UNIVERSITY of DELAWARE

Statistical Language Models

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE

Statistical Language Models

- A statistical language model is a probability distribution over word sequences.
- Given a sequence of words, it computes the probability of a sentence of words:

  $P(W) = P(w_1, w_2, w_3, w_4, w_5 \ldots w_n)$
- Examples

$$P(Today\ is\ Thursday) = 0.01$$
$$P(Today\ Thursday\ is) = 0.000000000000000001$$

UNIVERSITY of DELAWARE

Language models are useful in many applications

- Text categorization
  - P(topic = sports | baseball, baseball, baseball, game)
- Speech recognition
  - P(happy | John feels) >> P(habit | John feels)
- Information retrieval
  - P(an article about dog | puppy) > P(an article about cat | puppy)

UNIVERSITY of DELAWARE

Go back to the problem of computing joint probability of words in sentence

$$P(W) = P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P("its water is so transparent") =
P(its) × P(water|its) × P(is|its water)
× P(so|its water is) × P(transparent|its water is so)

UNIVERSITY of DELAWARE

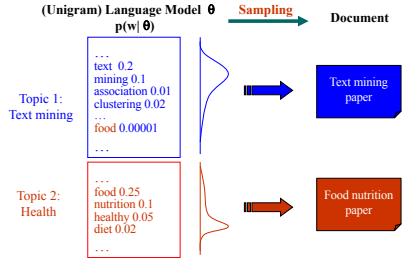
Simplest case: Unigram model

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

- Assumption:
  - Each word is generated independently.
- Parameters:
  - $\{p(w_i)\}$
  - $p(w_1)+\ldots+p(w_N)=1$ (N is voc. size)
- This is essentially a multinomial distribution over words
- A piece of text can be regarded as a sample drawn according to this word distribution.

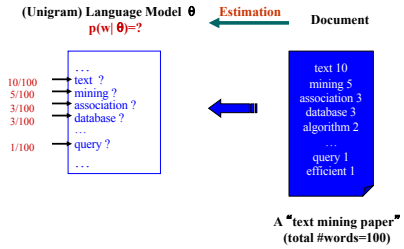
UNIVERSITY of DELAWARE

Text Generation with Unigram LM



UNIVERSITY of DELAWARE

Estimation of Unigram LM



A "text mining paper"
(total #words=100)

8

UNIVERSITY of DELAWARE

Maximum Likelihood (ML) Estimation

- Finding parameters that yield to maximum likelihood

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

9

UNIVERSITY of DELAWARE

Maximum Likelihood Estimation

Data: a document d with counts $c(w_1,d), \ldots, c(w_N,d)$, and length $|d|$
Model: multinomial distribution θ with parameters $\{\theta_i = p(w_i)\}$ $\sum_{i=1}^{N} \theta_i = 1$
Likelihood: $p(d|\theta)$
Maximum likelihood estimator: $\hat{\theta} = \arg\max_{\theta} P(d \mid \theta)$

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

$$\log P(d \mid \theta) = \log\left(\prod_{i=1}^{N} P(w_i \mid \theta)^{c(w_i,d)}\right) = \sum_{i=1}^{N} c(w_i,d) \log P(w_i \mid \theta) = \sum_{i=1}^{N} c(w_i,d) \log \theta_i$$

Subject to constraints: $\sum_{i=1}^{N} \theta_i = 1$

Use Lagrange multiplier approach $f(\theta \mid d) = \sum_{i=1}^{M} c(w_i,d) \log \theta_i + \lambda \left(\sum_{i=1}^{M} \theta_i - 1\right)$

11

UNIVERSITY of DELAWARE

Maximum Likelihood Estimation

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

$$\log P(d \mid \theta) = \log\left(\prod_{i=1}^{N} P(w_i \mid \theta)^{c(w_i,d)}\right) = \sum_{i=1}^{N} c(w_i,d) \log P(w_i \mid \theta) = \sum_{i=1}^{N} c(w_i,d) \log \theta_i$$

Subject to constraints: $\sum_{i=1}^{N} \theta_i = 1$

Use Lagrange multiplier approach $f(\theta \mid d) = \sum_{i=1}^{M} c(w_i,d) \log \theta_i + \lambda \left(\sum_{i=1}^{M} \theta_i - 1\right)$

Set partial derivatives to zero $\frac{\partial f(\theta \mid d)}{\partial \theta_i} = \frac{c(w_i,d)}{\theta_i} + \lambda = 0 \rightarrow \theta_i = -\frac{c(w_i,d)}{\lambda}$

$$\sum_{i=1}^{M} -\frac{c(w_i,d)}{\lambda} = 1 \rightarrow \lambda = -\sum_{i=1}^{M} c(w_i,d) \rightarrow \hat{\theta}_i = p(w_i \mid \hat{\theta}) = \frac{x(w_i,d)}{\sum_{i=1}^{M} c(w_i,d)} = \boxed{\frac{c(w_i,d)}{|d|}}$$

ML estimator

12

UNIVERSITY of DELAWARE

Problem with the ML Estimator

- What probability should we give a word that has not been observed?
- If we want to assign non-zero probabilities to unseen words, we need have to discount the probabilities of seen words
- This is what "smoothing" is about …

13

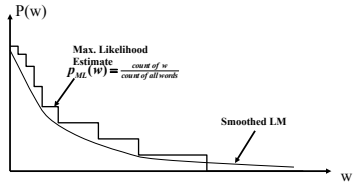
UNIVERSITY of DELAWARE

How to Smooth?

- All smoothing methods try to
  - discount the probability of words seen in a document
  - re-allocate the extra counts so that unseen words will have a non-zero count

UNIVERSITY of DELAWARE

Language Model Smoothing



UNIVERSITY of DELAWARE

Smoothing Method 1 - Additive Smoothing

- Add a constant δ to the counts of each word

$$p(w \mid d) = \frac{c(w,d)+1}{|d|+|V|}$$

Counts of w in d — "Add one", Laplace smoothing

Length of d (total counts) — Vocabulary size

- Any problems?

UNIVERSITY of DELAWARE

Using Reference Model for Smoothing

- Should all unseen words get equal probabilities?
- We can use a reference model to discriminate unseen words

$$p(w \mid d) = \begin{cases} p_{seen}(w \mid d) & \text{if } w \text{ is seen in } d \\ \alpha_d p(w \mid REF) & \text{otherwise} \end{cases}$$

Discounted ML estimate

Reference language model

$$\alpha_d = \frac{1 - \sum_{w \text{ is seen}} p_{seen}(w \mid d)}{\sum_{w \text{ is unseen}} p(w \mid REF)}$$

UNIVERSITY of DELAWARE

Smoothing Method 2 - Jelinek Mercer (JM) Smoothing

$$p(w \mid d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda\, p(w \mid REF)$$

ML estimate

parameter

UNIVERSITY of DELAWARE

Smoothing Method 3 - Dirichlet Smoothing

- Assume pseudo counts μp(w|REF)

$$p(w \mid d) = \frac{c(w;d)+\mu\, p(w|REF)}{|d|+\mu} = \frac{|d|}{|d|+\mu}\frac{c(w,d)}{|d|} + \frac{\mu}{|d|+\mu}p(w \mid REF)$$

parameter