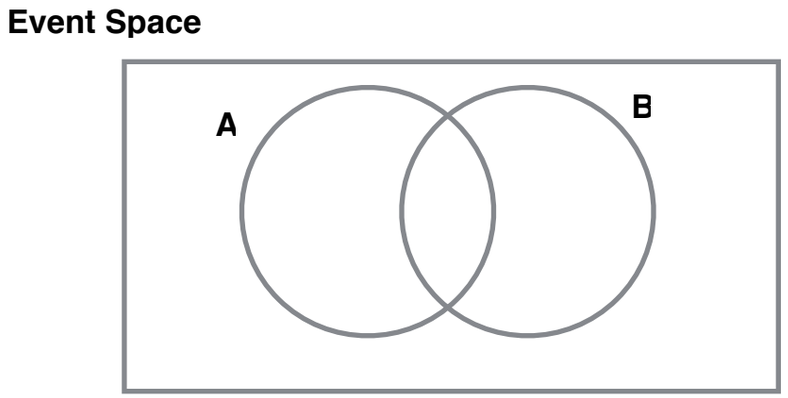
Ranking-I

Probability Basics and Language Models

Probability Basics : Independence and Conditionals

- Event space: All possible events
  - Coin: {T, H}, doc: {w_1, …. , $w_{|v|}$}, dice: {1,2,3,4,5,6}
- Total probability principle: Sum of prob. of events is 1

$$\sum_{x_i \in E} P(x_i) \;=\; 1$$



- Independence: A and B are independent if and only if
  - P (A and B) = P(A) . P(B)
  - independence means "unrelated" — consecutive throws of coins
- Dependence: A and B are influenced by each other
  - P (A|B) is read as prob. of A given that B is true or B is given

Probability Basics : Conditional Independence

A = X Studies CS
B = X knows programming
C = X plays cricket

A and B dependent

$$P(A,B) = P(A \mid B) \cdot P(B)$$
$$= P(B \mid A) \cdot P(A)$$

joint distribution

conditional

B and C independent

$$P(B,C) = P(B) \cdot P(C)$$
$$P(B \mid C) = P(B)$$

Probability Basics : Conditional Independence

A = X Studies CS
B = X knows programming
C = X plays cricket

A and B dependent

$$P(A,B) = P(A \mid B) \cdot P(B)$$
$$= P(B \mid A) \cdot P(A)$$

joint distribution

conditional

B and C independent

$$P(B,C) = P(B) \cdot P(C)$$
$$P(B \mid C) = P(B)$$

chain rule of prob.

$$P(A,B,C) = P(A \mid B, C) \cdot P(B,C) = P(A \mid B, C) \cdot P(B \mid C) \cdot P(C)$$

Probability Basics : Conditional Independence

A = X Studies CS
B = X knows programming
C = X plays cricket

A and B dependent

$$P(A,B) = P(A \mid B) \cdot P(B)$$
$$= P(B \mid A) \cdot P(A)$$

joint distribution

conditional

B and C independent

$$P(B,C) = P(B) \cdot P(C)$$
$$P(B \mid C) = P(B)$$

chain rule of prob.

$$\mathbf{P(A,B,C) = P(A \mid B, C) \cdot P(B,C) = P(A \mid B, C) \cdot P(B \mid C) \cdot P(C)}$$

A is **conditionally** independent to C given B → $P(A \mid B,C) = P(A \mid B)$

Probability Basics : Conditional Independence

A = X Studies CS
B = X knows programming
C = X plays cricket

A and B dependent

$$P(A,B) = P(A \mid B) \cdot P(B)$$
$$= P(B \mid A) \cdot P(A)$$

joint distribution

conditional

B and C independent

$$P(B,C) = P(B) \cdot P(C)$$
$$P(B \mid C) = P(B)$$

chain rule of prob.

$$\mathbf{P(A,B,C) = P(A \mid B, C) \cdot P(B,C) = P(A \mid B, C) \cdot P(B \mid C) \cdot P(C)}$$

A is **conditionally** independent to C given B → $P(A \mid B,C) = P(A \mid B)$

$$\mathbf{P(A,B,C) = P(A \mid B) \cdot P(B) \cdot P(C)}$$

Probability Basics : Bayes Theorem

$$P(A,B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

bayes' theorem

$$P(A \mid B) = P(B \mid A) \cdot P(A) / P(B)$$

posterior prob.

prior prob.

Basis for all bayesian inferencing

What is the P("holmes" | "sherlock") given that P ("sherlock", "holmes"), P ("sherlock") are known ?

Parameter Estimation

Given observed data D

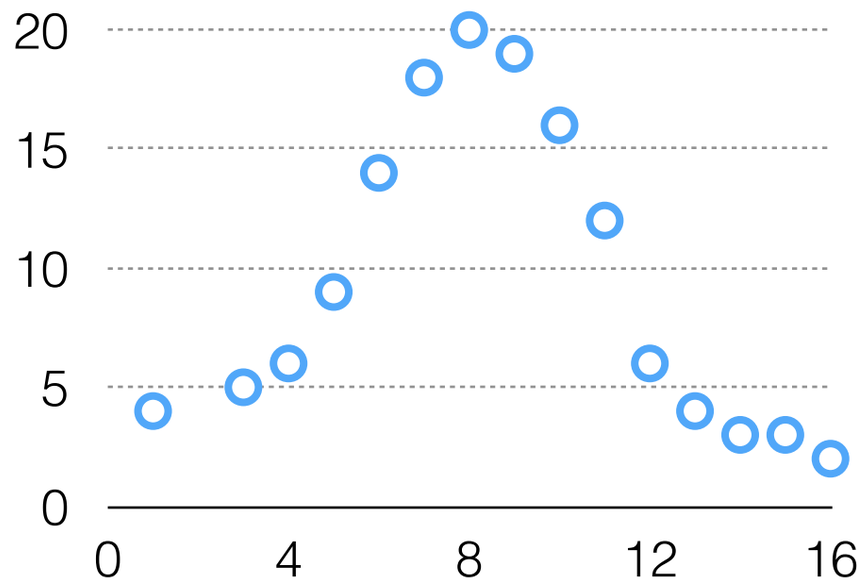
Assume that you have a probability density func. p(x) with parameters θ which generates it

parameter estimation problem

What are the value of the parameters which best explain my data ?

Say you feel that D is generated from a Normal distribution

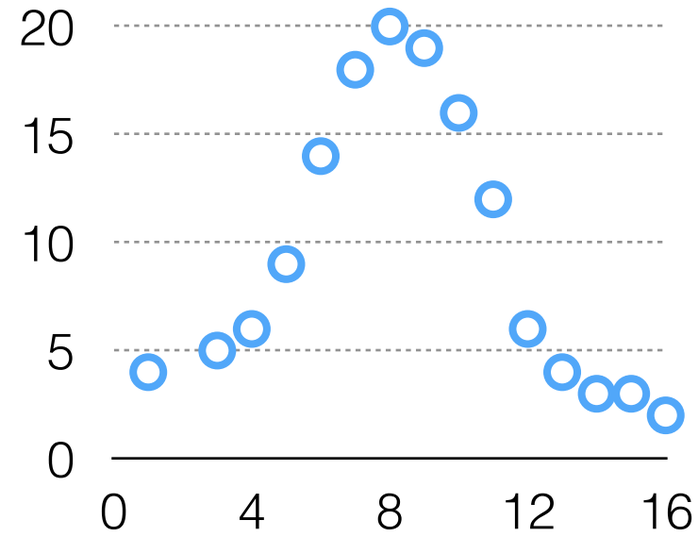
$$\theta = N(\mu, \sigma)$$



Maximum Likelihood Estimator (MLE)

Say you feel that D is generated from a Normal distribution

$$\theta = N(\mu, \sigma)$$



The **likelihood** of observing this independent and identically distributed sample is

$$\mathcal{L}(\theta \mid x = (x_1, x_2, \ldots, x_n)) = p(x \mid \theta)$$

parameters to be estimated

observed data

MLE chooses $\hat{\theta}$ maximizes this likelihood of generating the observed data

Maximum Likelihood Estimator (MLE)

- Desired probability distribution is the one that makes the observed data "most likely," which means that one must seek the value of the parameter vector $\hat{\theta}$ that maximizes the likelihood function $\mathcal{L}(\hat{\theta}|x)$

- For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function $ln(\,\mathcal{L}(\hat{\theta}|x)\,)$

$$ln(\,\mathcal{L}(\hat{\theta}|x)\,) = \sum_{x_i} ln\; p(x_i\,|\,\theta)$$

- If log-likelihood is differentiable find $\hat{\theta}$ partial derivatives

Maximum Likelihood Estimator (MLE)

- Desired probability distribution is the one that makes the observed data "most likely," which means that one must seek the value of the parameter vector $\hat{\theta}$ that maximizes the likelihood function $\mathcal{L}(\hat{\theta}|x)$

$$p(x = (x_1, x_2, \ldots, x_n) \mid \theta) = p(x_1|\theta) \, . \, p(x_2|\theta) \ldots p(x_n|\theta)$$

- For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function $ln(\, \mathcal{L}(\hat{\theta}|x) \,)$

$$ln(\, \mathcal{L}(\hat{\theta}|x) \,) = \sum_{x_i} ln \; p(x_i \,|\, \theta)$$

- If log-likelihood is differentiable find $\hat{\theta}$ partial derivatives

Maximum Likelihood Estimator (MLE)

$$\mathcal{L}(\theta \mid x = (x_1, x_2, \ldots, x_n)) = \prod_{x_i} p(x_i \mid \theta) \longrightarrow ln(\, \mathcal{L}(\hat{\theta}|x)\,) = \sum_{x_i} ln\; p(x_i \,|\, \theta)$$

likelihood

log likelihood

- Find maxima by using partial derivatives i.e. $\dfrac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \theta_i} = 0$

Maximum Likelihood Estimator (MLE)

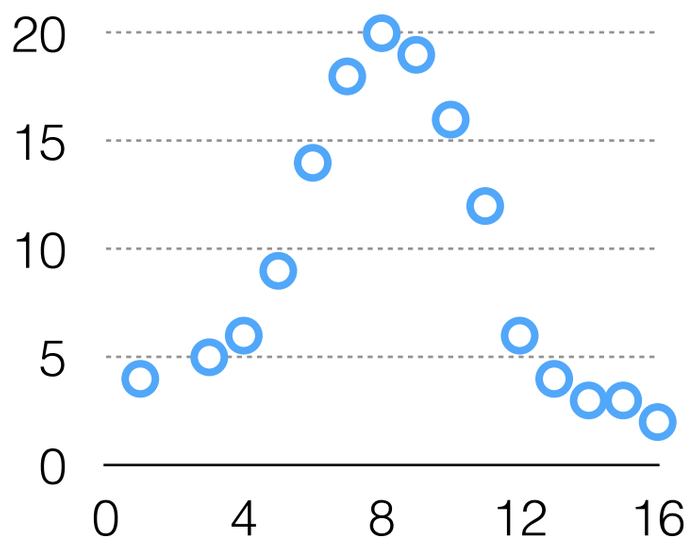
$$\mathcal{L}(\theta \mid x = (x_1, x_2, \ldots, x_n)) = \prod_{x_i} p(x_i \mid \theta) \longrightarrow ln(\, \mathcal{L}(\hat{\theta}|x)\,) = \sum_{x_i} ln\; p(x_i \,|\, \theta)$$

likelihood

log likelihood

- Find maxima by using partial derivatives i.e. $\dfrac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \theta_i} = 0$

- Example :



Maximum Likelihood Estimator (MLE)

$$\mathcal{L}(\theta \mid x = (x_1, x_2, \ldots, x_n)) = \prod_{x_i} p(x_i \mid \theta) \longrightarrow ln(\, \mathcal{L}(\hat{\theta}|x)\,) = \sum_{x_i} ln\; p(x_i \,|\, \theta)$$

likelihood

log likelihood

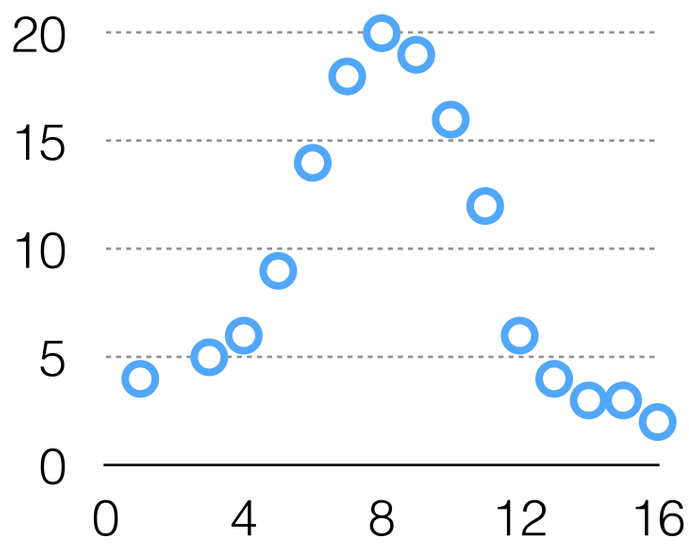
- Find maxima by using partial derivatives i.e. $\dfrac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \theta_i} = 0$

- Example :

$$p(x_i \mid (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

gaussian

generative model



Maximum Likelihood Estimator (MLE)

$$\mathcal{L}(\theta \mid x = (x_1, x_2, \ldots, x_n)) = \prod_{x_i} p(x_i \mid \theta) \longrightarrow ln(\, \mathcal{L}(\hat{\theta}|x)\,) = \sum_{x_i} ln\; p(x_i \,|\, \theta)$$

likelihood

log likelihood

- Find maxima by using partial derivatives i.e. $\dfrac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \theta_i} = 0$

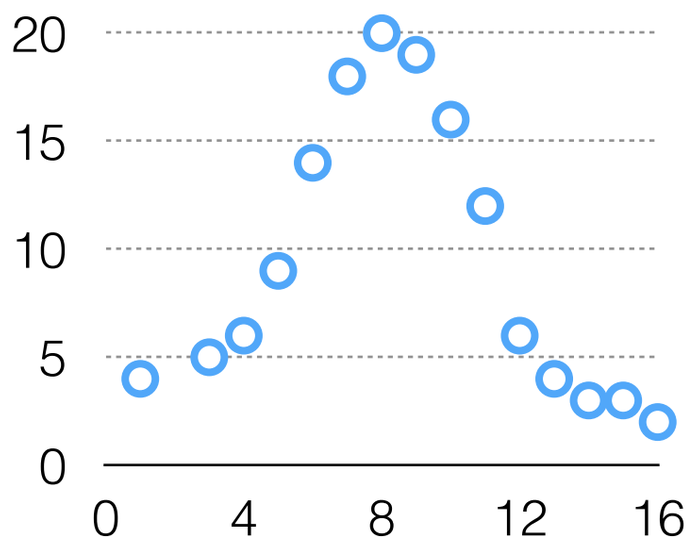
- Example :

$$p(x_i \mid (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

gaussian

generative model

$$\frac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \mu} = 0$$



Maximum Likelihood Estimator (MLE)

$$\mathcal{L}(\theta \mid x = (x_1, x_2, \ldots, x_n)) = \prod_{x_i} p(x_i \mid \theta) \longrightarrow ln(\, \mathcal{L}(\hat{\theta}|x)\,) = \sum_{x_i} ln\; p(x_i \,|\, \theta)$$

likelihood

log likelihood

- Find maxima by using partial derivatives i.e. $\dfrac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \theta_i} = 0$

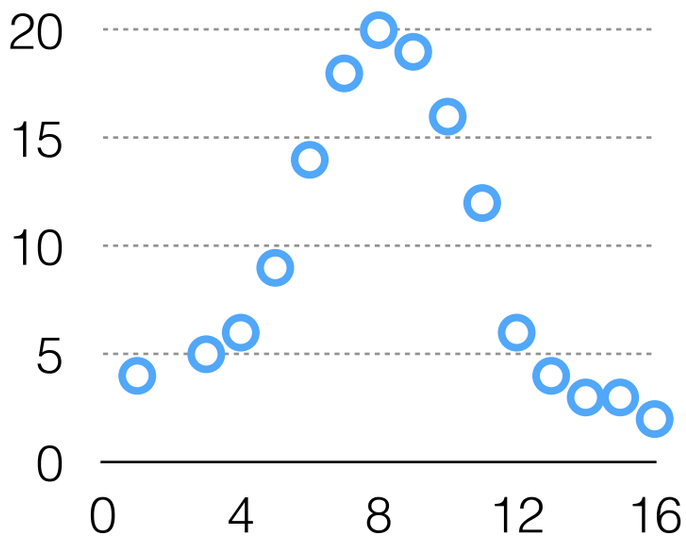
- Example :

$$p(x_i \mid (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

generative model

gaussian

$$\frac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \mu} = 0 \qquad \frac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \sigma} = 0$$



Maximum Likelihood Estimator (MLE)

$$\mathcal{L}(\theta \mid x = (x_1, x_2, \ldots, x_n)) = \prod_{x_i} p(x_i \mid \theta) \longrightarrow ln(\, \mathcal{L}(\hat{\theta}|x)\,) = \sum_{x_i} ln\; p(x_i \,|\, \theta)$$

likelihood

log likelihood

- Find maxima by using partial derivatives i.e. $\dfrac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \theta_i} = 0$

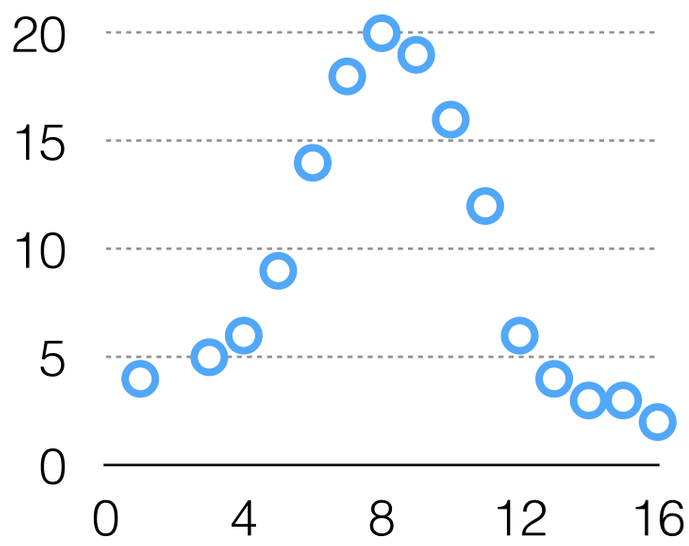
- Example :

$$p(x_i \mid (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

generative model

gaussian

$$\frac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \mu} = 0 \qquad \frac{\partial\, ln\, \mathcal{L}(\theta \mid x)}{\partial\, \sigma} = 0$$



- What are the estimated parameters of the gaussian ?

Find a distribution for Coin Tosses

- A coin E = {H,T} has a probability of p(H) = p, p(T) = 1-p
- Lets say in 3 tosses, what is the probability you get the following sequence:
  - H,T,T p (1-p) (1-p)
  - T,H,T p (1-p) (1-p)
  - What is the probability for getting exactly 1 head ? 3. p (1-p) (1-p)

- What is the probability of getting **k** heads in **n** throws ?

binomial distribution $\binom{n}{k} p^k \cdot (1-p)^{(n-k)}$

prob. of k heads

ways in which k heads are chosen

Multinomial Distribution

- Lets try to find a distribution for text, say a document
- You are given a document D with
  - tf ("game") = 5, tf ("of") = 15, tf ("thrones") = 5, tf ("arya") = 3, tf ("stark") = 4, ….
  - E = {all words in D} = {"game","thrones","stark",….}
- Multinomial distribution best encodes the generative process of text

multinomial distribution

$$\binom{n}{tf_1, tf_2, \ldots, tf_{|E|}} p(x_1)^{tf_1} \ldots . p(x_{|E|})^{tf_{|E|}}$$

term frequency of the word

prob. of the word occurrence

Language Model for a Document

- What are the parameters to be estimated and how can we estimate them ?

total words in Doc

$$\binom{|D|}{tf_1, tf_2, \ldots, tf_{|V|}} \; p(x_1)^{tf_1} \; \ldots . \; p(x_{|V|})^{tf_{|V|}}$$

multinomial distribution

term frequency of the word

Language Model for a Document

- What are the parameters to be estimated and how can we estimate them ?

total words in Doc

observed values

$$\binom{|D|}{tf_1, tf_2, \ldots, tf_{|V|}} p(x_1)^{tf_1} \ldots . p(x_{|V|})^{tf_{|V|}}$$

multinomial distribution

term frequency of the word

Language Model for a Document

- What are the parameters to be estimated and how can we estimate them ?

$$\binom{|D|}{tf_1, tf_2, \ldots, tf_{|V|}} p(x_1)^{tf_1} \ldots . p(x_{|V|})^{tf_{|V|}}$$

total words in Doc

observed values

multinomial distribution

term frequency of the word

parameter to be estimated

Language Model for a Document

- What are the parameters to be estimated and how can we estimate them ?

$$\binom{|D|}{tf_1, tf_2, \ldots, tf_{|V|}} \; p(x_1)^{tf_1} \; \ldots . \; p(x_{|V|})^{tf_{|V|}}$$

total words in Doc

observed values

multinomial distribution

term frequency of the word

parameter to be estimated

- The MLE for each of the word prob. p(x) is the most natural estimate

$$p(x_i) \; = \; \frac{tf_1}{|D|}$$

Statistical Language Models

- Models to describe language generation
- Traditional NLP applications: Assigns a probability value to a sentence
  - Machine Translation — P(**high** snowfall) > P(**large** snowfall)
  - Spelling Correction — P(in the vineyard) > P(in the vinyard)
  - Speech Recognition — P(I saw a van) >> P(eyes awe of an)
  - Question Answering
- Goal: compute the probability of a sentence or sequence of words:
  - P(S) = P(w1,w2,w3,w4,w5...wn)

A model that computes P(S) is called a language model

Statistical Language Models

- How do we compute P(S) ?
  - P(S) = P(w1,w2,w3,w4,w5,…)

chain rule of prob.

P(w1,w2,w3,w4, w5,..) = P(w1 I w2, w3…). P(w2 I w3,..)...

- Unigram Model
  - P(w1 | w2, w3…) = p (w1)
  - P(w1,w2,w3,w4, w5,..) = P(w1). P(w2)…
- Bi-gram Model
  - P(w1,w2,w3,w4, w5,..) = P(w1 | w2). P(w2 | w3)…
- n-gram models : store n-gram dependencies

small model

big model

Language Models in IR

- Unigram Model typically used in IR
  - P(w1,w2,w3,w4, w5,..) = P(w1). P(w2)…
- Build a language model for each document D — { P (w_i | D) }
- Ranking:
  - Query Likelihood: Given a query Q, find the relevance of D (rank acc to P (D|Q))
- KL-Divergence Model:
  - Build language model for the query P (w | Q)
  - Rank acc. to **KL (D || Q)** compares two distributions

Ranking using Query Likelihood

- Query Likelihood: Given a query Q, find the relevance of D (rank acc to P (D|Q))

$$P(D|Q) = \frac{P(Q \mid D).P(D)}{P(Q)}$$

$$P(D|Q) \propto P(Q \mid D).P(D)$$

Ranking using Query Likelihood

- Query Likelihood: Given a query Q, find the relevance of D (rank acc to P (D|Q))

$$P(D|Q) = \frac{P(Q \mid D).P(D)}{P(Q)}$$

constant for all docs

$$P(D|Q) \propto P(Q \mid D).P(D)$$

Ranking using Query Likelihood

- Query Likelihood: Given a query Q, find the relevance of D (rank acc to P (D|Q))

$$P(D|Q) = \frac{P(Q \mid D).P(D)}{P(Q)}$$

Authority of the docs. determined by Page Rank etc.

constant for all docs

$$P(D|Q) \propto P(Q \mid D).P(D)$$

Ranking using Query Likelihood

- Query Likelihood: Given a query Q, find the relevance of D (rank acc to P (D|Q))

Authority of the docs. determined by Page Rank etc.

$$P(D|Q) = \frac{P(Q \mid D).P(D)}{P(Q)}$$

constant for all docs

$$P(D|Q) \propto P(Q \mid D).P(D)$$

Assuming all documents are equally likely

$$P(D|Q) \propto P(Q \mid D)$$

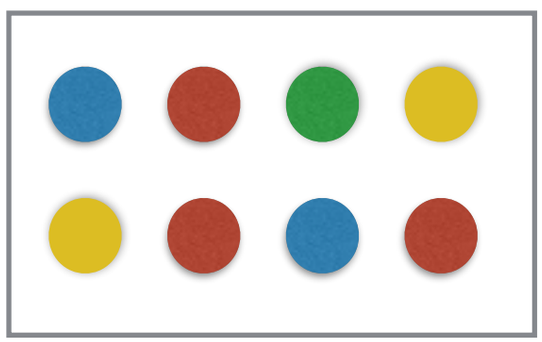
$$P(D|Q) \propto \prod_{w_i \in Q} P(w_i|D)$$

unigram language model

Language Model for a Document

- Unigram Language Model provides a probabilistic model for representing text in Information retrieval

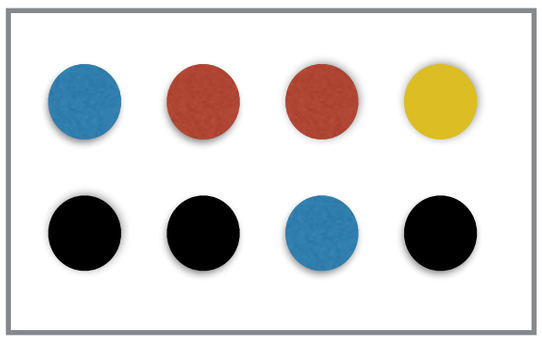
D_1



$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/4$

$p(\bullet | D_1) = 1/8$ $p(\bullet | D_1) = 3/8$

D_2



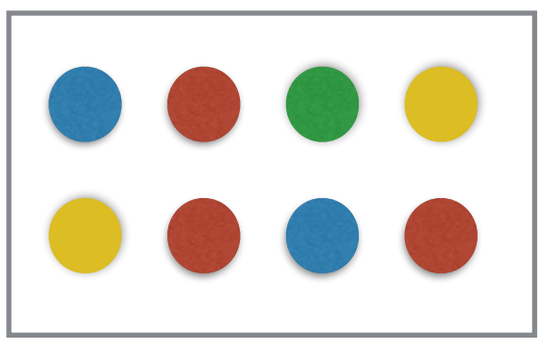
$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/8$

$p(\bullet | D_1) = 3/8$ $p(\bullet | D_1) = 1/4$

Language Model for a Document

- Unigram Language Model provides a probabilistic model for representing text in Information retrieval

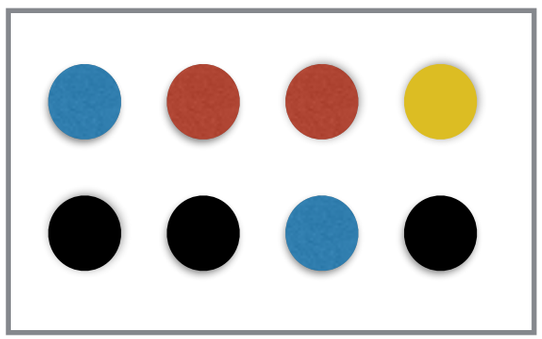
D_1



$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/4$

$p(\bullet | D_1) = 1/8$ $p(\bullet | D_1) = 3/8$

D_2



$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/8$

$p(\bullet | D_1) = 3/8$ $p(\bullet | D_1) = 1/4$

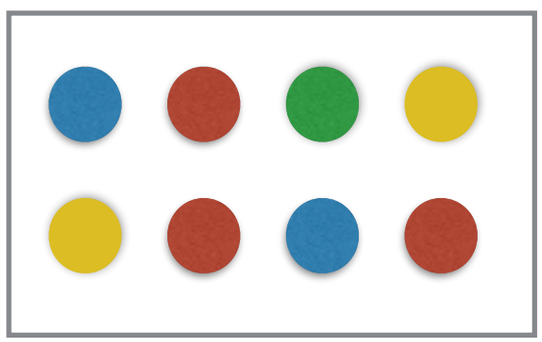
query

$p(\bullet\bullet | D_1) = 3/32$

Language Model for a Document

- Unigram Language Model provides a probabilistic model for representing text in Information retrieval

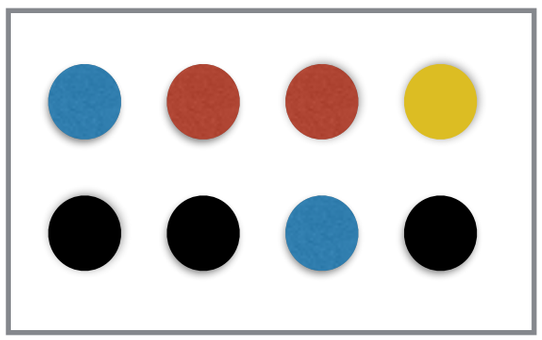
D_1



$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/4$

$p(\bullet | D_1) = 1/8$ $p(\bullet | D_1) = 3/8$

D_2



$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/8$

$p(\bullet | D_1) = 3/8$ $p(\bullet | D_1) = 1/4$

query

$p(\bullet\bullet | D_1) = 3/32$

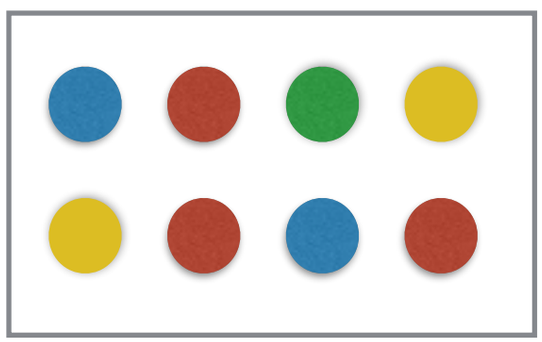
query

$p(\bullet\bullet | D_2) = 1/32$

Language Model for a Document

- Unigram Language Model provides a probabilistic model for representing text in Information retrieval

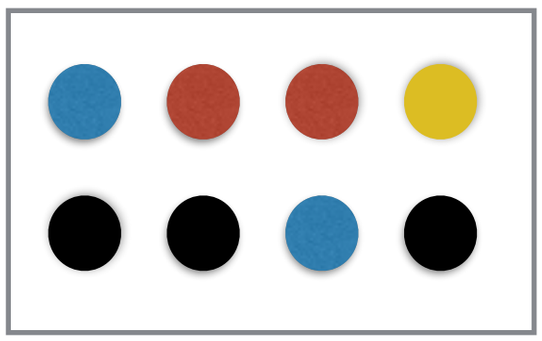
D_1



$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/4$

$p(\bullet | D_1) = 1/8$ $p(\bullet | D_1) = 3/8$

D_2



$p(\bullet | D_1) = 1/4$ $p(\bullet | D_1) = 1/8$

$p(\bullet | D_1) = 3/8$ $p(\bullet | D_1) = 1/4$

query

$p(\bullet\bullet | D_1) = 3/32$

query

$p(\bullet\bullet | D_2) = 1/32$ $D_1 > D_2$

Example - Language Model + MLE

- Unigram Language Model provides a probabilistic model for representing text

Model M_1		Model M_2	
the	0.2	the	0.15
a	0.1	a	0.12
frog	0.01	frog	0.0002
toad	0.01	toad	0.0001
said	0.03	said	0.03
likes	0.02	likes	0.04
that	0.04	that	0.04
dog	0.005	dog	0.01
cat	0.003	cat	0.015
monkey	0.001	monkey	0.002
...	...	...	...

s	frog	said	that	toad	likes	that	dog
M_1	0.01	0.03	0.04	0.01	0.02	0.04	0.005
M_2	0.0002	0.03	0.04	0.0001	0.04	0.04	0.01

$$P(s|M_1) = 0.0000000000048$$
$$P(s|M_2) = 0.00000000000000384$$

- The MLE for each of the word prob. p(x) is the most natural estimate

$$p(x_i) = \frac{tf_1}{|D|}$$

Zero Probability Problem

- What if some of the queried terms are absent in the document ?
- MLE based estimation results in a zero probability for query generation

Model M_1		Model M_2	
the	0.2	the	0.15
a	0.1	a	0.12
frog	0.01	frog	0.0002
toad	0.01	toad	0.0001
said	0.03	said	0.03
likes	0.02	likes	0.04
that	0.04	that	0.04
dog	0.005	dog	0.01
cat	0.003	cat	0.015
monkey	0.001	monkey	0.002
…	…	…	…

- P("frog","ape" | M_1) = 0.01x **0**
- P("frog","ape" | M_2) = 0.00021x **0**

- Need to smooth the probability estimates for terms to avoid zero probabilities
- Take the prob. mass from each term and redistribute among missing terms

Smoothing Methods

- **Jelinek-Mercer Smoothing :** Linear combination of document and corpus statistics to estimate term probabilities

$$P(Q|D) = \prod_{w_i \in Q} \underbrace{\lambda.P(w_i|D)}_{\text{doc. contrib.}} + \underbrace{(1-\lambda).P(w_i|C)}_{\text{corpus contrib.}}$$

- Collection frequency: fraction of occurence of term in the entire collection
- Document frequency: fraction of document occurence of term in the entire collection

Smoothing Methods

- **Jelinek-Mercer Smoothing :** Linear combination of document and corpus statistics to estimate term probabilities

$$P(Q|D) = \prod_{w_i \in Q} \lambda.P(w_i|D) + (1-\lambda).P(w_i|C)$$

doc. contrib.

corpus contrib.

collection freq. or document fréquency

- Collection frequency: fraction of occurence of term in the entire collection
- Document frequency: fraction of document occurence of term in the entire collection

Smoothing Methods

- **Jelinek-Mercer Smoothing :** Linear combination of document and corpus statistics to estimate term probabilities

$$P(Q|D) = \prod_{w_i \in Q} \lambda.P(w_i|D) + (1-\lambda).P(w_i|C)$$

doc. contrib.

corpus contrib.

param. regulates contribution

collection freq. or document fréquency

- Collection frequency: fraction of occurence of term in the entire collection
- Document frequency: fraction of document occurence of term in the entire collection

Smoothing Methods

- Smoothing with **Dirichlet Prior:**

term freq of word in Doc

$$P(Q|D) = \prod_{w_i \in Q} \frac{tf(w_i; D) + \mu.P(w_i|C)}{|D| + \mu}$$

- Takes the corpus distribution as a prior to estimating the prob. for terms
- works well for short queries

Smoothing Methods

- Smoothing with **Dirichlet Prior:**

$$P(Q|D) = \prod_{w_i \in Q} \frac{tf(w_i; D) + \mu.P(w_i|C)}{|D| + \mu}$$

term freq of word in Doc

dirichlet prior

- Takes the corpus distribution as a prior to estimating the prob. for terms
- works well for short queries

Smoothing Methods

- Smoothing with **Dirichlet Prior:**

$$P(Q|D) = \prod_{w_i \in Q} \frac{tf(w_i; D) + \mu.P(w_i|C)}{|D| + \mu}$$

term freq of word in Doc

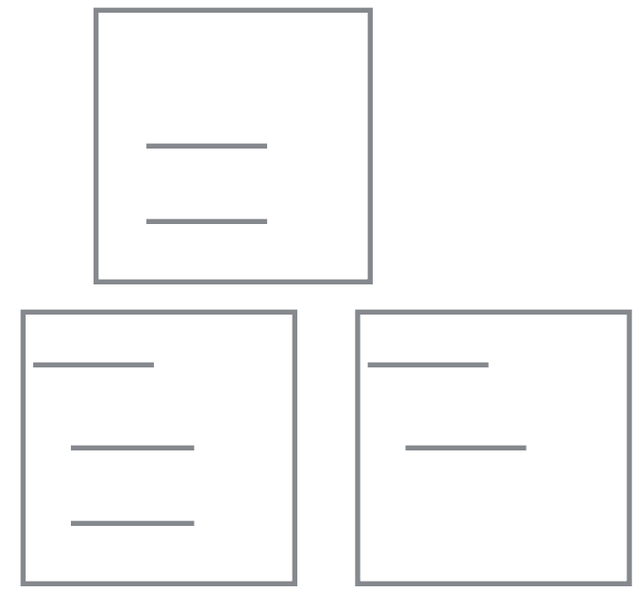
collection freq. or document fréquency

dirichlet prior

- Takes the corpus distribution as a prior to estimating the prob. for terms
- works well for short queries

Temporal Ranking

- Queries with temporal expressions
  - fifa world cup in 1998
- Documents also mention temporal expressions
  - "Clinton was the president in 1990s"
  - "Interstellar was slated for release in july 2014"



$$P(Q|D) = P(Q_{text}|D_{text}).P(Q_{time}|D_{time})$$

- Assume time mentions are independent of text
- Assume temporal expressions are independent of each other

How do we compute $P(Q_{time}|D_{time})$

References and Further Readings

- Information retrieval: (http://www.ir.uwaterloo.ca/book/)
  - Stefan Büttcher, Google Inc. , Charles L. A. Clarke, Univ. of Waterloo, Gordon V. Cormack, Univ. of Waterloo

- Foundations of Information retrieval: Manning, Schutze, Raghavan