

Machine Learning

Topic Models

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 30, 2018

<http://www.carlhenrik.com>

Introduction

Guest Lecture



Dr. Niall Twomey

Lecture: 4th of December

Process → Distribution → value

- Each evaluation of a process is a distribution

$$\mathcal{N}(\mu, \Sigma) \sim \mathcal{GP}(m(x), k(x, x))$$

- Each evaluation of a distribution is a value

$$y \sim \mathcal{N}(y|0, \Sigma)$$

- Parametric models are defined by distributions and non-parametric by processes

Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

- Due to binary outcome

$$p(x = 0|\mu) = 1 - \mu$$

Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

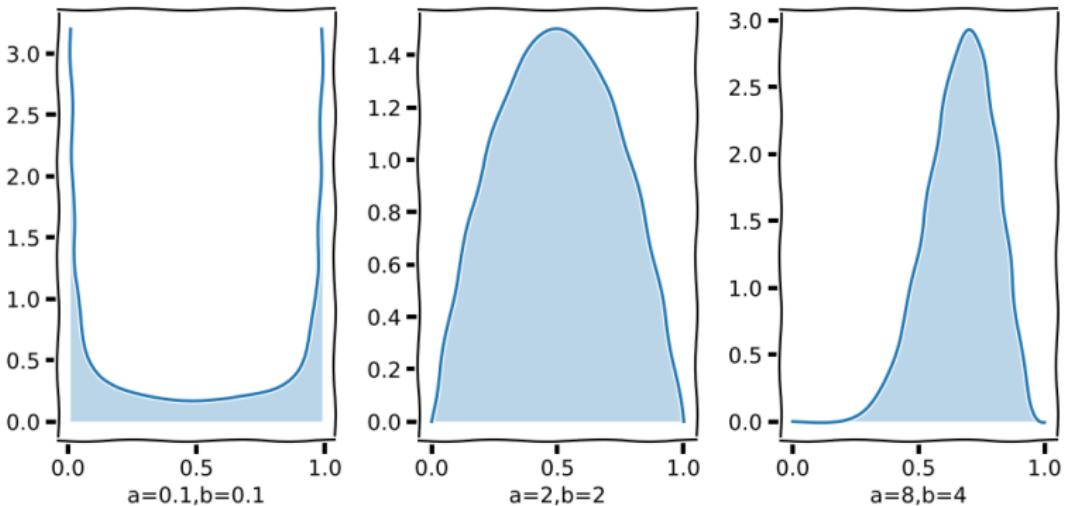
- Due to binary outcome

$$p(x = 0|\mu) = 1 - \mu$$

- Distribution

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}$$

Beta Distribution



$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

Multinomial

- If we have a variable that can take K different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^T$$

Multinomial

- If we have a variable that can take K different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^T$$

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

$$\boldsymbol{\mu} = [\mu_1, \dots, \mu_k]^T, \sum_k \mu_k = 1$$

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

Dirichlet

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

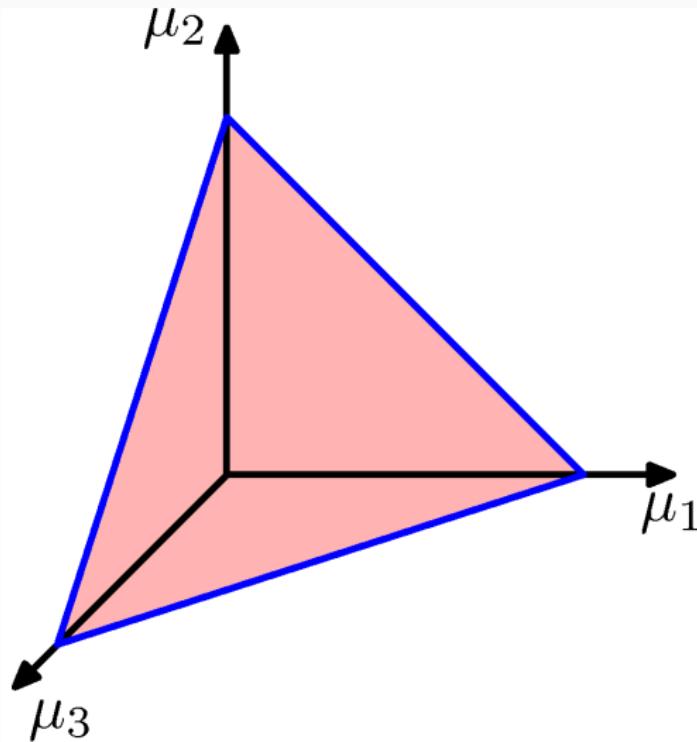
- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

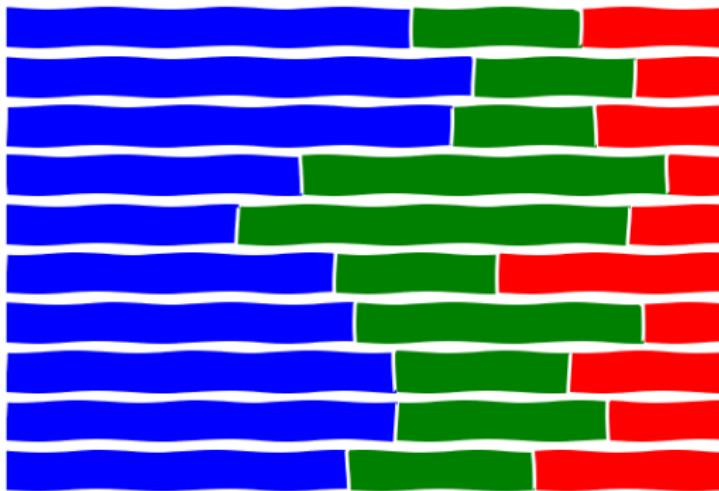
- Dirichlet Distribution

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \dots \cdot \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

Dirichlet Prior

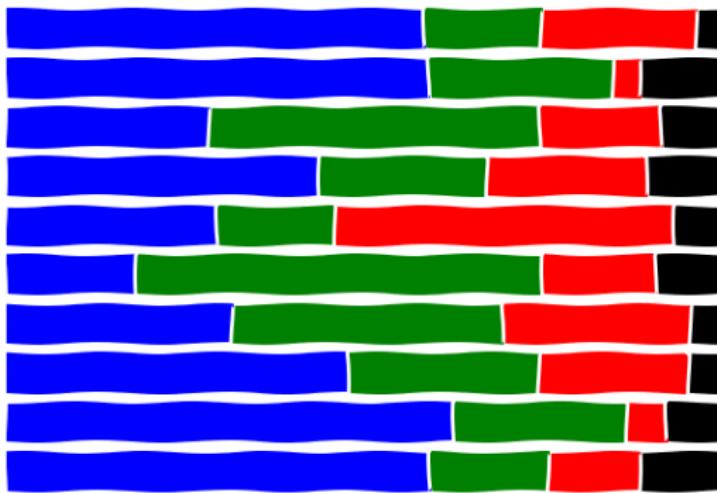


Dirichlet Distribution



$\text{Dir}(10, 5, 3)$

Dirichlet Distribution

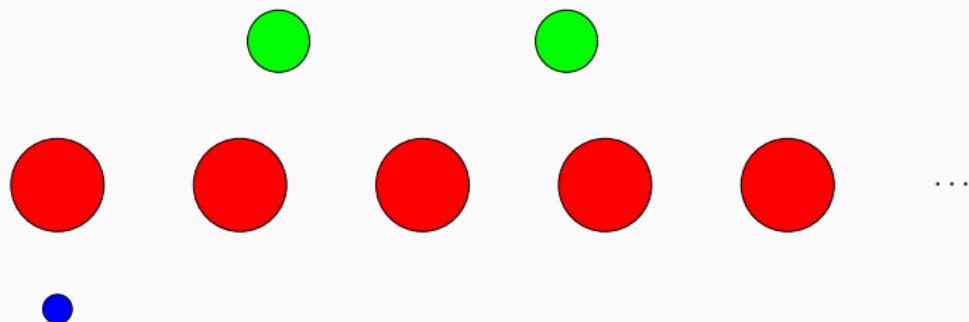


$\text{Dir}(7, 5, 3, 2)$

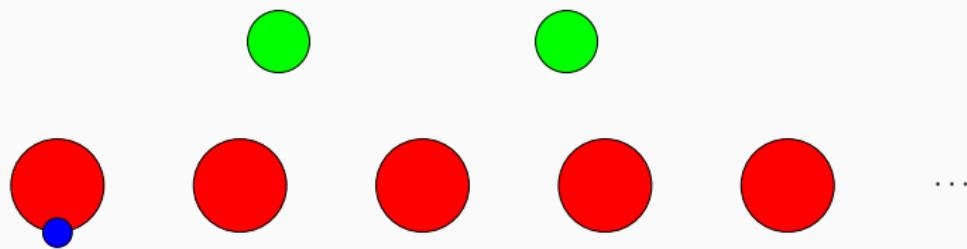
Dirichlet Process

- Is the infinite dimensional generalisation of a Dirichlet distribution
 - just as Gaussian process is of Gaussian distribution
- Generates a partitioning of (possibly) infinite number of elements
- Not as intuitive to write down
- Best written constructively

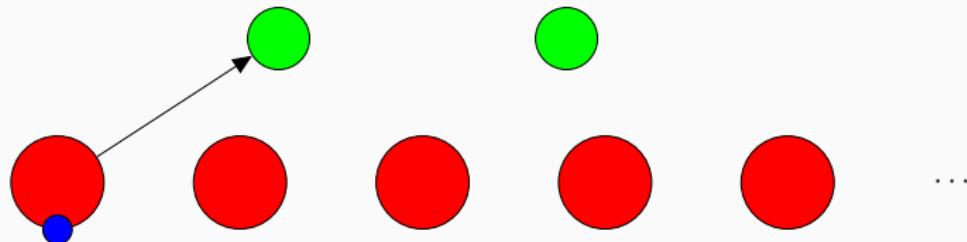
Chinese Restaurant Process



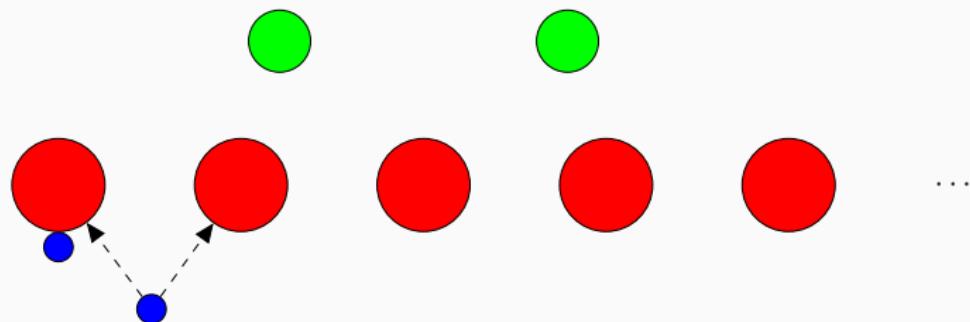
Chinese Restaurant Process



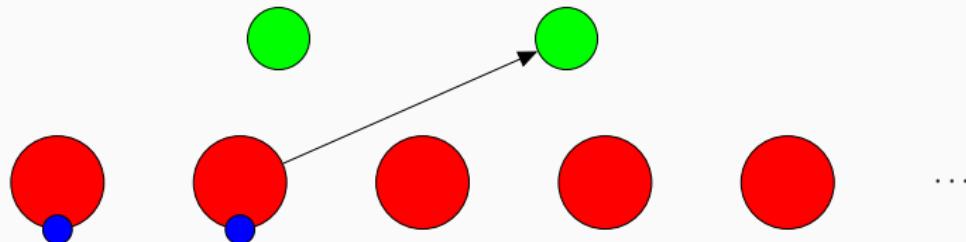
Chinese Restaurant Process



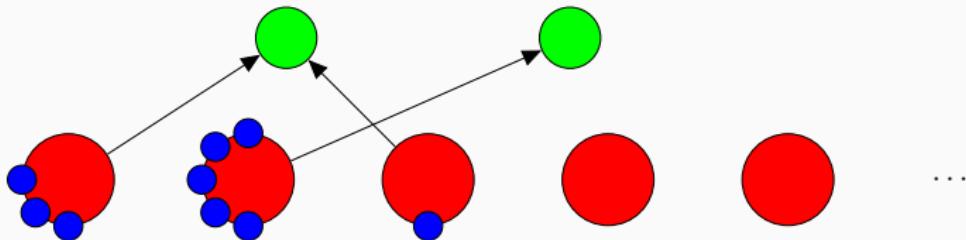
Chinese Restaurant Process



Chinese Restaurant Process

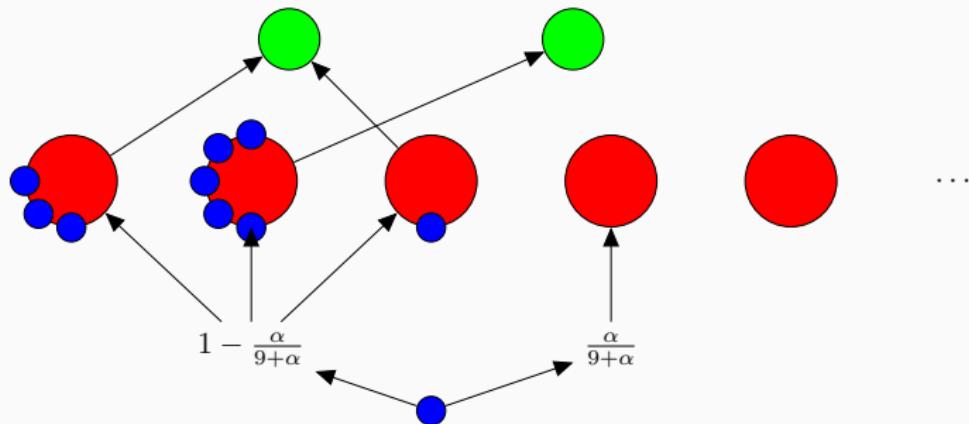


Chinese Restaurant Process

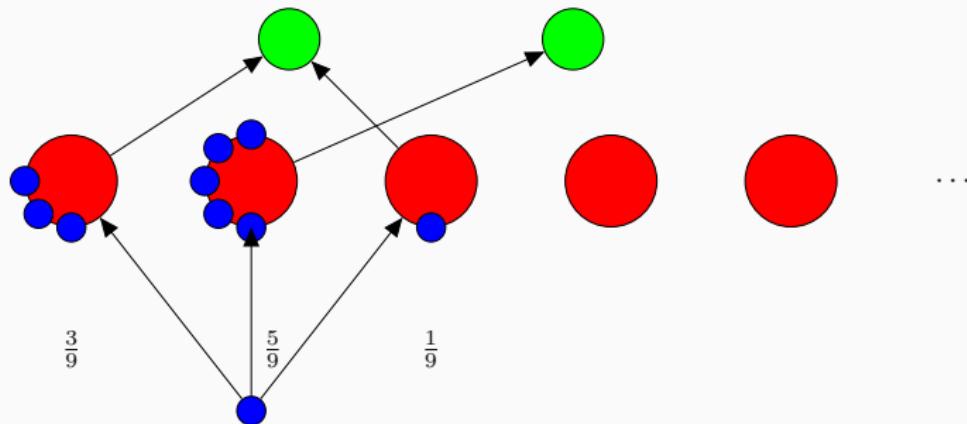


- Go to new table $\frac{\alpha}{N-1+\alpha}$
- If not choose table as $\frac{n_i}{N}$ where n_i number of diners at table i

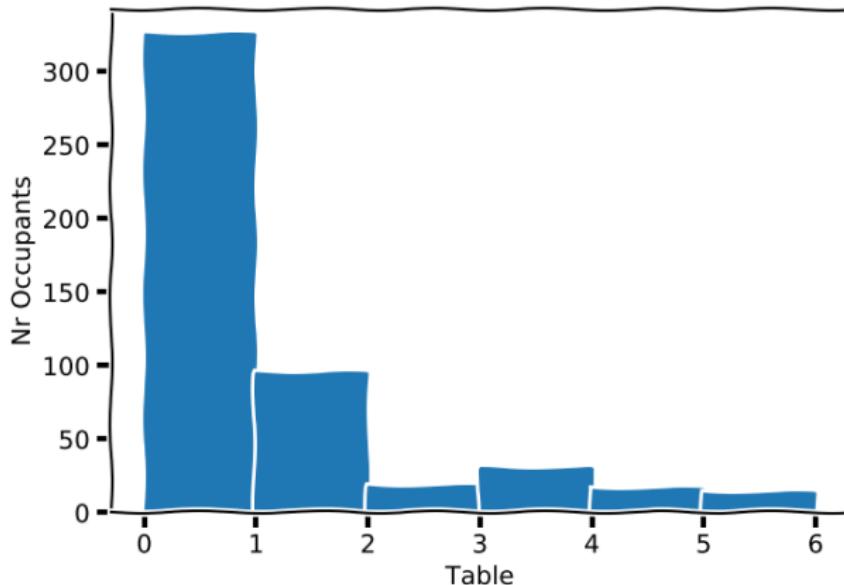
Chinese Restaurant Process



Chinese Restaurant Process

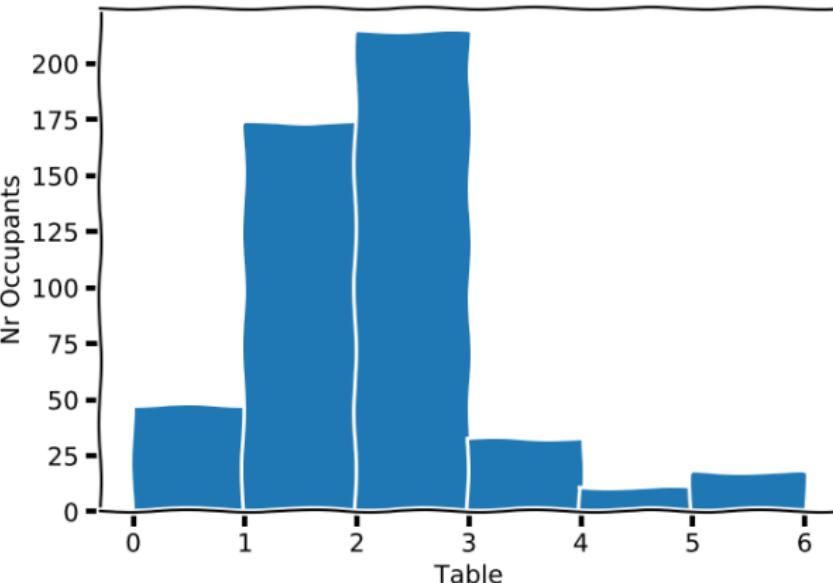


Chinese Restaurant Process



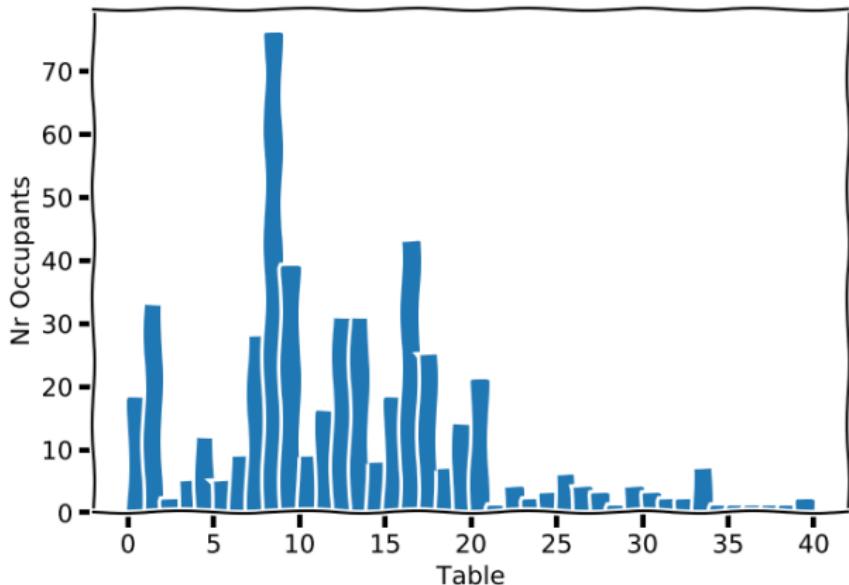
$$N = 500 \quad \alpha = 1.0$$

Chinese Restaurant Process



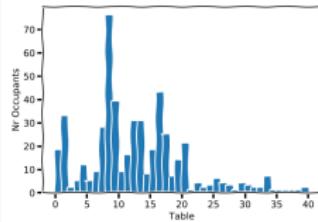
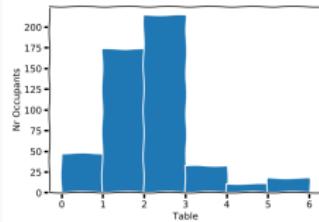
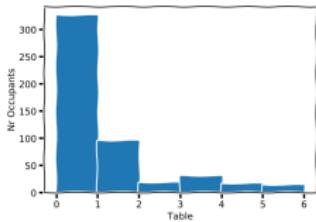
$$N = 500 \quad \alpha = 2.0$$

Chinese Restaurant Process



$$N = 500 \quad \alpha = 10.0$$

Generative Model: Chinese Restaurant Process



- Given a variable α we can **generate** data
- α specifies a distribution over "table" assignments
- Think of this as,

$$p(\mathbf{Y}|\alpha)$$

- We can make α random as well

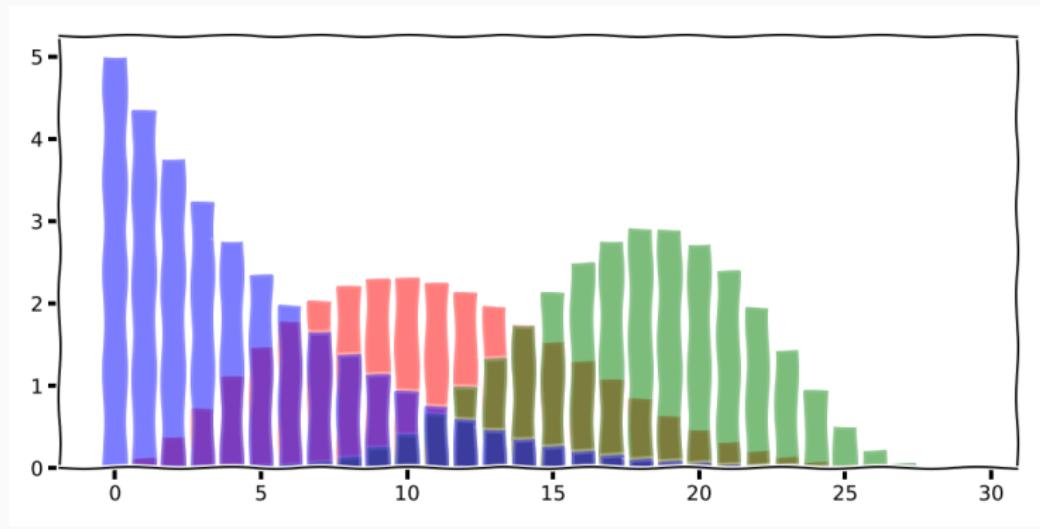
$$\alpha \sim \text{Lognormal}(\mu, \sigma^2)$$

Latent Dirichlet Allocation

Text Data

Etiam vel semper nullam, etiam vel semper nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Bag-of-Words



Assumptions

- Documents contains a blend of topics
- Each topic has a characteristic distribution of words
- The order of the words is irrelevant

Topic Distribution

Topic	$p(\text{Topic})$
Computer Science	0.3
Pugs	0.5
Bristol City	0.1
C64	0.1

$$p(\text{Topic}) = p(\beta)$$

Word Topic Distribution

Topic	8-bit	Walsh	AI	Hugs
Computer Science	0.6	0.09	0.3	0.01
Pugs	0.1	0.1	0.3	0.5
Bristol City	0.1	0.8	0.05	0.05
C64	0.8	0.01	0.15	0.04

$$p(\text{Word}|\text{Topic}) = p(w_{d,n}|\beta_k)$$

Document Topic Distribution

Document	Computer Science	Pugs	Bristol City	C64
One Team in Bristol	0.1	0.2	0.69	0.01
Coursework report	0.8	0.1	0.07	0.03
Dr. Doobs	0.6	0.1	0.1	0.2
Pug Weekly	0.1	0.8	0.05	0.05

$$p(\theta_d)$$

Generative Process

1. Choose a random distribution over topics
2. Randomly choose a topic from the topic distribution
3. Randomly choose a word from the word distribution of this topic

We have specified a *generative model* for text data

Topic Model

$w_{d,n}$ n :th word in d :th document

Topic Model

$w_{d,n}$ n :th word in d :th document

β_k topic-word distribution for k :th topic

Topic Model

$w_{d,n}$ n :th word in d :th document

β_k topic-word distribution for k :th topic

θ_d topic distribution for d :th document

Topic Model

$w_{d,n}$ n :th word in d :th document

β_k topic-word distribution for k :th topic

θ_d topic distribution for d :th document

$z_{d,n}$ topic assignment for n :th word in d :th document

Topic Model

$w_{d,n}$ n :th word in d :th document

β_k topic-word distribution for k :th topic

θ_d topic distribution for d :th document

$z_{d,n}$ topic assignment for n :th word in d :th document

Now lets write down the joint distribution of these according to our generative model

Topic Model

- To generate the n :th word in the d :th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n} | \beta, z_{d,n})$$

Topic Model

- To generate the n :th word in the d :th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

- To generate a topic assignment $z_{d,n}$ I need to know the topic distribution for document d

$$p(z_{d,n}|\theta_d)$$

Topic Model

- To generate the n :th word in the d :th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

- To generate a topic assignment $z_{d,n}$ I need to know the topic distribution for document d

$$p(z_{d,n}|\theta_d)$$

- If we know the topic assignment and the topic-word distribution we can assume the words to be independent

$$p(w_d|\beta, z_d) = \prod_{n=1}^N p(w_{d,n}|\beta, z_{d,n})$$

Topic Model

- We can assume that the topic-document distribution is not dependent between documents

$$p(\theta) = \prod_{d=1}^D p(\theta_d)$$

Topic Model

- We can assume that the topic-document distribution is not dependent between documents

$$p(\theta) = \prod_{d=1}^D p(\theta_d)$$

- We can assume that topic word distribution is independent

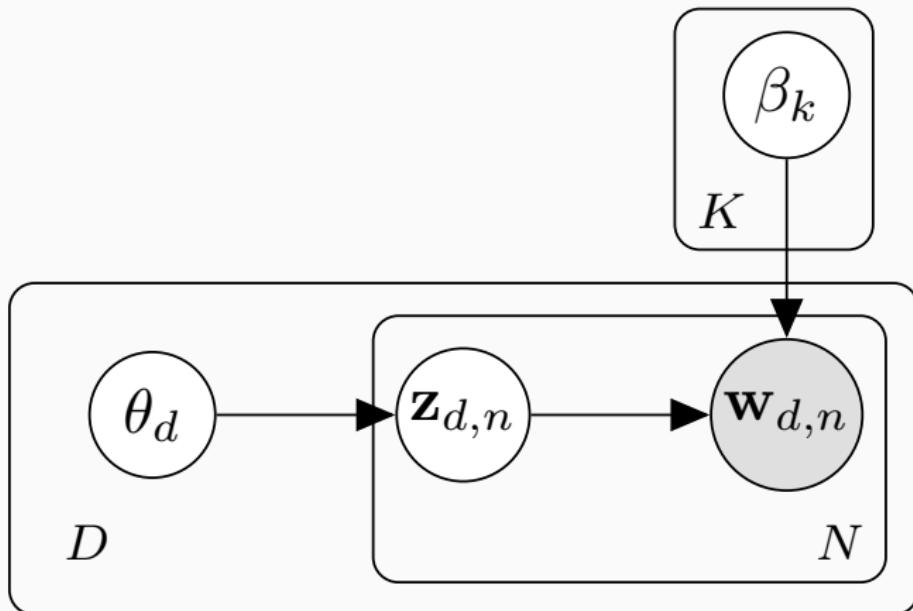
$$p(\beta) = \prod_{k=1}^K p(\beta_k)$$

Topic Model

$$p(w, z, \theta, \beta) = \prod_{k=1}^K p(\beta_k) \prod_{d=1}^D p(\theta_d) \underbrace{\prod_{n=1}^N p(w_{d,n} | \beta, z_{d,n}) p(z_{d,n} | \theta_d)}_{\text{word}} \\ \underbrace{\quad}_{\text{document}} \\ \underbrace{\quad}_{\text{corpus}}$$

- This is the joint distribution of a specific topic model
- The assumptions we have made are called a Latent Dirichlet Allocation [1]
- This specifies a specific probability distribution by our assumptions

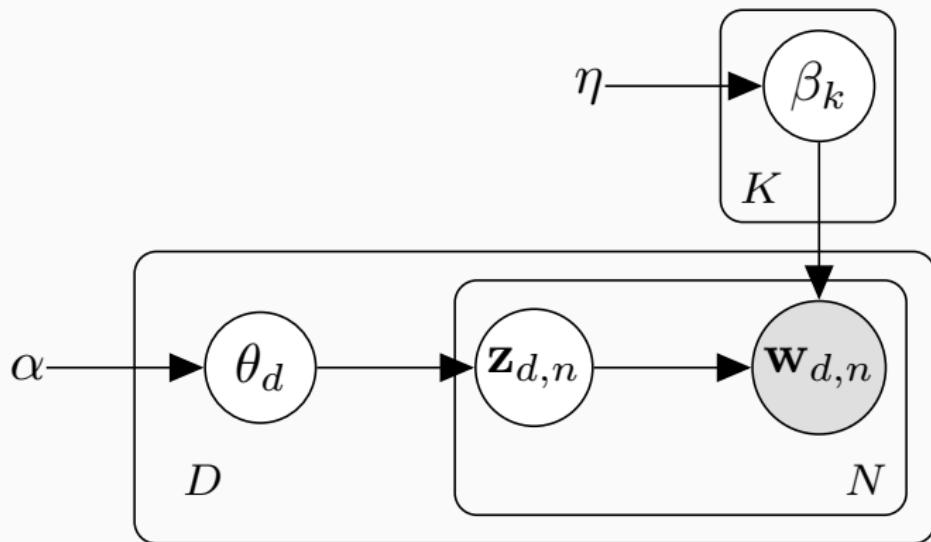
Graphical Model



Topic Model: Distributions

- We still have to choose the specific form of the distribution
- Topic-word: $\beta_k \sim \text{Dir}(\eta)$
- Topic-proportions: $\theta_d \sim \text{Dir}(\alpha)$
- Topic assignment: $z_{d,n} \sim \text{Mult}(\theta_d)$
- Word assignment: $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$

Topic Model: Graphical Model



Topic Model: Learning

$$p(z, \theta, \beta | w) = \frac{p(w, z, \theta, \beta)}{p(w)}$$

- We only observe the words
- Want to infer all the distributions

Topic Model: Evidence

$$\begin{aligned} p(w) &= \int p(w, z, \theta, \beta) d\beta d\theta dz \\ &= \int \prod_{k=1}^K p(\beta_k) \prod_{d=1}^D p(\theta_d) \prod_{n=1}^N p(w_{d,n} | \beta, z_{d,n}) p(z_{d,n} | \theta_d) d\beta d\theta dz \end{aligned}$$

- Distribution over corpus for *any topic model* weighted by our assumptions

Topic Model: Learning

$$p(z, \theta, \beta | w) = \frac{p(w, z, \theta, \beta)}{p(w)}$$

- Completely intractable
- Efficient approximations (Part III)
- Possible to learn from huge amounts of data

Experiments

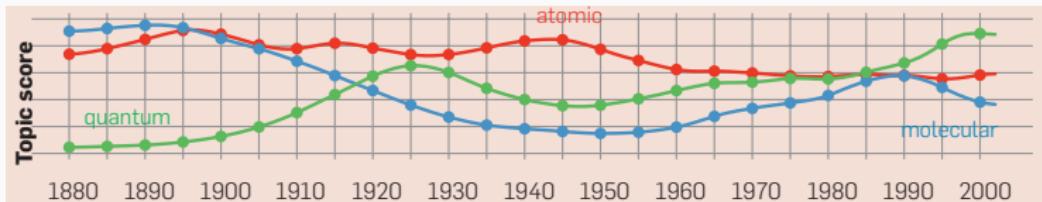
Science [2]

“Genetics”	“Evolution”	“Disease”	“Computers”
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

4	10
tax	labor
income	workers
taxation	employees
taxes	union
revenue	employer
estate	employers
subsidies	employment
exemption	work
organizations	employee
year	job
treasury	bargaining
consumption	unions
taxpayers	worker
earnings	collective
funds	industrial

3	13
women	contract
sexual	liability
men	parties
sex	contracts
child	party
family	creditors
children	agreement
gender	breach
woman	contractual
marriage	terms
discrimination	bargaining
male	contracting
social	debt
female	exchange
parents	limited

Science [2]



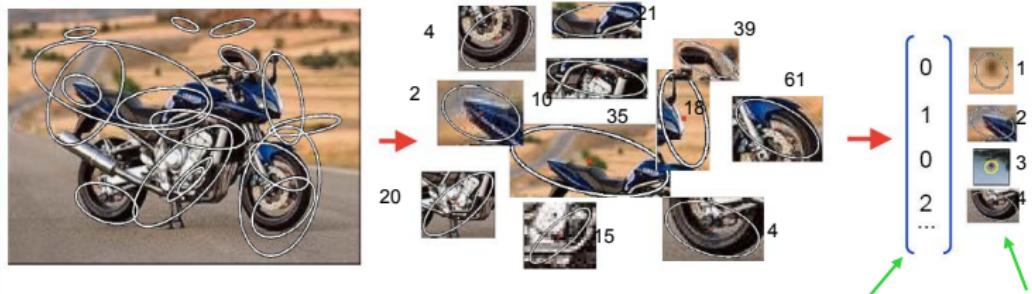
Beyond words



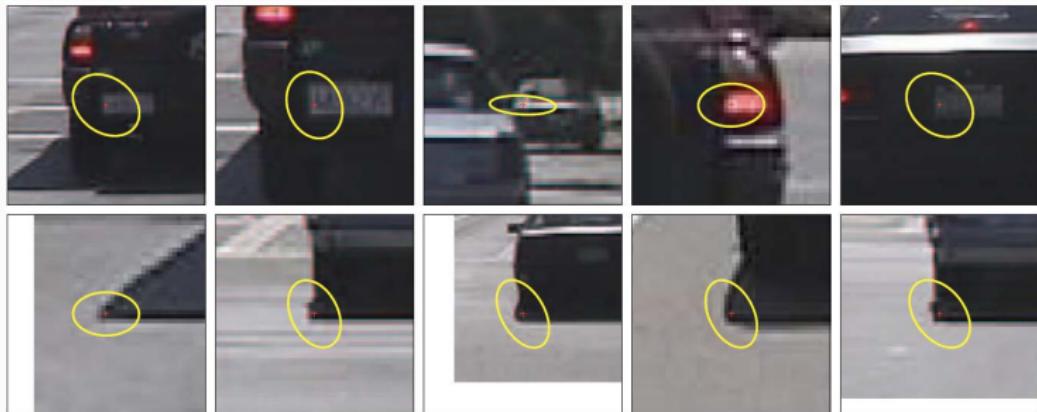
Computer Vision



Computer Vision



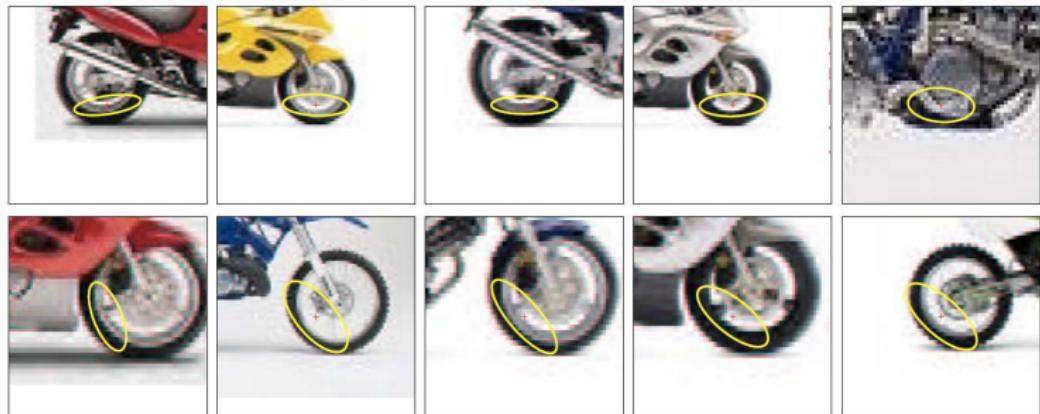
Computer Vision

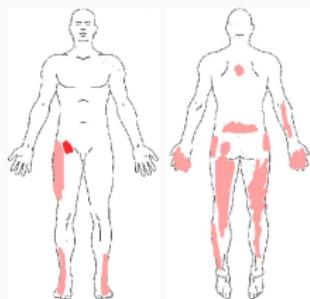


Computer Vision



Computer Vision



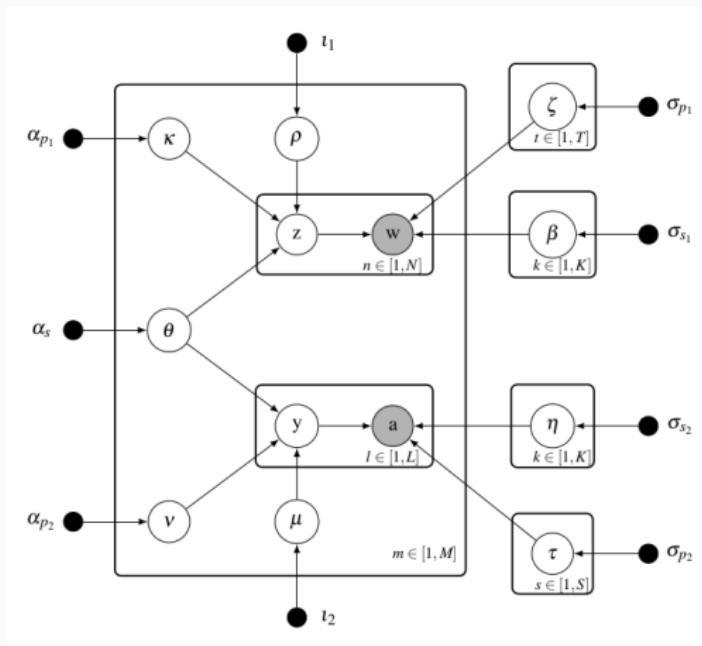


Symptom diagnoses: Interscapular discomfort; R arm discomfort; B hands discomfort; Lumbago; B crest of the ilium discomfort; L side thigh discomfort; B back thigh discomfort; B calf discomfort; B achilles tendinitis; B shin discomfort; R inguinal discomfort;

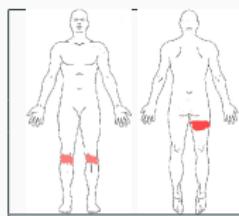
Pattern diagnoses B L5 Radiculopathy; B S1 Radiculopathy; B C7 Radiculopathy;

Pathophysiological diagnoses DLI L4-L5; DLI S1-S2; DLI C6-C7

Health [3]



Health [3]



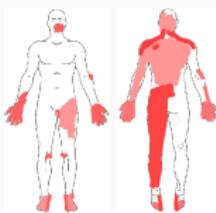
6 Prd: R back thigh dcf; L PFS (Patellofemoral pain syndrome); R PFS;

L L5 Rdc; R L5 Rdc; DLI L4-L5;

6 GT: R back thigh dcf; L PFS; R PFS;

L L5 Rdc; R L5 Rdc; DLI L4-L5;

Health [3]



50 Prd: Headache; L neck dcf; R neck dcf; Neck def; L upper trapezius dcf; R upper trapezius dcf; L shoulder def; L hand def; R hand def; Interscapular dcf; Lumbago; Lateral abdominal def; L groin dcf; L side thigh dcf; R side thigh dcf; L calf def; L back thigh dcf; L crest of the ilium dcf; R crest of the ilium def; R foot arch def; L toe joint dcf; R toe joint dcf; L medial elbow def; L ankle def; R ankle def; L foot arch def; L PFS; L dorsal knee dcf; R medial knee def;

L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; R C6 Rdc; L C6 Rdc; L C7 Rdc; R C7 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rde; R S1 Rde; DLI C2-C3; DLI C3-C4; DLI C5-C6; DLI C6-C7; DLI L4-L5; DLI L5-S1; OB;

50 GT: L back headache; R back headache; Neck def; L jaw def; L upper trapezius def; R upper trapezius def; L arm def; R arm def; L lateral elbow dcf; R lateral elbow def; L hand joint def; R hand joint def; L hand def; R hand def; L thumb def; R thumb def; L finger def; R finger def; Lumbago; L groin def; L back thigh def; L calf def; L medial knee def; L ankle def; R ankle def; R medial knee def; R big toe def; L big toe def;

L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; L C5 Rdc; R C5 Rdc; L C6 Rdc; R C6 Rdc; L C7 Rdc; R C7 Rdc; L L4 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rde; R S1 Rde; Craniocervical joint injury; DLI C4-C5; DLI L3-L4; DLI L4-L5; DLI L5-S1;

Summary

Summary

- How to build up a more complicated model from scratch
- Thinking about conditional distributions as generative models
- Topic models

Next time

- We are done with Part II
- This was the **big** part of the unit
- How to build models
- Monday: Summary on Part II
- Final part: how to perform inference

eof

References

 David M Blei, Andrew Y Ng, and Michael I Jordan.

Latent dirichlet allocation.

3:993–1022, March 2003.

 David M. Blei.

Probabilistic topic models.

Commun. ACM, 55(4):77–84, April 2012.

 Cheng Zhang, Hedvig Kjellström, Carl Henrik Ek, and Bo C. Berglindson.

Diagnostic prediction using discomfort drawings with IBTM.

In *Proceedings of the 1st Machine Learning in Health Care,
MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016,*
pages 226–238, 2016.