Lecture 7: The Dirichlet Distribution and Text

4F13: Machine Learning

Joaquin Quiñonero-Candela and Carl Edward Rasmussen

Department of Engineering
University of Cambridge

http://mlg.eng.cam.ac.uk/teaching/4f13/

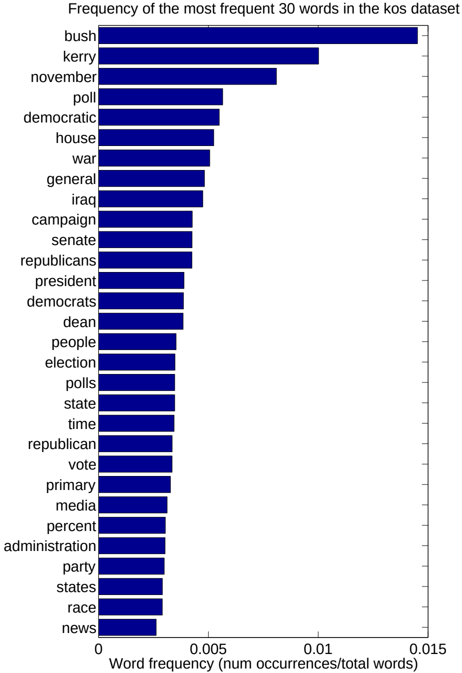
Example: word counts in text

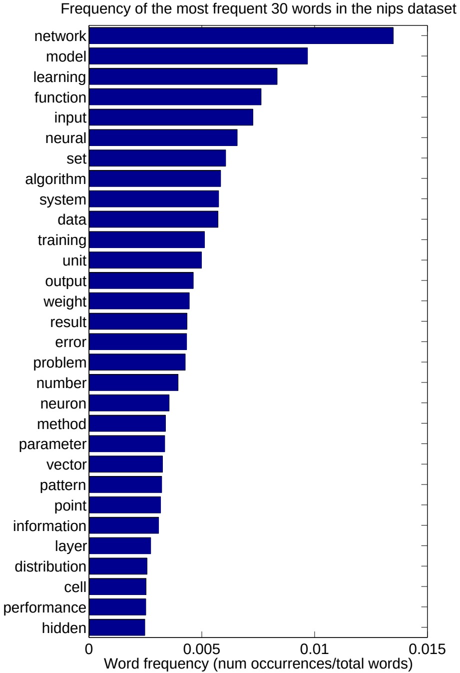
Consider describing a text document by the frequency of occurrence of every distinct word.

The UCI *Bag of Words* dataset from the University of California, Irvine. [1] For illustration consider two collections of documents from this dataset:

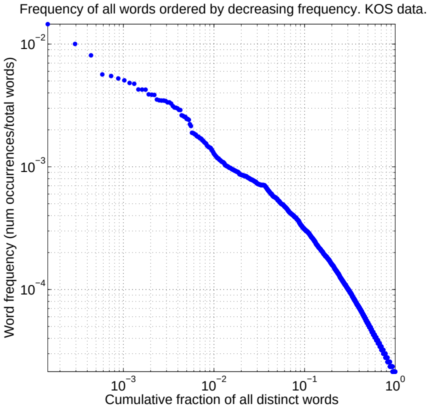
- KOS (political blog — `http://dailykos.com`):
  - $n = 353,160$ words
  - $m = 6,906$ *distinct* words
  - $D = 3,430$ documents (blog posts)
- NIPS (machine learning conference — `http://nips.cc`):
  - $n = 746,316$ words
  - $m = 12,375$ *distinct* words
  - $D = 1,500$ documents (conference papers)

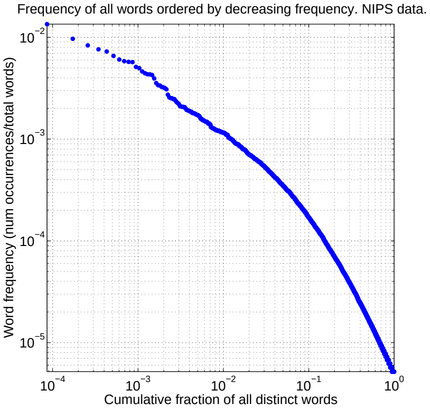
[1] `http://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/`


Frequency of the most frequent 30 words in the kos dataset
bush
kerry
november
poll
democratic
house
war
general
iraq
campaign
senate
republicans
president
democrats
dean
people
election
polls
state
time
republican
vote
primary
media
percent
administration
party
states
race
news
0
0.005
0.01
0.015
Word frequency (num occurrences/total words)


Frequency of the most frequent 30 words in the nips dataset
network
model
learning
function
input
neural
set
algorithm
system
data
training
unit
output
weight
result
error
problem
number
neuron
method
parameter
vector
pattern
point
information
layer
distribution
cell
performance
hidden
0
0.005
0.01
0.015
Word frequency (num occurrences/total words)

Different text collections, similar behaviour



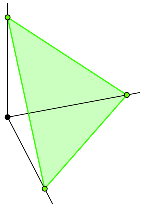


- Zipf's law states that *the frequency of any word is inversely proportional to its rank in the frequency table*.
- In these graphs the frequencies decay even faster.
- These words seem to be drawn from very particularly *sparse* Multinomials.

Priors on Multinomials: The Dirichlet distribution

The Dirichlet distribution is to the Multinomial what the Beta is to the Binomial. It is a generalisation of the Beta defined on the $m-1$ dimensional simplex.

- Consider the vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$, with $\sum_{i=1}^{m} \pi_i = 1$ and $\pi_i \in (0,1)\ \ \forall i$.
- Vector $\boldsymbol{\pi}$ lives in the open standard $m-1$ simplex.
- $\boldsymbol{\pi}$ could for example be the parameter vector of a Multinomial. [Figure on the right $m = 3$.]

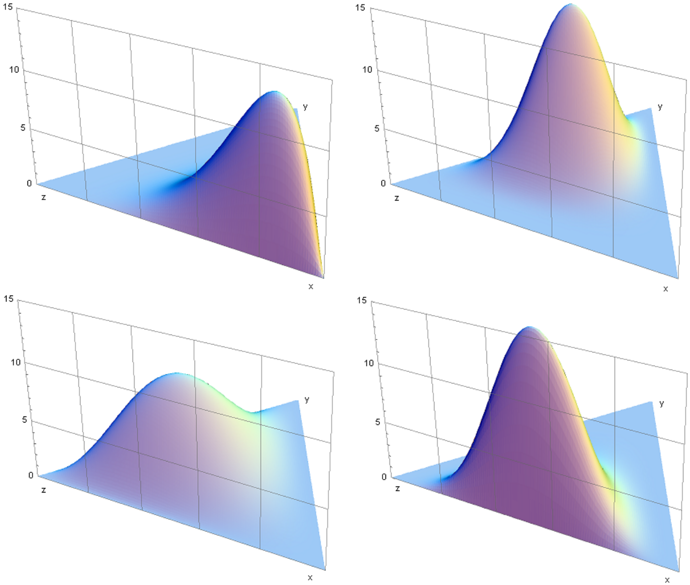


The Dirichlet distribution is given by $(0,1)$

$$\mathrm{Dir}(\boldsymbol{\pi}|\alpha_1, \ldots, \alpha_m) = \frac{\Gamma(\sum_{i=1}^{m} \alpha_i)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m} \pi_i^{\alpha_i - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{m} \pi_i^{\alpha_i - 1}$$

- $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m]^\top$ are the shape parameters.
- $B(\boldsymbol{\alpha})$ is the multinomial beta function.
- $E(\pi_j) = \frac{\alpha_j}{\sum_{i=1}^{m} \alpha_i}$ is the mean for the j-th element.

Dirichlet Distributions from Wikipedia



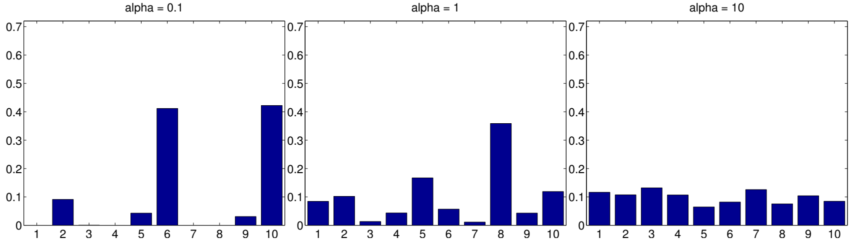
The symmetric Dirichlet distribution

In the symmetric Dirichlet distribution all parameters are identical: $\alpha_i = \alpha, \ \forall i.$

`en.wikipedia.org/wiki/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif`

To sample from a symmetric Dirichlet in D dimensions with concentration α use:

```
w = randg(alpha,D,1); bar(w/sum(w));
```



Distributions drawn at random form symmetric 10 dimensional Dirichlet distributions with various concentration parameters.