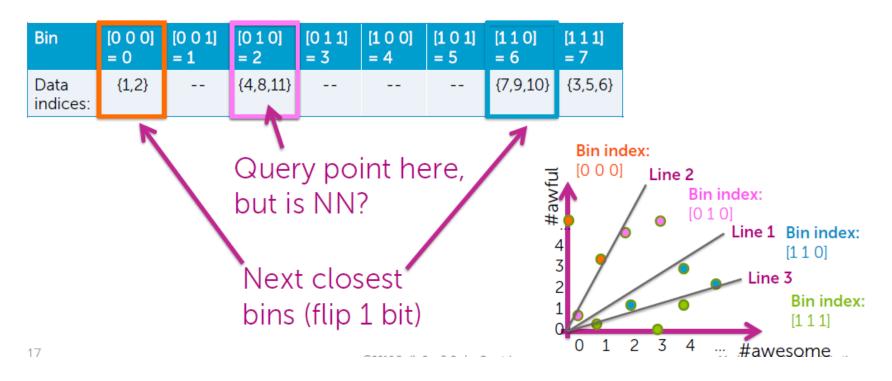
Locality sensitive hashing

15



16

LSH for approximate NN search



What we've learned so far

17

k-means and MapReduce

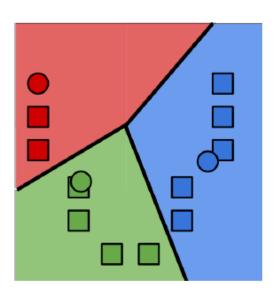


Discover clusters of related documents



k-means algorithm

- 0. Initialize cluster centers
- 1. Assign observations to closest cluster center
- 2. Revise cluster centers as mean of assigned observations
- 3. Repeat 1.+2. until convergence



A coordinate descent algorithm

1. Assign observations to closest cluster center

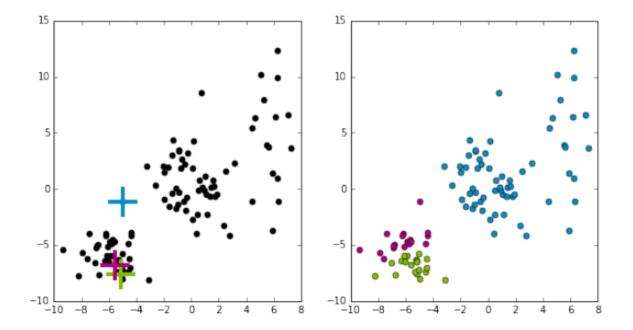
$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

2. Revise cluster centers as mean of assigned observations

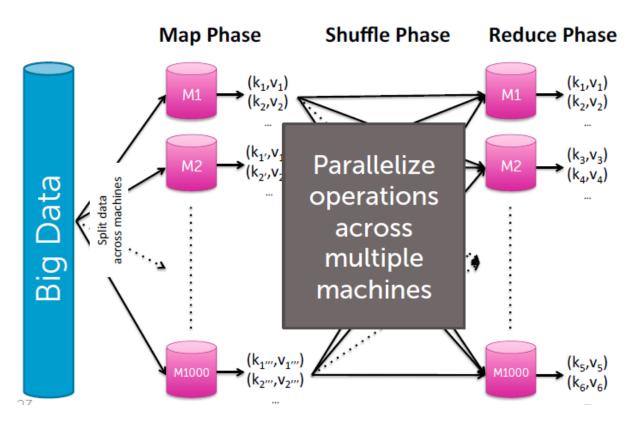
$$\mu_j \leftarrow \arg\min_{\mu} \sum_{i:z_i=j} ||\mu - \mathbf{x}_i||_2^2$$

Alternating minimization 1. (z given μ) and 2. (μ given z) = coordinate descent

Convergence of k-means to local mode



MapReduce framework



MapReduce abstraction

 Map: Data-parallel over elements, e.g., documents Generate (key,value) pairs "value" can be any data type 	Word count example: map(doc) for word in doc emit(word,1)
 Reduce: Aggregate values for each key Must be commutative-associative operation Data-parallel over keys Generate (key,value) pairs 	<pre>reduce(word, counts_list) c = 0 for i in counts_list c += counts_list[i] emit(word, c)</pre>

MapReduce has long history in functional programming

- Popularized by Google, and subsequently by open-source Hadoop implementation from Yahoo!

MapReducing 1 iteration of k-means

Classify: Assign observations to closest cluster center

 $z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$

Map: For each data point, given $({\mu_j}, \mathbf{x}_i)$, emit (z_i, \mathbf{x}_i)

Recenter: Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i:z_i = k} \mathbf{x}_i$$

Reduce: Average over all points in cluster j ($z_i = k$)

What we've learned so far

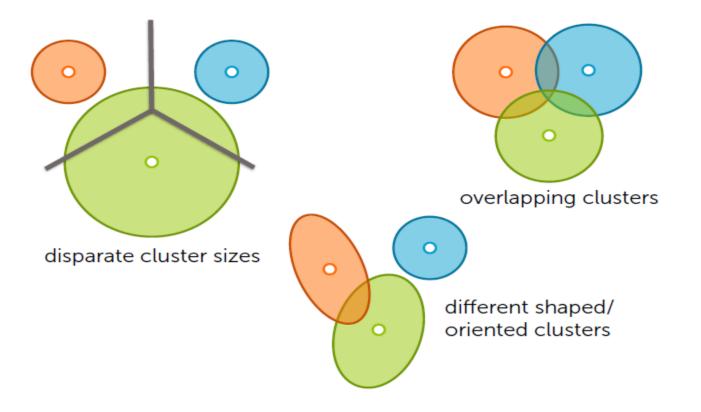
Mixture models

Probabilistic clustering model



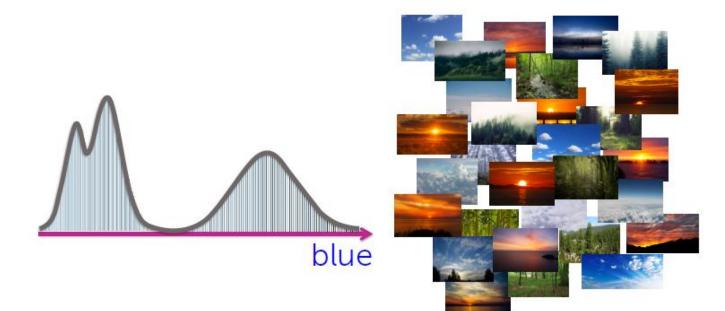
captures uncertainty in clustering

Failure modes of k-means

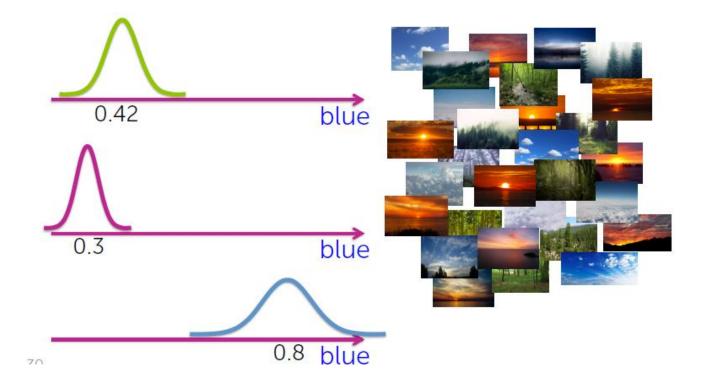




Jumble of unlabeled images

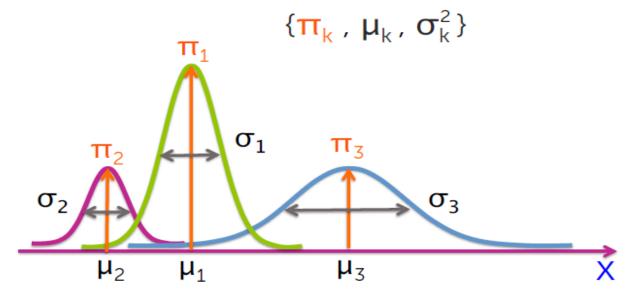


Model of jumble of unlabeled images



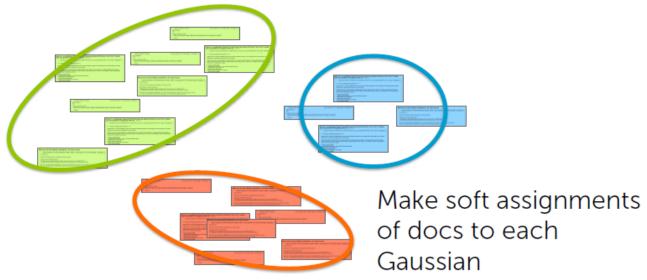
Mixture of Gaussians (1D)

Each mixture component represents a unique cluster specified by:



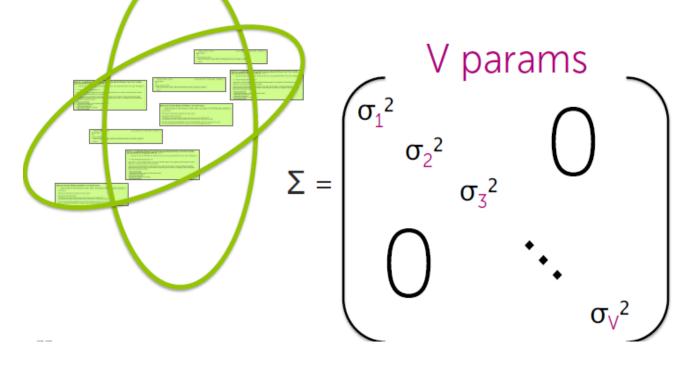
Mixture of Gaussians for clustering documents

Space of all documents (really lives in \mathbf{R}^{V} for vocab size V)



Restricting to diagonal covariance

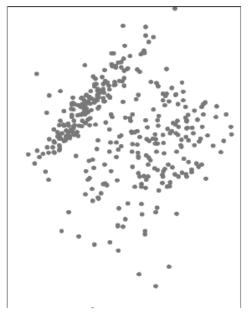


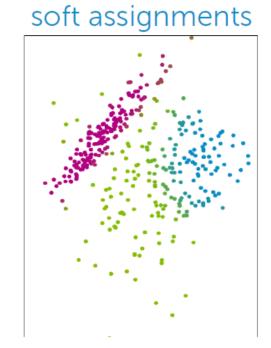


Inferring cluster labels

EM algorithm \rightarrow

Data





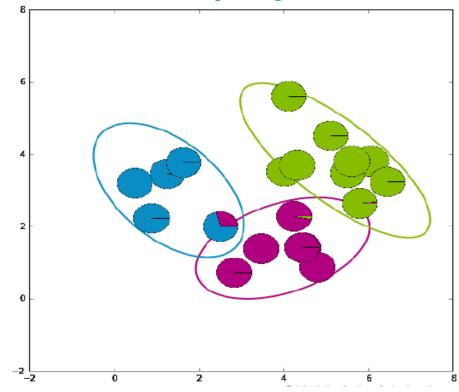
Expectation maximization (EM): An iterative algorithm

1. E-step: <u>e</u>stimate cluster responsibilities given current parameter estimates

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^K \hat{\pi}_j N(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j)}$$

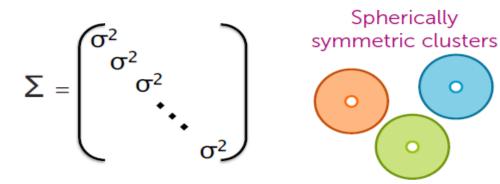
2. M-step: <u>m</u>aximize likelihood over parameters given current responsibilities $\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}_k \mid \{\hat{r}_{ik}, x_i\}$

EM for mixtures of Gaussians in pictures - replay



Relationship to k-means

Consider Gaussian mixture model with



and let the variance parameter $\sigma \rightarrow 0$

Datapoint gets fully assigned to nearest center, just as in k-means

What we've learned so far

Latent Dirichlet allocation



Topic vocab
distributions:

SCIENCE		
experiment	0.1	
test	0.08	
discover	0.05	
hypothesize	0.03	١
climate	0.01	ľ
		L

TECH	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02

SPORTS	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^e, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA ^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA ^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

with enilepsy can manifest short, sub-clinical epileptic "bursts" in blown clinical sciences. We believe the relationship between s of events something not previously studied quantitativelyeld important insights into the nature and intrinsic dynamics of our work is to parse these complex epileptic events A challenge posed by the intracranial EEG dynamic regimes. data we study is the fact that the number and placement of electrodes between patients. We develop a Bayesian nonparametric Markov thing process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of MEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three dec des of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Clustering:

One topic indicator z_i per **document** i

All words come from (get scored under) same topic z_i

Distribution on prevalence of topics in **corpus** $\mathbf{\pi} = [\pi_1 \ \pi_2 \dots \pi_K]$

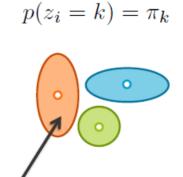
39

Comparing and contrasting

Previously

Prior topic probabilities

Likelihood under each topic



compute likelihood of **tf-idf** vector under each **Gaussian**

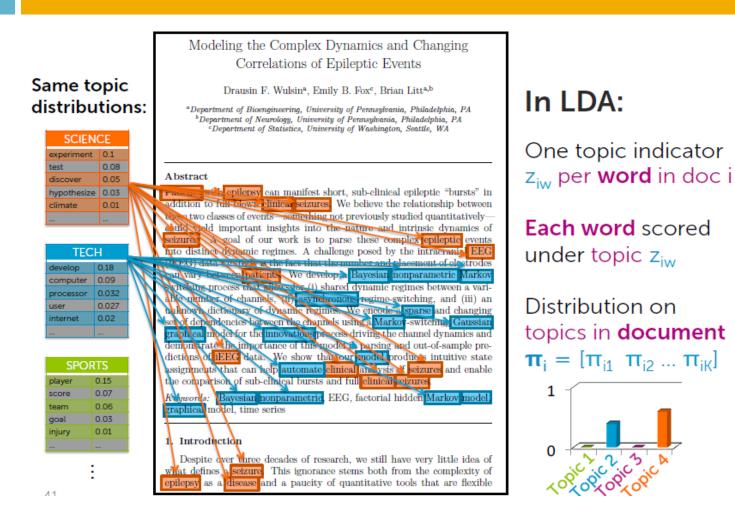


 $p(z_i = k) = \pi_k$

Now

{modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic...}

compute likelihood of the collection of words in doc under each topic distribution



Gibbs sampling for LDA

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01

TOPIC 2	
develop	0.18
computer	0.09
processor	0.032
user	0.027
internet	0.02

TOPIC 3	
player	0.15
score	0.07
team	0.06
goal	0.03
injury	0.01

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pernsylvania, Philadelphia, PA ^bDepartment of Neurology, University of Pernsylvania, Philadelphia, PA ^cDepartment of Statistics, University of Washington, Scattle, WA

Abstract

Patients with epikepsy can manifest short, sub-clinical epikeptic "bursts" in addition to full-blown clinical seizures We believe the relationship between these two classes of events—something not previously studied quantitatively could yield important insights into the nature and intrinsic dynamics of seizures A goal of our work is to parse these complex epikeptia events into distinct dynamic regimes. A challenge posed by the intracrania EEG (iEEC) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable <u>number</u> of channels, (ii) asynchronous regimes witching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a<u>Markor</u>-switching<u>Gaussian</u> maphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of <u>iEEG</u> data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures *Keynords:* Bayesian nonparametric EEG, factorial hidden <u>Markov model</u> graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Step 1: Randomly reassign all z_{iw} based on - doc topic proportions - topic vocab distributions Draw randomly from responsibility vector [r_{iw1} r_{iw2} ... r_{iwK}]

42

Gibbs sampling for LDA

TOPIC 1	
experiment	0.1
test	0.08
discover	0.05
hypothesize	0.03
climate	0.01

TOPIC 2		
develop	0.18	
computer	0.09	
processor	0.032	
user	0.027	
internet	0.02	

TOPIC 3	

÷

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

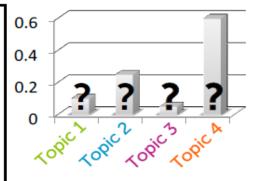
Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA ^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA ^cDepartment of Statistics, University of Washington, Scattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical scizures. We believe the relationship between these two classes of events-something not previously studied quantitativelycould yield important insights into the nature and intrinsic dynamics of seizures A goal of our work is to parse these complex epileptic into distinct dynamic regimes. A challenge posed by the in (iEEG) data we study is the fact that the number and placement of elecan vary between patients. We develop a Bayesian nonparametric switching process that allows for (i) shared dynamic regimes between able number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switchi raphical model for the innovations process driving the channel dynamics and nstrate the importance of this model in parsing and out-of-sample predictions of iNEC data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures Bayesian nonparametric, EEG, factorial hidden graphical model, time series 1. Introduction

Despite over three decades of research, we still have very little idea of at defines a seizure. This ignorance stems both from the complexity of lepsy as a disease and a paucity of quantitative tools that are flexible



Step 2: Randomly reassign doc topic proportions based on assignments z_{iw} in **current doc**

Step 3: Repeat for all docs

43

Gibbs sampling for LDA

Modeling the Complex Dynamics and Changing

Correlations of Epileptic Events

TOPIC 1		
Word 1	?	
Word 2	?	
Word 3	?	
Word 4	?	
Word 5	?	

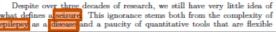
TOPIC 2		
Word 1	?	
Word 2	?	
Word 3	?	
Word 4	?	
Word 5	?	

TOPIC 3		
Word 1	?	
Word 2	?	
Word 3	?	
Word 4	?	
Word 5	?	

Drausin F. Wulsin^{*}, Emily B. Fox^c, Brian Litt^{*,b} ^aDepartment of Bioengineering, University of Permsylvania, Philadelphia, PA ^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA ^cDepartment of Statistics, University of Washington, Seattle, WA stract

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown dinical scizures. We believe the relationship between these two classes of events-something not previously studied quantitativelycould yield important insights into the nature and intrinsic dynamics of seizures A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracrania (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric switching process that allows for (i) shared dynamic regimes between a vari-able number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse et of dependencies between the channels using a Markor-switchin rephica model for the innovations process driving the channel dynamics and emonstrate the importance of this model in parsing and out-of-sample predictions of iEEC data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures nonparametric, EEG, factorial hidden Markov model, time series 1. Introduction





Step 4: Randomly reassign topic vocab distributions based on assignments z_{iw} in **entire corpus**

Collapsed Gibbs sampling for LDA

	0.1	

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^e, Brian Litt^{a,b}

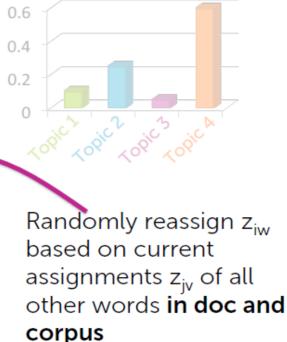
^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA *Department of Neurology, University of Pennsylvania, Philadelphia, PA +Department of Statistics, University of Washington, Scattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events-something not previously studied quantitativelycould yield important insights into the nature and intrinsic dyn A goal of our work is to parse these complex supplic events into distinct dynamic regimes. A challenge posed (iEEG) data we study is the fact that the number can vary between patients. We develop a Bayesian switching process that allows for (i) shared dy able number of channels, (ii) asynchronous regime-switching, and (iii) an and changing unknown dictionary of dynamic regimes. We encode a set of dependencies between the channels using a Markor-switching ruphical model for the innovations process driving the channel dyna process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample pre-dictions of **iEEG** data. We show that our **model** produces intuitive state assignments that can help automate dinical analysis of sciences and enable the comparison of sub-clinical bursts and full clinical seizures Bayesian nonparametric, EEG, factorial hidden Markov model (eywords: model, time series

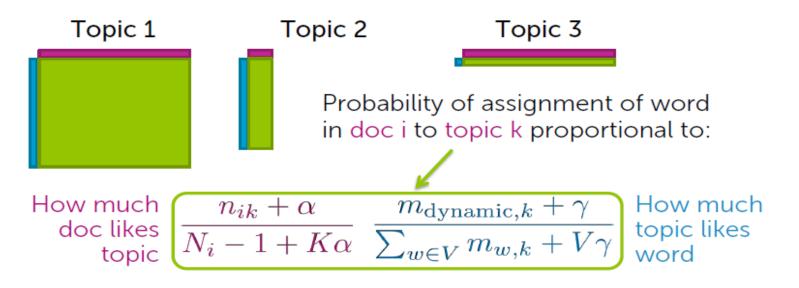
1. Introduction

Despite over three decades of research, we still have very little idea of shat defines a seizure. This ignorance stems both from the complexity of pilepsy as a disease and a paucity of quantitative tools that are flexible



Collapsed conditional distribution

3	?	1	3	1
epilepsy	dynamic	Bayesian	EEG	model



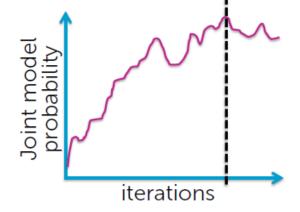
What to do with sampling output?

Predictions:

- 1. Make prediction for each snapshot of randomly assigned variables/parameters (full iteration)
- 2. Average predictions for final result

Parameter or assignment estimate:

 Look at snapshot of randomly assigned variables/parameters that maximizes "joint model probability"



Summary of what we have learned

