Topic Mining and Analysis: Overview of Statistical Language Models

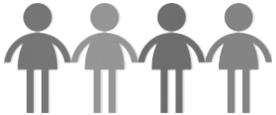
ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Probabilistic Topic Models: Overview of Statistical Language Models

5. Text-based prediction

3. Topic mining and analysis

Real World



Perceive

(Perspective)

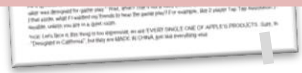
Observed World

Express

(English)

Text Data

1. Natural language processing and text representation

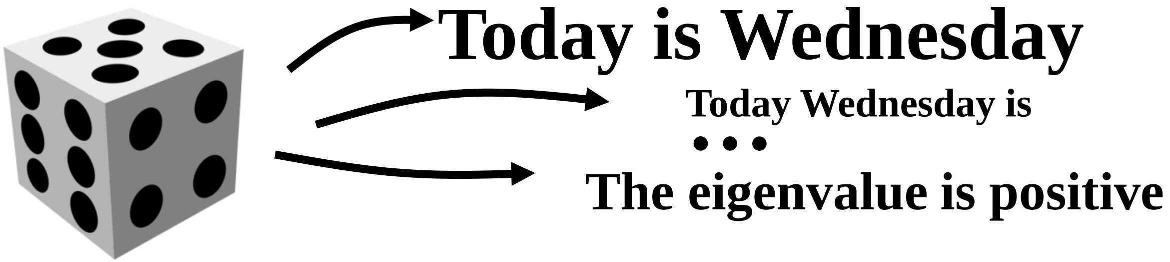


4. Opinion mining and sentiment analysis

2. Word association mining and analysis

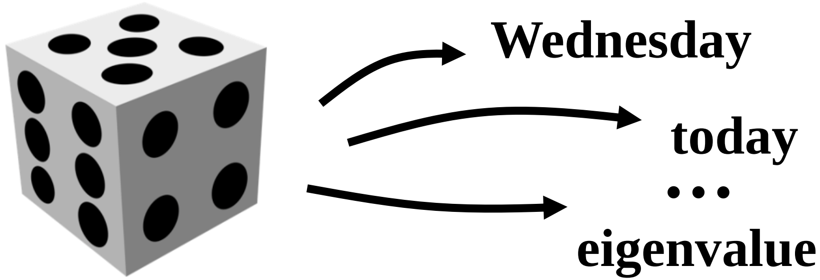
What Is a Statistical Language Model (LM)?

- A probability distribution over word sequences
  - p("*Today is Wednesday*") ≈ 0.001
  - p("*Today Wednesday is*") ≈ 0.0000000000001
  - p("*The eigenvalue is positive*") ≈ 0.00001
- Context-dependent!
- Can also be regarded as a probabilistic mechanism for "generating" text – thus also called a "generative" model



The Simplest Language Model: Unigram LM

- Generate text by generating each word INDEPENDENTLY
- Thus, $p(w_1\ w_2\ \ldots\ w_n)=p(w_1)p(w_2)\ldots p(w_n)$
- Parameters: $\{p(w_i)\}$ $p(w_1)+\ldots+p(w_N)=1$ (N is voc. size)
- Text = sample drawn according to this **word distribution**



p("today is Wed")

= p("today")p("is")p("Wed")

= 0.0002 × 0.001 × 0.000015

Text Generation with Unigram LM

Unigram LM $p(w|\theta)$ — **Sampling** → **Document d**

$p(d|\theta)$=?

Topic 1:
Text mining

…
text 0.2
mining 0.1
association 0.01
clustering 0.02
…
food 0.00001
…

→ **Text mining paper**

Topic 2:
Health

…
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
…

→ **Food nutrition paper**

Estimation of Unigram LM

Unigram LM $p(w|\theta)=?$ ← **Estimation** ← **Text Mining Paper d**

Total #words=**100**

Maximum Likelihood Estimate	Unigram LM
	…
10/100	text ?
5/100	mining ?
3/100	association ?
3/100	database ?
	…
1/100	query ?
	…

Maximum Likelihood Estimate

text 10
mining 5
association 3
database 3
algorithm 2
…
query 1
efficient 1

Is this our best estimate?

How do we define "best"?

Maximum Likelihood vs. Bayesian

- Maximum likelihood estimation
  - "Best" means "data likelihood reaches maximum"

$$\hat{\theta} = \arg\max_{\theta} P(X \mid \theta)$$

  - Problem: Small sample
- Bayesian estimation:

Bayes Rule $p(X \mid Y) = \dfrac{p(Y \mid X)p(X)}{p(Y)}$

  - "Best" means being consistent with our "prior" knowledge and explaining data well

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid X) = \arg\max_{\theta} P(X \mid \theta)P(\theta)$$

  - Problem: How to define prior?

Maximum a Posteriori (MAP) estimate

Illustration of Bayesian Estimation

Bayesian inference: $f(\theta)=?$

$$\hat{f}(\theta) = \sum_{\theta} f(\theta) p(\theta \mid X)$$

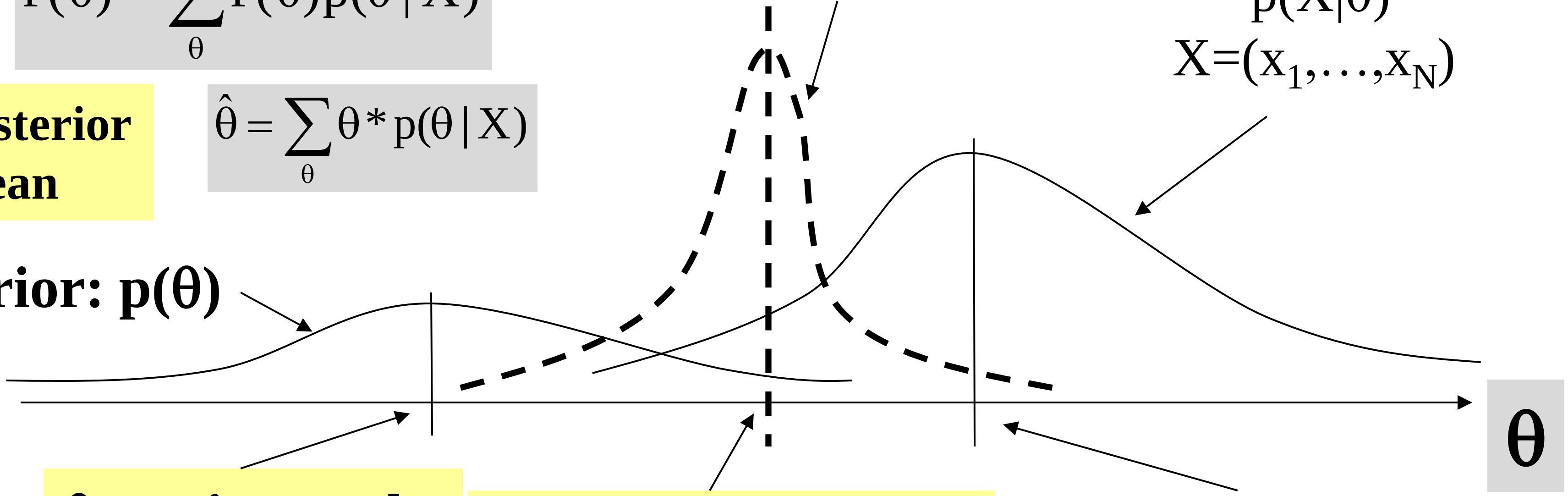
Posterior Mean

$$\hat{\theta} = \sum_{\theta} \theta * p(\theta \mid X)$$

Posterior:
$p(\theta|X) \propto p(X|\theta)p(\theta)$

Likelihood:
$p(X|\theta)$
$X=(x_1,\ldots,x_N)$

Prior: $p(\theta)$



θ_0: prior mode

θ_1: posterior mode

θ_{ml}: ML estimate

Summary

- **Language Model** = probability distribution over text = generative model for text data
- **Unigram** Language Model = **word distribution**
- **Likelihood** function: **p(X|θ)**
  - **Given θ** ➔ which X has a higher likelihood?
  - **Given X** ➔ which θ maximizes p(X| θ)? **[ML estimate]**
- **Bayesian** estimation/inference
  - Must define a **prior: p(θ)**
  - **Posterior** distribution**: p(θ|X)∝ p(X|θ)p(θ)**

  ➔ **Allows for inferring any "derived value" from θ!**