-10-2005

.M-60's

M-90's

Retrieval Models: Language Model

① Create a probability model for each document
② Often create a probability model for query
[① & ② often depend on prob model for corpus]
③ Compare ① & ② - docs which are "closer" are ranked higher

Simple view

$$\vec{P}_{t;d} \quad \frac{tf_{t;d}}{\sum_t tf_{t;d}}$$

eg. The quick brown dog dog
t: the, quick, brown, dog
$(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{2}{5})$

q = "brown dogs"

? $Pr[(brown, dog)|d] = \prod_{t \in \Sigma} Pr[t|d] = \frac{1}{5} \cdot \frac{2}{5} = \frac{2}{25}$

Mathematical Aside

$Pr(s)$ - probability of ("event") s
$Pr(s|M)$ - probability of S given M
Independence: $Pr(S_1, S_2) = Pr(S_1) \cdot Pr(S_2)$
$Pr(q_1, q_2, q \ldots | M) = Pr(q_1|M) \cdot P(q_2|M) \ldots = \prod_{q_i} Pr(q_i|M)$
Bayes Rule: $Pr(A|B) = \frac{Pr(B|A) \cdot Pr(A)}{P(B)} = \frac{Pr(B|A) \cdot P(A)}{\sum_A Pr(B|A) \cdot Pr(A)}$

Coding / Information Theory

① Coding & Entropy:
- Suppose events generated by dist $(P_1, P_2, \ldots P_n)$
- optimum code length for shortest avg code length: $\log \frac{1}{P_i}$
- avg code: $= P_1 \cdot \log \frac{1}{P_1} + P_2 \cdot \log \frac{1}{P_2} \ldots = \sum_i P_i \log \frac{1}{P_i} = H(\vec{P}) =$ entropy of dist

Relative Entropy (cross-entropy, KL-distance)
$KL(\vec{P} \| \vec{q}) = \sum_i P_i \log \frac{1}{q_i} - \sum_i P_i \log \frac{1}{P_i} = \sum_i P_i \log \frac{P_i}{q_i}$
- how many bits are wasted by encoding p with q's encoding
- not symmetric

all $\log_2 = \log$

$\log_a b = \frac{\log_c b}{\log_c a}$

$\log_2 = \lg$
$\log$
$\ln = \log_3$

Language Models for Documents

Unigram model: $Pr(t|d)$ = relative frequency of term in doc

$$= \frac{tf_{t,d}}{\sum_{t \in d} tf_{t,d}}$$

Higher order models
- n-gram : 2-gram - look at word pairs ...
- caching : www w... based on a window
- grammar : based on a parse tree

- don't help significantly

We will consider unigram models & "smooth" them later

Now, assume we don't model queries, just compute prob of generating query from model (s).

Two ways
(1) Multi-nomial : $Pr[q_1, q_2 \ldots q_n | M] = \prod Pr(q_i | M)$
(2) Multiple Bernoulli : $Pr[q_1, q_2 \ldots q_n | M] =$

$$\prod_{w \in q_1 \ldots q_n} P(w|M) \prod_{w \notin q_1 \ldots q_n} (1 - P(w|M))$$

query matches but query

Generate doc from query

$$P[D|q] = \frac{P(q|D) \cdot P(D)}{P(q)}$$

$P(D)$ - belief in how similar to D in corpus

doesn't affect rank of docs

L.M. cont.

5/11/2005

Issues w/ L.M.
① How to model docs? - prob dist over words
② How to model queries?
③ How do "assess" similarity of doc to query (or vice versa)
  L 2 methods: $P[q|D]$ *
  $P[D|q]$
  third: $sim(d,q)$ - use KL-distance

* - likely to perform the best

Let D be a dist corresponding to a doc
Let Q be a dist corresponding to a query
$\Rightarrow KL(Q||D)$ - "distance" from query to doc *
$= \sum q_i \lg \frac{q_i}{p_i}$

Smoothing
Issue: how to gen dist over document
- probably not enough data to build good model just on doc itself
- should incorporate some "prior belief" based on dist in entire corpus
- have to move smoothly from prior belief to totally evidence based

Generalize using a coin

Laplace Smoothing - determining estimates for probabilities from data
- Problem: Given lots of data the maximum likelihood estimate
  eg: $Pr(\text{heads}) = \frac{\#\text{heads}}{\#\text{trials}}$ is good. for limited data, this is often poor

Slowly...
Consider 5 possibilities: Evidence: E: the sun has risen each of the last N days

Hypotheses
- $H_1 = Pr = 0 = h_1$
- $H_2 = Pr = \frac{1}{4} = h_2$
- $H_3 = Pr = \frac{1}{2} = h_3$
- $H_4 = Pr = \frac{3}{4} = h_4$
- $H_5 = Pr = 1 = h_5$

$P[E|H_i] = h_i^N$

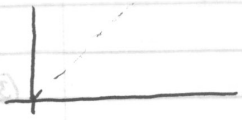
$P[H_i] = \frac{1}{5}$ uniform, no assumptions

$$Pr[H_i|E] = \frac{Pr[E|H_i]\,Pr[H_i]}{Pr[E]} = \frac{Pr[E|H_i]\cdot Pr[H_i]}{\sum_i Pr[E|H_i]\cdot Pr[H_i]}$$

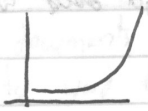
$$= \frac{h_i^N \cdot \frac{1}{5}}{\sum_i h_i^N \cdot \frac{1}{5}} = \frac{h_i^N}{\sum_i h_i^N}$$

eg: $N=0$ $P[H_i|E] = \frac{1}{5}$

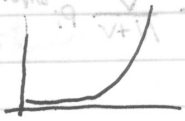
$N=1$ $P[H_i|E] = \frac{h_i}{\sum_i h_i} = \frac{h_i}{0+\frac{1}{4}+\frac{1}{2}+\frac{3}{4}+1} = \frac{h_i}{5/2} = \frac{2}{5}h_i$



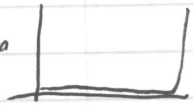
$N=2$



$N=3$



$N=1{,}000{,}000$

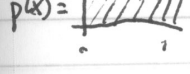


What's the answer: $1\cdot(.4) + (\frac{3}{4})\cdot(.3) + (\frac{1}{2})(.2) + (\frac{1}{4})\cdot(.1) + 0(0) = \frac{3}{4}$

- break even point

Now consider <u>all</u> possible hypotheses $X \in [0,1]$

$p(x) = 1$



$$Pr[x|E] = \frac{Pr[E|x]\cdot 1}{\int_{x=0}^{1} Pr[E|x]\cdot 1\,dx} = \frac{x^N \cdot 1}{\int_{x=0}^{1} x^N dx} = \frac{x^N}{\frac{x^{N+1}}{N+1}\Big|_0^1} = \frac{x^N}{\frac{1}{N+1}} = (N+1)x^N$$

<u>Expected Answer</u>:

$$\int_0^1 (N+1)x^N \cdot x\,dx = (N+1)\int_0^1 x^{N+1}dx = (N+1)\frac{x^{N+2}}{N+2}\Big|_0^1 = \frac{N+1}{N+2} = p_i$$

So, if N successes in N trials, belief in success is $\frac{N+1}{N+2}$ for coin

<u>Generalization</u>: ① If M successes in N trials, belief in success is $\frac{M+1}{N+2}$

② if M_i occurences of outcome i in N trials for K total possible outcomes:

$$\frac{M_i+1}{N+K} = p_i$$

Max Likelihood = $\frac{M_i}{N} = p_i$

Generalizations of Laplace Smoothing

① Jelinek-Mercer - take into account non-uniform prior (eg. corpus prior)

$$p_i = \lambda p_i^{ML} + (1-\lambda) p_i^{corpus} \qquad \lambda = 1 - \frac{k}{N+k}$$

② Dirichlet Smoothing : μ - arbitrary - generalization of k

$$\frac{N}{N+\mu} p_i^{ML} + \frac{\mu}{N+\mu} p_i^{corpus}$$

③ Witten-Bell

$$\frac{N}{N+V} p_i^{ML} + \frac{V}{N+V} p_i^{corpus}$$

N = length of doc

V = # unique words in doc

5-24-2004

Language Model Example

D1 The cat in the hat
D2 one fish two fish red fish blue fish
D3 The fish in the red hat

3 documents in corpus
Corpus Vocab - 9 words (unique)

	The	cat	in	hat	one	fish	two	red	blue	Doc Length
D1	2	1	1	1	0	0	0	0	0	5
D2	0	0	0	0	1	4	1	1	1	8
D3	2	0	1	1	0	1	0	1	0	6

Max Likelihood Model

D1	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0	0	0	0	0
D2	0	0	0	0	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
D3	$\frac{2}{6}$	0	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0

Query: "red fish"
- Score associated w/ doc is prob. query words would be generated when drawing words at random according to LM.

$Score_{ML}(D1) = 0 \cdot 0 = 0$
$Score_{ML}(D2) = \frac{1}{8} \cdot \frac{4}{8} = \frac{4}{64} = \frac{1}{16}$
$Score_{ML}(D3) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

Smoothing

Laplace: $\frac{\#\text{ Successes} + 1}{\#\text{ trials} + \#\text{ outcomes (possible)}}$

	the	cat	in	hat	one	fish	two	red	blue
D1	$\frac{2+1}{5+9}$	$\frac{1+1}{5+9}$	$\frac{1+1}{5+9}$	$\frac{1+1}{5+9}$	$\frac{0+1}{5+9}$	$\frac{2+1}{5+9}$	$\frac{0+1}{5+9}$	$\frac{0+1}{5+9}$	$\frac{0+1}{5+9}$
D2	...								
D3	...								

Laplace Smoothing

red fish

$\text{Score}_{\text{Laplace}}(D1) = \frac{1}{14} \cdot \frac{1}{14} =$ small but not 0

$\text{Score}_{\text{Laplace}}(D2) = \frac{1+1}{8+9} \cdot \frac{4+1}{8+9} = \frac{2}{17} \cdot \frac{5}{17}$

$\text{Score}_{\text{Laplace}}(D3) =$

Witten-Bell smoothing

$$LM(D) = \frac{\#\text{ words in doc}}{\#\text{ words in doc} + \#\text{ unique words in doc}} \cdot LM_{ML} + \frac{\#\text{ unique words in doc}}{\#\text{ words in doc} + \#\text{ unique words in doc}} \cdot LM_{corpus}$$

(Max likelihood model)

need corpus language model...

often corpus LM is average of ML LM's

	the	cat	in	hat	one	fish	two	red	blue
corpus	$\frac{\frac{2}{5}+0+\frac{2}{10}}{3}$	$\frac{\frac{1}{5}+0+0}{3}$							

	the	cat	in	hat
$LM_{WB}(D1)$	$\frac{5}{5+4} \cdot \frac{2}{5} + \frac{4}{5+4} \cdot \frac{\frac{2}{5}+0+\frac{2}{10}}{3}$			

Rational for Witten-Bell smoothing

$V = \#$ unique words in doc Heaps Law: $V = KN^{\beta}$ $\beta \approx \frac{1}{2}$

$N = \#$ words in doc

$\frac{N}{N+V} \propto \frac{N}{N+KN^{\beta}}$ (grows with N) $\frac{V}{N+V} = \frac{KN^{\beta}}{N+KN^{\beta}}$ - shrinks w/ N

5-24-2004

Language Model Example

D1 The cat in the hat
D2 one fish two fish red fish blue fish
D3 The fish in the red hat

3 documents in corpus
Corpus Vocab - 9 words (unique)

	The	cat	in	hat	one	fish	two	red	blue	Doc Length
D1	2	1	1	1	0	0	0	0	0	5
D2	0	0	0	0	1	4	1	1	1	8
D3	2	0	1	1	0	1	0	1	0	6

Max Likelihood Model

	The	cat	in	hat	one	fish	two	red	blue
D1	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0	0	0	0	0
D2	0	0	0	0	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
D3	$\frac{2}{6}$	0	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0

Query: "red fish"
- Score associated w/ doc is prob. query words would be generated when drawing words at random according to LM.

$$Score_{ML}(D1) = 0 \cdot 0 = 0$$
$$Score_{ML}(D2) = \frac{1}{8} \cdot \frac{4}{8} = \frac{4}{64} = \frac{1}{16}$$
$$Score_{ML}(D3) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Smoothing

Laplace: $\frac{\#\text{ successes} + 1}{\#\text{ trials} + \#\text{ outcomes (possible)}}$

	the	cat	in	hat	one	fish	two	red	blue
D1	$\frac{2+1}{5+9}$	$\frac{1+1}{5+9}$	$\frac{1+1}{5+9}$	$\frac{1+1}{5+9}$	$\frac{0+1}{5+9}$	$\frac{0+1}{5+9}$	$\frac{0+1}{5+9}$	$\frac{0+1}{5+9}$	$\frac{0+1}{5+9}$
D2	. . .								
D3	. . .								

Laplace Smoothing

red fish

$\text{Score}_{\text{Laplace}}(D1) = \frac{1}{14} \cdot \frac{1}{14} =$ small but not 0

$\text{Score}_{\text{Laplace}}(D2) = \frac{1+1}{8+9} \cdot \frac{4+1}{8+9} = \frac{2}{17} \cdot \frac{5}{17}$

$\text{Score}_{\text{Laplace}}(D3) =$

Witten-Bell smoothing

$$LM(D) = \frac{\#\text{ words in doc}}{\#\text{ words in doc} + \#\text{ unique words in doc}} \cdot LM_{ML} + \frac{\#\text{ unique words in doc}}{\#\text{ words in doc} + \#\text{ unique words in doc}} \cdot LM_{\text{corpus}}$$

(Max likelihood model)

need corpus language model...

often corpus LM is average of ML LMs

corpus:

the	cat	in	hat	one	fish	two	red	blue
$\frac{\frac{2}{5}+0+\frac{2}{10}}{3}$	$\frac{\frac{1}{5}+0+0}{3}$							

$LM_{WB}(D1)$:

the	cat	in	hat
$\frac{5}{5+4} \cdot \frac{2}{5} + \frac{4}{5+4} \cdot \frac{\frac{2}{5}+0+\frac{2}{10}}{3}$			

---

rational for Witten-Bell smoothing

V = # unique words in doc

N = # words in doc

Heaps Law: $V = KN^{\beta}$ $\beta \approx \frac{1}{2}$

$$\frac{N}{N+V} \propto \frac{N}{N+KN^{\beta}}$$ grows with N

$$\frac{V}{N+V} = \frac{KN^{\beta}}{N+KN^{\beta}}$$ - shrinks w/ N

generate query — query — document model

perform better — Two ways: (1) Multinomial : $Pr[q_1 q_2 \cdots q_k | M] = \prod_{i=1}^{k} Pr[q_i | M]$

came first — (2) Multiple Bernoilli

generate these query words and no other word

$Pr[q_1 q_2 \cdots q_k | M] = \prod_{w_i = q_1 q_2 \cdots q_k} Pr(w_i | M) \cdot \prod_{w_i \neq q_1 q_2 \cdots q_k} (1 - Pr(w_i | M))$

generate query and nothing but query

Smoothing in language modeling = parameter estimation in statistics.

→ Why calculate prob. of generating a query given a doc. but not prob. of a document given a query?

Generate doc from query instead:

$Pr(D|q) = \frac{Pr(q|D) \boxed{Pr(D)}}{(Pr(q))}$ → doesn't affect ranking of docs

or

↳ can generate probability models for queries also. Then, u can use KL-distance to see how two distributions are different from each other.

Lecture #5

Issues w/ L.M.

1) How to model docs?
2) How to model queries?
3) How to "assess" sim. of doc. to query (or vice versa)?
4) ↳ two methods: $P(q|D)$
   $P(D|q)$

third: $sim(d, q)$ - use KL-distance

Let D be a distr. corresponding to a doc.
Let Q be a distr. corresponding to a query
$\Rightarrow$ KL(Q||D) → "distance" from query to doc.

$$= \sum_i q_i \log \frac{q_i}{d_i}$$

Laplace Smoothing

Determining estimate for probabilities from data.

- Problem: Given lots of data, the maximum likelihood estimate, e.g.

$$\Pr[\text{Heads}] = \frac{\#\text{heads}}{\#\text{trials}}$$ is good.

- for limited data, this is often poor.

Will sun rise tomorrow?

Consider 5 possibilities

hypotheses
- H_1 - $\Pr = 0$ $= h_1$
- H_2 - $\Pr = 1/4$ $= h_2$
- H_3 - $\Pr = 1/2$ $= h_3$
- H_4 - $\Pr = 3/4$ $= h_4$
- H_5 - $\Pr = 1$ $= h_5$

Evidence E: The sun has risen each of the last N days.

$$\Pr[H_i|E] = \frac{\Pr[E|H_i]\cdot\Pr[H_i]}{\Pr[E]}$$

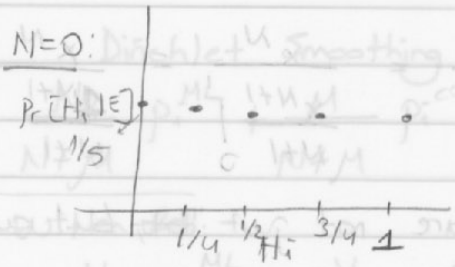
$$= \frac{\Pr[E|H_i]\Pr[H_i]}{\sum_i \Pr[E|H_i]\Pr[H_i]}$$

$\sim$ uniform

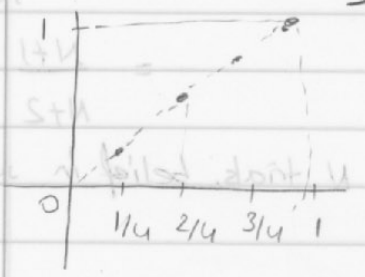
$$= \frac{h_i^N \cdot 1/5}{\sum_i h_i^N \cdot 1/5} = \frac{h_i^N}{\sum_i h_i^N}$$

Emine Yilmaz

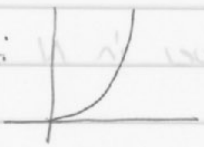
e.g. N=0:



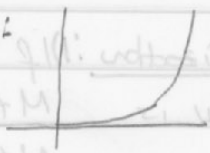
N=1: $Pr[H_i|E] = \frac{h_i}{\sum h_i} = \frac{h_i}{0+1/4+1/2+3/4+1} = \frac{h_i}{5/2} = \frac{2}{5} h_i$



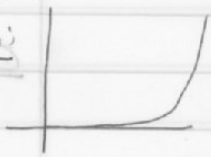
N=2:



N=3:



N=4:



N=3: $Pr[H_i|E] = \frac{h_i^N}{\sum_i h_i^N} = \frac{h_i^3}{\sum_i h_i^3} = \frac{h_i^3}{0^3+(1/2)^3+}$

Problem: Laplace (true) considers only a small set that has been seen. However, in reality, there can be much more different events happening.

Now consider all possible hypotheses $x \in [0, 0.$

$$Pr[H^x|E] = \frac{Pr[E|x] \cdot (1)}{\int_{x=0}^{1} Pr[E|x] \cdot 1 \, dx} \longrightarrow Pr(x) = \text{uniform}$$

$$= \frac{x^N \cdot 1}{\int_0^1 x^N dx} = \frac{x^N}{\left.\frac{x^{N+1}}{N+1}\right|_0^1} = \frac{x^N}{\frac{1^{N+1}}{N+1}} = (N+1)x^N$$

My beliefs now are not just dots, but curves.

Expected answer:

$$\int_0^1 (N+1)x^N \cdot x\,dx = (N+1)\int_0^1 x^{N+1} dx = \left.(N+1)\frac{x^{N+2}}{N+2}\right|_0^1$$

$$= \frac{N+1}{N+2}$$

So, if N successes in N trials, belief in success is $\frac{N+1}{N+2}$.

Generalization: 1) If M successes in N trials, belief in success is $\frac{M+1}{N+2}$

2) If M_i occurrences of outcome i in N trials for K total possible outcomes

$$\frac{M_i+1}{N+k}$$

how many times I see the word ← M_i

docs in my db ← N

Max. Likelihood $= \frac{M_i}{N} = p_i^{ML}$

Generalizations of Laplace Smoothing

1) Jelinek - Mercer : take into account non-uniform prior (e.g. corpus prior)

$$p_i = \lambda \cdot p_i^{ML} + (1-\lambda)\, p_i^{corpus} \qquad \lambda = 1 - \frac{1}{N+k}$$

Dirichlet Smoothing

$$\frac{N}{N+\mu} p_i^{ML} + \frac{\mu}{N+\mu} p_i^{corpus}$$

Witten-Bell:

$$\frac{N}{N+V} p_i^{ML} + \frac{V}{N+V} p_i^{corpus}$$

N = length of doc.

V = # unique words in doc.