The Dirichlet Categorial Model

Christoph Teichmann, Antoine Venant

Goals

Motivation

A Simple Approach

The Dirichlet Distribution

The Dirichlet Process

Chinese Restaurant Process

Summary

The Dirichlet Categorial Model

Christoph Teichmann   Antoine Venant

Goals

Teaching Goals

- Basic properties of Dirichlet Distribution
- Dirichlet Categorial Model
- Chinese Restaurant Process
- How to build a simple language model

Goals

Motivation

A Simple Approach

The Dirichlet Distribution

The Dirichlet Process

Chinese Restaurant Process

Summary

Motivation

Categorial Variables in Computationl Linguistics

We want probabilites $P(\text{outcome}_1)$, $P(\text{outcome}_2)$, ... for random variables with discrete outcomes:

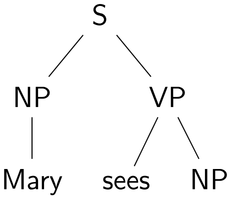
- Next word given previous ones
- Children given parent in binary constituent tree
- Next part-of-speech tag given previous ones

Mary sees { the, something, John, . } ...

Categorial Variables in Computationl Linguistics

We want probabilites $P(\text{outcome}_1)$, $P(\text{outcome}_2)$, $\ldots$ for random variables with discrete outcomes:

- Next word given previous ones
- Children given parent in binary constituent tree
- Next part-of-speech tag given previous ones



$NP \rightarrow something$

$NP \rightarrow John$

Goals

Motivation

A Simple Approach

The Dirichlet Distribution

The Dirichlet Process

Chinese Restaurant Process

Summary

A Simple Approach

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
- Reduce your problem to estimating a random quantity of interest
- Build joint probabilistic model of the data and the quantity to estimate

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
  - sentences from websites (tokenized) ✓
- Reduce your problem to estimating a random quantity of interest
- Build joint probabilistic model of the data and the quantity to estimate

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
  - sentences from websites (tokenized) ✓
- Reduce your problem to estimating a random quantity of interest
  - Probabilities for next word in context ✓
  - We want a probability $P(P(W_{\text{next}} = Mary) = 0.4)$, $P(P(W_{\text{next}} = \text{Mary}) = 0.5)$, $P(P(W_{\text{next}} = \text{Mary}) = 0.6)$ ...
- Build joint probabilistic model of the data and the quantity to estimate

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
  - sentences from websites (tokenized) ✓
- Reduce your problem to estimating a random quantity of interest
  - Probabilities for next word in context ✓
  - We want a probability $P(P(W_{\text{next}} = \mathit{Mary}) = 0.4)$, $P(P(W_{\text{next}} = \text{Mary}) = 0.5)$, $P(P(W_{\text{next}} = \text{Mary}) = 0.6)$ ...
- Build joint probabilistic model of the data and the quantity to estimate
  - uh oh

Bayes Theorem

v - value of interesting variable, d - value of data, we are interesting in $P(v|d)$

$$P(v,d) = P(v|d)P(d)$$
$$P(v,d) = P(d|v)P(v)$$
$$P(v|d) = \frac{P(d|v)P(v)}{P(d)}$$
$$P(v|d) = \frac{\overbrace{P(d|v)}^{likelihood}\ \overbrace{P(v)}^{prior}}{\underbrace{\sum_{v'} P(v',d)}_{}\!\!\!\!\!\!\!\!\overbrace{\phantom{\sum_{v'} P(v',d)}}^{normalizer}}$$

What Do We Need

- $P(\text{text}|\text{probabilities})$
- $P(\text{probabilities})$

Simple Model $P(text|probabilities)$

- Text generated from n-Gram model
- $P(\overbrace{W_i}^{\text{random variable for i-th word}} | \overbrace{w_{i-1}}^{\text{concrete value for word}}, \ldots, w_0) = P(W_i | w_{i-1}, \ldots, w_{i-(n-1)})$
- Let us start with $n = 1$, i.e., Unigram Model: $P(W_i | w_{i-1}, \ldots, w_0) = P(W)$
- Text draw word by word from $P(W)$

We have model for data given word probabilities – need a model to give us probabilities for $P(W = \text{Mary}) = 0.01$ or $P(W = \text{Mary}) = 0.5$?

Simplest Choice – Set-UP

- Assume finite lexicon L – possible outcomes
- Random variable over distributions P^X, P^x single outcome
  - P^x could be $P^x(\text{Mary}) = 0.1$, $P^x(\text{sees}) = 0.3 \ldots$
- Probability word is "Mary" $P^x(W = \text{Mary})$
- **Full joint model** $\underline{\mathbf{P}\left(\mathbf{P^X}, \mathbf{W_1}, \ldots, \mathbf{W_n}\right)}$

Simplest Choice – Everything Equally Likely

- Let us start with an uniform model, i.e. $P\left(P^X = P^x\right) = \frac{1}{Z}$
- Z is a constant to ensure integration to 1
- We think $P^x(\text{the}) = 0.9$ is as likely as $P^x(\text{the}) = 0.2$ or $P^x(\text{the}) = 0.03$
- We think $P^x(\text{Mary}) = 0.9$ and $P^x(\text{the}) = 0.1$ is as likely as $P^x(\text{Mary}) = 0.2$ and $P^x(\text{the}) = 0.8$

Similar to model from last session

How Does This Work – Example

- Lexicon: $L = \{\text{Mary}, \text{sees}, ., \text{something}, \text{John}\}$
- Dataset: $D =$"Mary sees something"

Relevant Questions:

How Does This Work – Example

- Lexicon: $L = \{\text{Mary}, \text{sees}, ., \text{something}, \text{John}\}$
- Dataset: $D =$"Mary sees something"

Relevant Questions:

- What is the posterior, i.e., what is $P(P^x|\text{something}, \text{sees}, \text{Mary})$ for a given P^x?

How Does This Work – Example

- Lexicon: $L = \{\text{Mary}, \text{sees}, ., \text{something}, \text{John}\}$
- Dataset: $D =$"Mary sees something"

Relevant Questions:

- What is the posterior, i.e., what is $P(P^x|\text{something}, \text{sees}, \text{Mary})$ for a given P^x?
- What is $P(W_{next}|\text{something}, \text{sees}, \text{Mary})$? - predictive distribution

Posterior

The posterior is proportional to this (why?):

$$\overbrace{P(P^X = P^x | w_3, w_2, w_1)}^{P(\text{Posterior})} \propto \left(\overbrace{\prod_{i \in \{1,2,3\}} P^x(w_i)}^{P(\text{text}|\text{probs})} \right) \overbrace{\frac{1}{Z}}^{P(\text{probs})}$$

Posterior

The posterior is proportional to this (why?):

$$\overbrace{P(P^X = P^x | w_3, w_2, w_1)}^{P(\text{Posterior})} \propto \left(\overbrace{\prod_{i \in \{1,2,3\}} P^x(w_i)}^{P(\text{text}|\text{probs})} \right) \overbrace{\frac{1}{Z}}^{P(\text{probs})}$$

- What is the normalizer?
- What is the mean?
- Where is the maximum of this posterior?

Posterior

The posterior is proportional to this (why?):

$$\overbrace{P(P^X = P^x|w_3, w_2, w_1)}^{P(\text{Posterior})} \propto \left(\overbrace{\prod_{i \in \{1,2,3\}} P^x(w_i)}^{P(\text{text}|\text{probs})} \right) \overbrace{\frac{1}{Z}}^{P(\text{probs})}$$

- What is the normalizer?
- What is the mean?
- Where is the maximum of this posterior?
- Maybe map this to known distribution?

The Dirichlet Distribution

Is This A Known Distribution?

Dirichlet Distribution – distribution over probability distributions P^x with different outcomes $w_1 \in L$:

$$P_{Dir}(P^x) = \frac{\prod_{w \in L} P^x(w)^{\alpha_i - 1}}{B(\boldsymbol{\alpha})}$$

- Here $\alpha_i > 0$ are parameters – there are lots of Dirichlet Distributions
- $B(\boldsymbol{\alpha})$ is a normalizer depending on all the a_i – look it up on Wikipedia

Is This A Known Distribution?

Dirichlet Distribution – distribution over probability distributions P^x with different outcomes $w_1 \in L$:

$$P_{Dir}(P^x) = \frac{\prod_{w \in L} P^x(w)^{\alpha_i - 1}}{B(\boldsymbol{\alpha})}$$

- Mean a probability distribution with $P^x(w) = \frac{\alpha}{\sum_{w' \in L} \alpha_{w'}}$
- Maximum = mean if αs all greater 1

It is a Dirichlet Distribution!

Let us work through this on the table:

$$P_{Dir}(P^x) = \frac{\prod_{w \in L} P^x(w)^{\alpha_i - 1}}{B(\boldsymbol{\alpha})}$$

$$P(P^x | w_3, w_2, w_1) \propto \left(\prod_{i \in \{1,2,3\}} P^x(w_i) \right) \frac{1}{Z}$$

$$\propto \frac{\prod_{w \in L} Px(w)^{\sum_{i \in [1,3]} 1(w_i = w)}}{Z}$$

It is a Dirichlet Distribution!

Let us work through this on the table:

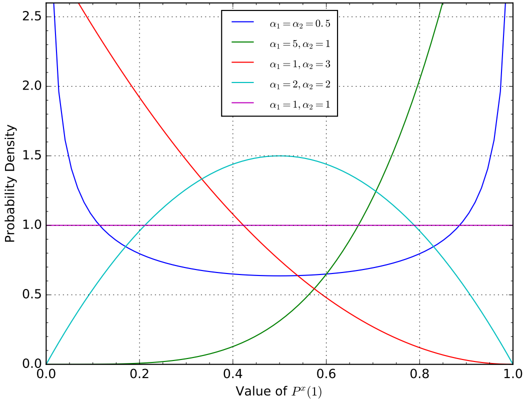
$$P_{Dir}(P^x) = \frac{\prod_{w \in L} P^x(w)^{\alpha_i - 1}}{B(\boldsymbol{\alpha})}$$

$$P(P^x | w_3, w_2, w_1) = \frac{\prod_{w \in L} P_X(w)^{\sum_{i \in [1,3]} 1(w_i = w)}}{B(\text{sums} + 1)}$$

αs correspond to frequency of each word $+1$

Visualizing a Dirichlet Distribution With Two Outcomes

Say we have $P^x(1)$ and $P^x(2)$

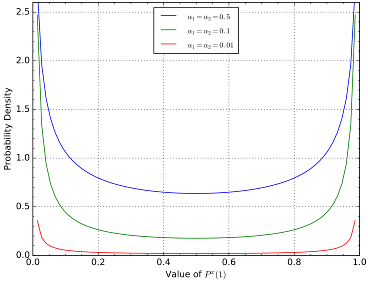


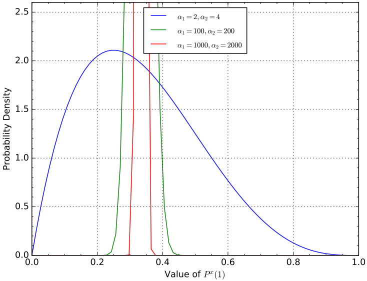
Generated with code based on code found at:

`commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg` **by user Horas**

Understanding the Behaviour of the Dirichlet

- Increase αs $\rightarrow$ concentrates at values proportional to αs
- If all α are less than 1, then we favour "sparse distributions" few outcomes likely





Back to the Example

For our example our posterior is (table):

$$P(P^x) = \frac{\left(\prod_{w \in L} P^x(w)^{\alpha_{w_i}-1}\right)}{B(\boldsymbol{\alpha})}$$

Where $\alpha_{\text{Mary}} = \alpha_{\text{sees}} = \alpha_{\text{something}} = 2$ and $\alpha_{\text{John}} = \alpha_{.} = 1$

Next Step

- Have posterior probability for the different P^x
- Know it is a Dirichlet Distribution
- What are the probabilities for next word – what is predictive distribution?

For a given P^x:

$$P(w_{next}, P^x | w_1, \ldots, w_n) = \overbrace{P^x(w_{next})}^{P(\text{text}|\text{probs})} \overbrace{P(P^x | w_1, \ldots, w_n)}^{\text{posterior for } P^x}$$

But we need to sum (actually integrate) over all possible P^x

Predictive Distribution

$$P(w_{next}) = \int_{P^x} \overbrace{P^x(w_{next})}^{Categorial} \overbrace{\frac{\left(\prod_{w \in L} P^x(w)^{\alpha_w - 1}\right)}{B(\boldsymbol{\alpha})}}^{Dirichlet} dP^x$$

Combines a Categorial and Dirichlet Distribution – called a Dirichlet Categorial Model, well known

Predictive Distribution

$$P(w_{next}) = \int_{P^x} \overbrace{P^x(w_{next})}^{Categorial} \overbrace{\frac{\left(\prod_{w \in L} P^x(w)^{\alpha_w - 1}\right)}{B(\boldsymbol{\alpha})}}^{Dirichlet} dP^x$$

Let us just look it up, e.g., the Dirichlet write up at `https://people.eecs.berkeley.edu/~stephentu/`

Predictive Distribution

$$P(w_{\text{next}}) = \int_{P^x} P^x(w_{\text{next}}) \frac{\left(\prod_{w \in L} P^x(w)^{\alpha_w - 1}\right)}{B(\boldsymbol{\alpha})} dP^x$$

$$P(w_{\text{next}}) = \frac{\alpha_{w_{\text{next}}}}{\sum_{w \in L} \alpha_w}$$

- Probability of outcome is α for that outcome divided by sum of αs for all outcomes (finish example on table)

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
  - sentences from websites (tokenized) ✓
- Reduce your problem to estimating a random quantity of interest
  - Probabilities for next word in context ✓
- Build joint probabilistic model of the data and the quantity to estimate
  - Unigram model with uniform prior ✓

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
  - sentences from websites (tokenized) ✓
- Reduce your problem to estimating a random quantity of interest
  - Probabilities for next word in context ✓
- Build joint probabilistic model of the data and the quantity to estimate
  - Unigram model with uniform prior ✓
- Get a posterior distribution
  - We get a Dirichlet Distribution posterior, which is well understood ✓
  - We can predict the next word using knowledge about Dirichlet Categorial Model ✓

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
  - sentences from websites (tokenized) ✓
- Reduce your problem to estimating a random quantity of interest
  - Probabilities for next word in context ✓
- Build joint probabilistic model of the data and the quantity to estimate
  - Unigram model with uniform prior ✓
- Get a posterior distribution
  - We get a Dirichlet Distribution posterior, which is well understood ✓
  - We can predict the next word using knowledge about Dirichlet Categorial Model ✓
- Check model quality and revise if necessary

Is Our Model Reasonable?

Predictive probabilities for any word position:

$$P(W_{\text{next}} = .) = \frac{1}{8}$$
$$P(W_{\text{next}} = \text{Mary}) = \frac{2}{8}$$

Data may support this, but it seems unlikely given what we know about English

A Better Model

- We assumed that we have unigram word probabilities – switch to e.g. a bigram model
- Instead of assuming uniform probability for different P^x we could use Dirichlet as prior with different αs

New Model

- We will have a Dirichlet Prior with αs proportional to observed word frequencies in some corpus
- We now have to find probabilities $P^x(W_i = w_i|w_{i-1})$

New Model

- We will have a Dirichlet Prior with αs proportional to observed word frequencies in some corpus
- We now have to find probabilities $P^x(W_i = w_i | w_{i-1})$
- Let us work through that for the data:
  - "Mary saw something ."
  - "John saw"

Conjugate Model

What is the general posterior with the Dirichlet model after words $w_1, \dots, w_n$?

$$P(P^x) = \frac{\prod_{w \in L} P^x(w)^{\alpha_w - 1 + \sum_{i \in [1,n]} 1(w_i = w)}}{Z}$$

- Another Dirichlet Distribution
- Prior and piosterior same type $\rightarrow$ conjugate model
- Dirichlet Distribution prior is conjugate for Categorial Distribution likelihood
- Understanding prior – understanding posterior

The Next Word – The Bayesian Approach

- Find a dataset relevant to your problem
  - sentences from websites (tokenized) ✓
- Reduce your problem to estimating a random quantity of interest
  - Probabilities for next word in context ✓
- Build joint probabilistic model of the data and the quantity to estimate
  - unigram model with uniform prior ✓
- Get a posterior distribution
  - We get a Dirichlet Distribution posterior, which is well understood ✓
  - We can predict the next word using knowledge about Dirichlet Categorial model ✓
- Check model quality and revise if necessary
  - Larger n-grams with better prior ✓

Goals

Motivation

A Simple Approach

The Dirichlet Distribution

The Dirichlet Process

Chinese Restaurant Process

Summary

The Dirichlet Process

Further Model Problems

"Mary saw a cake ."

- More than 5 words in language
- Maybe even infinitely many words

Moving to the Dirichlet Process

- $P(X)$ single outcome probability distribution
- $P(X_1, \ldots, X_n) \rightarrow$ multiple outcome probability distribution
- $P(X_1, X_2, X_3 \ldots) \rightarrow$ stochastic process / random function

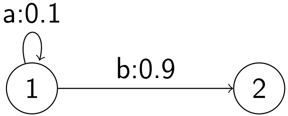
Extension to Infinite Number of Variables

We need a prior for P^x when L is infinite $\rightarrow$ infinitely many probabilities $\rightarrow$ need a stochastic process

- Use what we know about Dirichlet Distribution
- How can we define an infinite list of αs?
- How to define a process?

How Can We Define an Infinite List of αs?

- When you have a hammer ...
- Define a single concentration parameter α and a base probability distribution (measure) $P(X)$
- $\alpha_x = \alpha P(X = x)$
- different αs always sum to concentration parameter



If $\alpha = 10$ what is α_{ab}?

Extension to Infinite Number of Variables

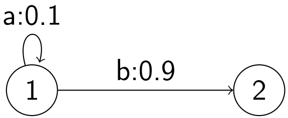
- How can we define an infinite list of αs?
  - Concentration parameter + probability distribution ✓
- How to define a process?

How to Define a Process

- No need to define probability for a full outcome $P(X_1 = x_1, X_2 = x_2, \dots)$
- If we pick a few values for a few variables and do not care about all the others – what is the probability?
- Probability Distributions for any $X^1, X^2, \dots, X^m$ follow a Dirichlet

Making it Concrete

The Dirichlet Categorial Model
Christoph Teichmann, Antoine Venant
Goals
Motivation
A Simple Approach
The Dirichlet Distribution
The Dirichlet Process
Chinese Restaurant Process
Summary



$$P_{Dir}(P_X) = \frac{\prod_{w \in L} P_X(w)^{\alpha_w}}{B(\boldsymbol{\alpha})}$$

$$\alpha = 10$$

$$P(P^x(b) = 0.2, P^x(ab) = 0.5) = \frac{\overbrace{0.2^9}^{b} \times \overbrace{0.5^{0.9}}^{ab} \times \overbrace{0.3^{0.1}}^{\text{everything else}}}{B(\langle 9, 0.9, 0.1 \rangle)}$$

Extension to Infinite Number of Variables

- How can we define an infinite list of αs?
  - Concentration parameter + probability distribution ✓
- How to define a process?
  - Define probability distribution for any finite selection of variables ✓

Applying for Next Word Prediction

Christoph Teichmann, Antoine Venant

Goals
Motivation
A Simple Approach
The Dirichlet Distribution
The Dirichlet Process
Chinese Restaurant Process
Summary

- Bigram Model
- Infinitely many possible words w_i
- Probabilities for each $P(W_i|w_{i-1})$ come from Dirichlet Process
- Pick some concentration parameter α
- use a Markov Chain over morphems to define base probabilities $P(X)$

What is the predictive density / what is the posterior?

Predictive Density

Still a Dirichlet Categorical Distribution!

$$P(W_0 = ab) = \frac{\alpha_{ab}}{\alpha}$$

Posterior

Still a Dirichlet Process! – table

Seen “ab” as first word - now $\alpha'_{ab} = \alpha_{ab} + 1$, $\alpha' = \alpha + 1$

Goals

Motivation

A Simple Approach

The Dirichlet Distribution

The Dirichlet Process

Chinese Restaurant Process

Summary

Chinese Restaurant Process

Can We Make This Simpler?

- Chinese Restaurant Process – metaphor for Dirichlet Categorial Model
- Model just for words
- Every word is a table
- Number of people at table x is α_x
- People are likely to sit at popular tables

Presentation in Papers

- Usual presentation: tables created when first costumer sits at them
- New table created with $\frac{\alpha}{\alpha + \text{seen}}$
- New table label – according to base probability distribution
- Sitting at old table $\frac{\text{costumers old table}}{\alpha + \text{seen}}$

Goals

Motivation

A Simple Approach

The Dirichlet Distribution

The Dirichlet Process

Chinese Restaurant Process

Summary

Summary

Take Away

- Important for NLP: categorial probability distributions
- Handy prior: Dirichlet Distribution/Process
- Works for infinite choices as well
- Posterior is also Dirichlet Distribution/Process
- Predictive distribution easy to read off α
- Chinese Restaurant metaphor used frequently

Questions!