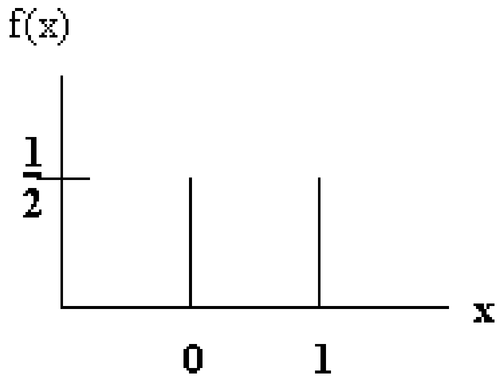
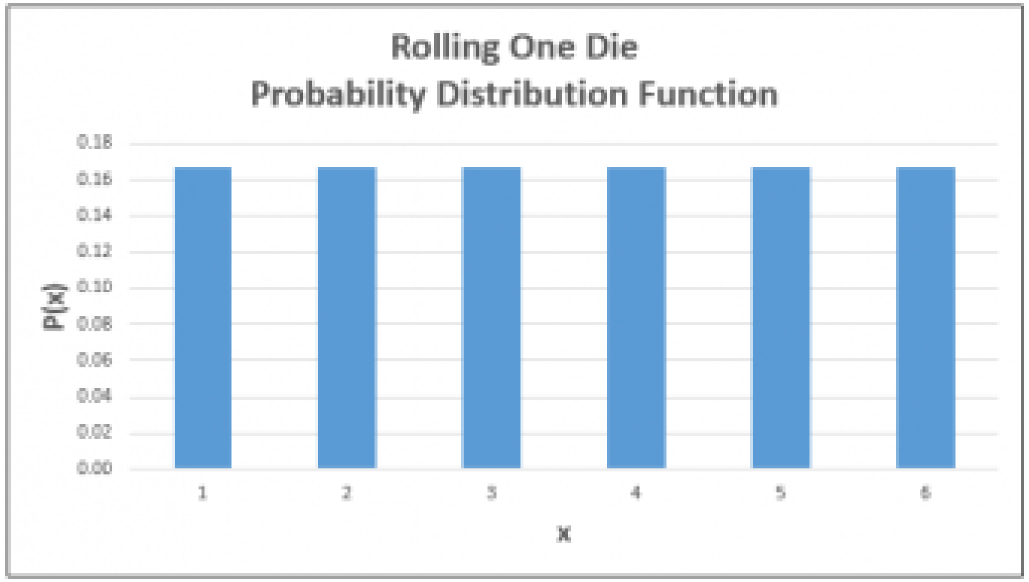
Discussion Week 10

Reminders

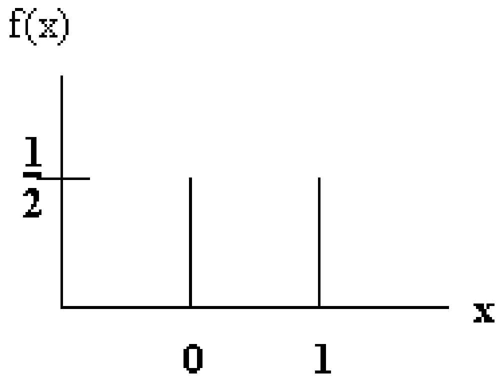
- Project report/code due Sunday December 10 (11:59 PM)
- Final Exam: Wednesday December 13 (11:30 AM – 2:30 PM)
  - Royce Hall 362

Naïve Bayes

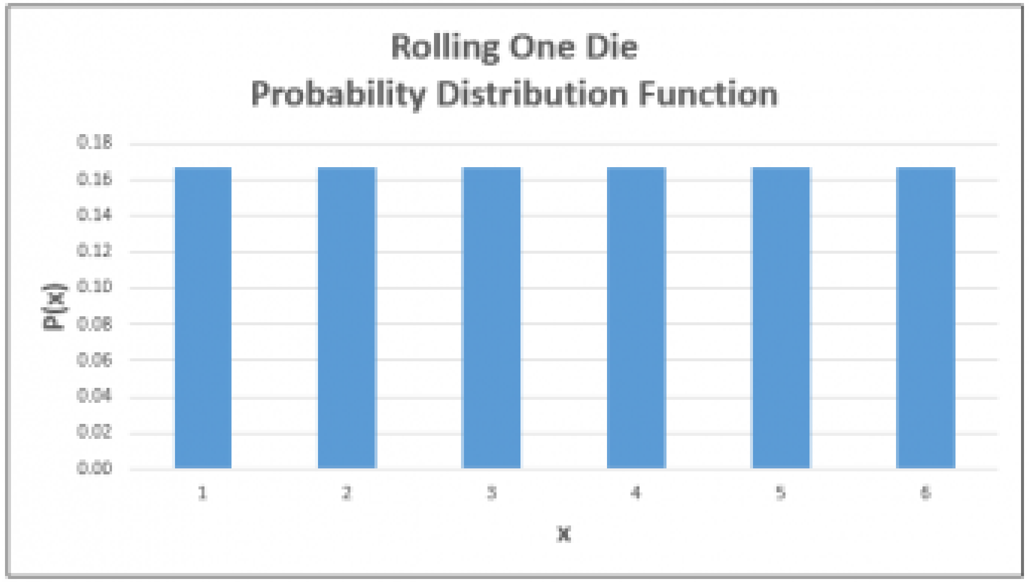

f(x)
1
2
x
0
1


Rolling One Die
Probability Distribution Function
P(x)
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
6
x

Naïve Bayes

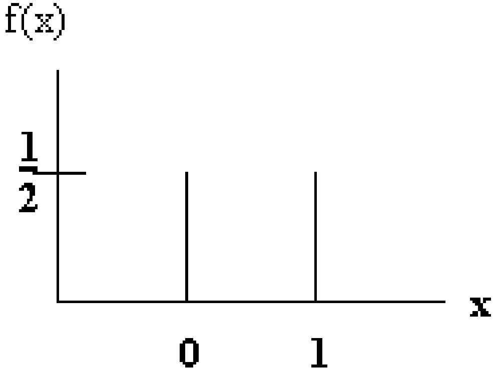


Bernoulli

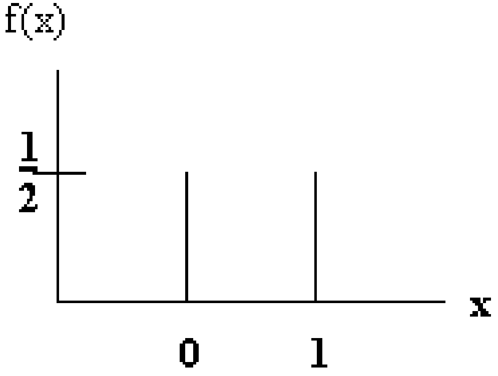


Categorical

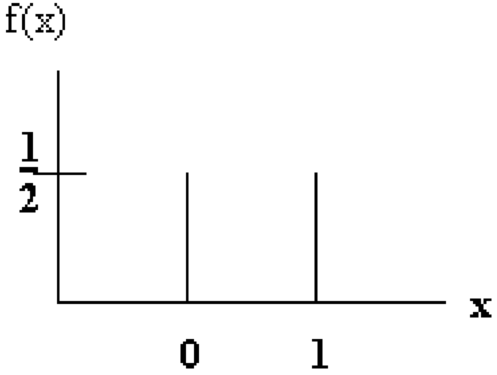
Naïve Bayes

- X ~ Binomial(5, )?

Naïve Bayes

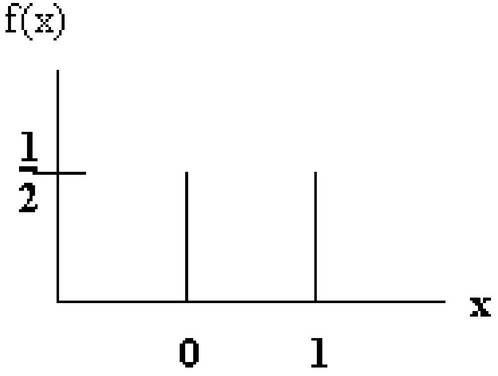
- X ~ Binomial(5, 

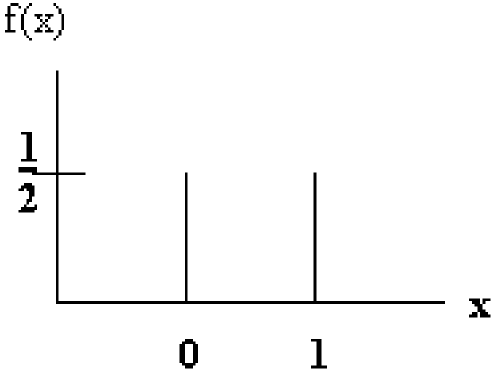
)?

- Pr(<3,2>; 5, 

)?

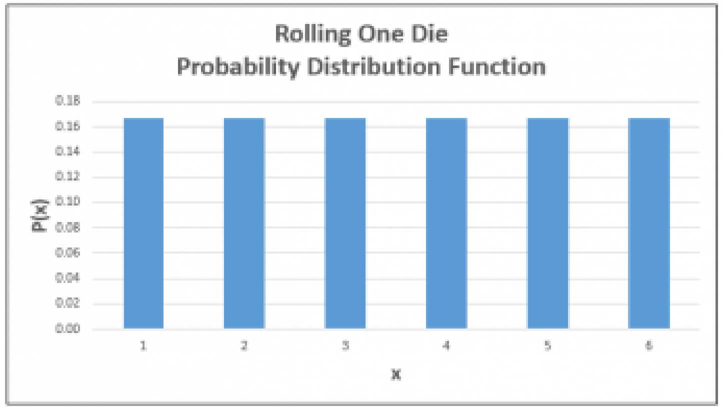
Naïve Bayes

- X ~ Binomial(5, )?

- Pr(<3,2>; 5, )?

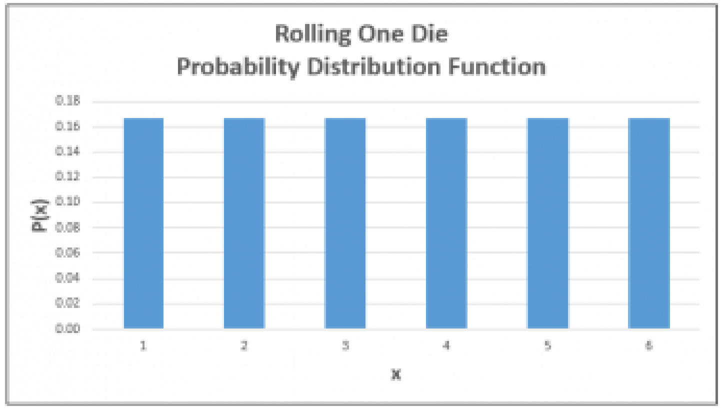
  - 0.3125

Naïve Bayes

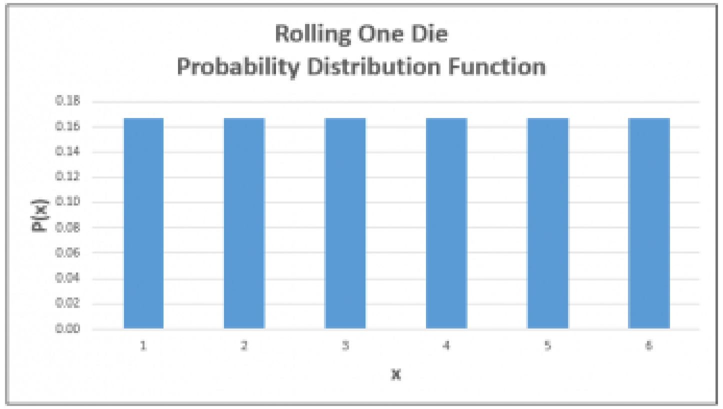
- X ~ Multinomial(5, 

)?

Naïve Bayes

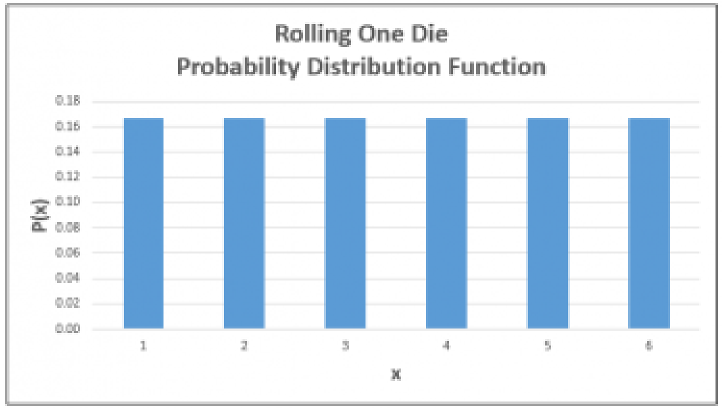
- X ~ Multinomial(5, 

)?

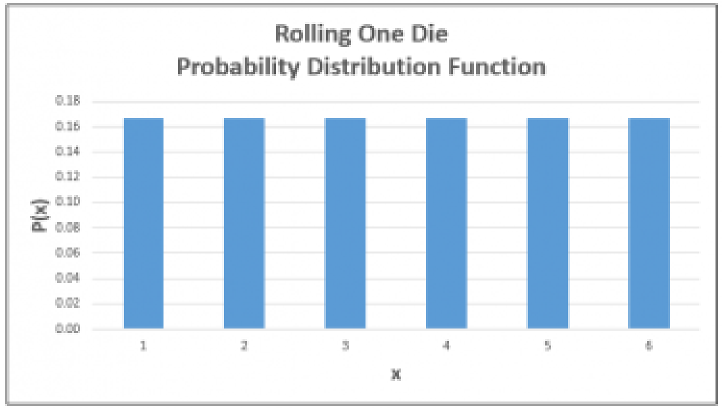
- Pr(<3,2,0,0,0,0>; 5, 

)?

Naïve Bayes

- X ~ Multinomial(5, 

)?

- Pr(<3,2,0,0,0,0>; 5, 

)?
  - 0.001286008

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- P(C|D)?

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- $P(C|D) \propto 0.020 \cdot 0.30 = 0.006$

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- $P(C|D) \propto 0.020 \cdot 0.30 = 0.006$
- $P(B|D) \propto 0.010 \cdot 0.35 = 0.0035$
- $P(A|D) \propto 0.015 \cdot 0.35 = 0.00525$

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- $P(C|D) \propto 0.020 \cdot 0.30 = 0.006$
- $P(B|D) \propto 0.010 \cdot 0.35 = 0.0035$
- $P(A|D) \propto 0.015 \cdot 0.35 = 0.00525$
- $P(A|D) = \frac{0.006}{0.006+0.0035+0.00525} = 0.40677966101$

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- P(C|D) $\propto 0.020 \cdot 0.30 = 0.006$
- P(B|D) $\propto 0.010 \cdot 0.35 = 0.0035$
- P(A|D) $\propto 0.015 \cdot 0.35 = 0.00525$
- Most likely factory if defective?

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- P(C|D) $\propto 0.020 \cdot 0.30 = 0.006$
- P(B|D) $\propto 0.010 \cdot 0.35 = 0.0035$
- P(A|D) $\propto 0.015 \cdot 0.35 = 0.00525$
- Most likely factory if defective?
  - Maximum a posteriori

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- $P(C|D) \propto 0.020 \cdot 0.30 = 0.006$
- $P(B|D) \propto 0.010 \cdot 0.35 = 0.0035$
- $P(A|D) \propto 0.015 \cdot 0.35 = 0.00525$
- Most likely factory if defective?
  - Maximum a posteriori = C
  - Maximum likelihood

Naïve Bayes

Factory	% of total production	Pr of defective lamps
A	0.35 = P(A)	0.015 = P(D|A)
B	0.35 = P(B)	0.010 = P(D|B)
C	0.30 = P(C)	0.020 = P(D|C)

- P(C|D) $\propto 0.020 \cdot 0.30 = 0.006$
- P(B|D) $\propto 0.010 \cdot 0.35 = 0.0035$
- P(A|D) $\propto 0.015 \cdot 0.35 = 0.00525$
- Most likely factory if defective?
  - Maximum a posteriori = C
  - Maximum likelihood = D

Example

- Data:

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

- Vocabulary:

Index	1	2	3	4	5	6
Word	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan

- Learned parameters (with smoothing):

$$\hat{\beta}_{c1} = \frac{5+1}{8+6} = \frac{3}{7} \qquad \hat{\beta}_{j1} = \frac{1+1}{3+6} = \frac{2}{9}$$

$$\hat{\beta}_{c2} = \frac{1+1}{8+6} = \frac{1}{7} \qquad \hat{\beta}_{j2} = \frac{0+1}{3+6} = \frac{1}{9}$$

$$\hat{\beta}_{c3} = \frac{1+1}{8+6} = \frac{1}{7} \qquad \hat{\beta}_{j3} = \frac{0+1}{3+6} = \frac{1}{9}$$

$$\hat{\beta}_{c4} = \frac{1+1}{8+6} = \frac{1}{7} \qquad \hat{\beta}_{j4} = \frac{0+1}{3+6} = \frac{1}{9}$$

$$\hat{\beta}_{c5} = \frac{0+1}{8+6} = \frac{1}{14} \qquad \hat{\beta}_{j5} = \frac{1+1}{3+6} = \frac{1}{9}$$

$$\hat{\beta}_{c6} = \frac{0+1}{8+6} = \frac{1}{14} \qquad \hat{\beta}_{j6} = \frac{1+1}{3+6} = \frac{2}{9}$$

$$\hat{\pi}_c = \frac{3}{4}$$

$$\hat{\pi}_j = \frac{1}{4}$$

Maximum Likelihood Estimation

- Since objects are assumed to be generated independently, for a data set D = {x_1, …, x_n}, we have,

$$P(D) = \prod_i P(x_i) = \prod_i \sum_j w_j f_j(x_i)$$

$$\Rightarrow logP(D) = \sum_i logP(x_i) = \sum_i log \sum_j w_j f_j(x_i)$$

- Task: Find a set *C* of *k* probabilistic clusters s.t. *P*(*D*) is maximized

Gaussian Mixture Model

- Generative model
  - For each object:
    - Pick its cluster, i.e., a distribution component: $Z \sim Multinoulli(w_1, \ldots, w_k)$
    - Sample a value from the selected distribution: $X|Z \sim N(\mu_Z, \sigma_Z^2)$
- Overall likelihood function
  - $L(D|\,\theta) = \prod_i \sum_j w_j p(x_i|\mu_j, \sigma_j^2)$

  s.t. $\sum_j w_j = 1$ and $w_j \geq 0$

Multinomial Mixture Model

- For documents with bag-of-words representation
  - $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \dots, x_{dN})$, x_{dn} is the number of words for nth word in the vocabulary
- Generative model
  - For each document
    - Sample its cluster label $z \sim Multinoulli(\boldsymbol{\pi})$
      - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$, π_k is the proportion of kth cluster
      - $p(z = k) = \pi_k$
    - Sample its word vector $\boldsymbol{x}_d \sim multinomial(\boldsymbol{\beta}_z)$
      - $\boldsymbol{\beta}_z = (\beta_{z1}, \beta_{z2}, \dots, \beta_{zN})$, β_{zn} is the parameter associate with nth word in the vocabulary
      - $p(\boldsymbol{x}_d | z = k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}} \propto \prod_n \beta_{kn}^{x_{dn}}$

Mixture of Unigrams

- For documents represented by a sequence of words
  - $\boldsymbol{w}_d = (w_{d1}, w_{d2}, \dots, w_{dN_d})$, N_d is the length of document d, w_{dn} is the word at the nth position of the document
- Generative model
  - For each document
    - Sample its cluster label $z \sim Multinoulli(\boldsymbol{\pi})$
      - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$, π_k is the proportion of kth cluster
      - $p(z = k) = \pi_k$
    - For each word in the sequence
      - Sample the word $w_{dn} \sim Multinoulli(\boldsymbol{\beta}_z)$
      - $p(w_{dn} | z = k) = \beta_{kw_{dn}}$

Question

- Are multinomial mixture model and mixture of unigrams model equivalent? Why?

Likelihood Function

- For a set of M documents

$$L = \prod_d p(\boldsymbol{x}_d) = \prod_d \sum_k p(\boldsymbol{x}_d, z = k)$$

$$= \prod_d \sum_k p(\boldsymbol{x}_d | z = k) p(z = k)$$

$$\propto \prod_d \sum_k p(z = k) \prod_n \beta_{kn}^{x_{dn}}$$

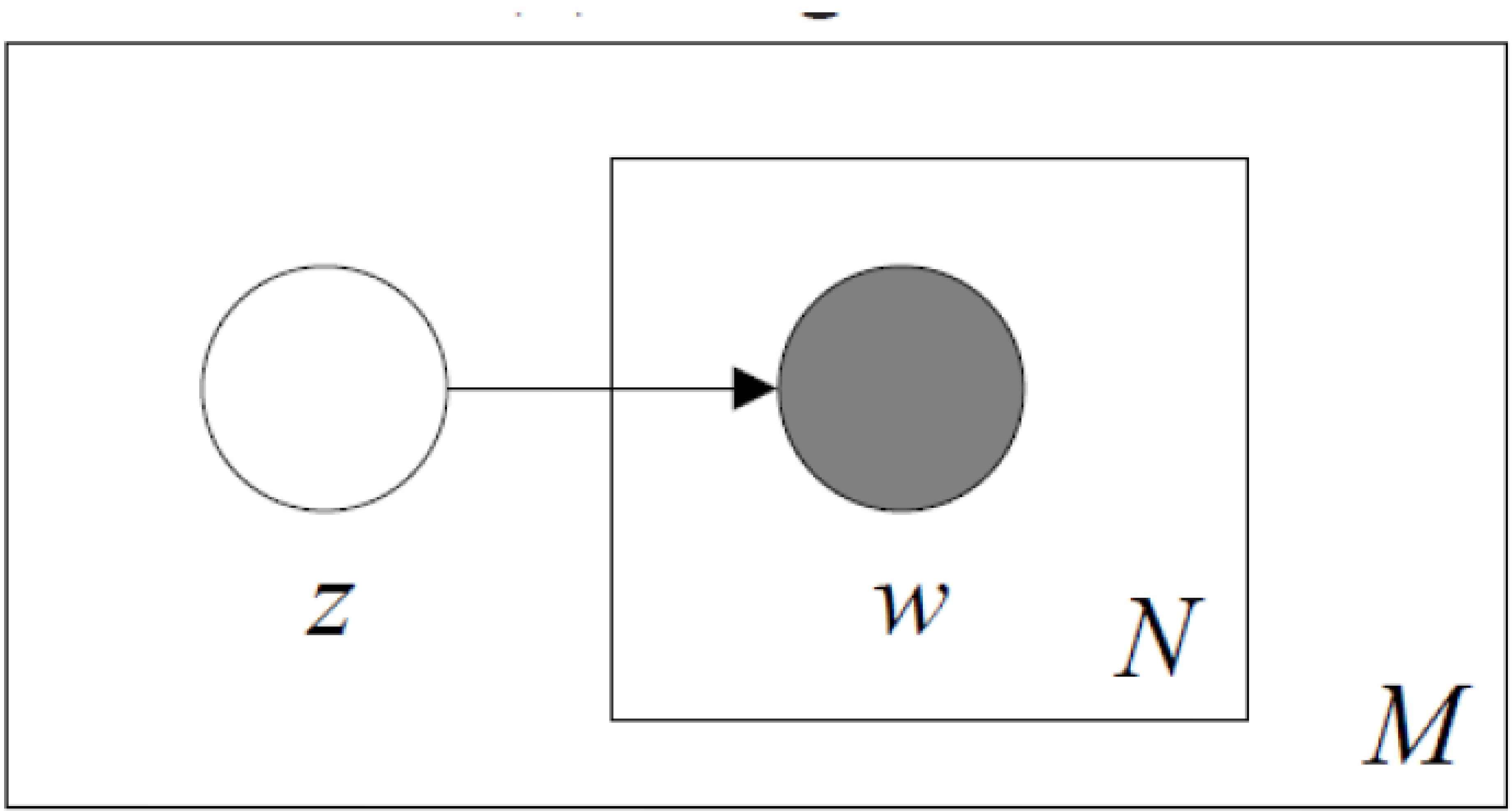
Likelihood Function

- For a set of M documents

$$L = \prod_d p(\boldsymbol{w}_d) = \prod_d \sum_k p(\boldsymbol{w}_d, z = k)$$

$$= \prod_d \sum_k p(\boldsymbol{w}_d | z = k) p(z = k)$$

$$= \prod_d \sum_k p(z = k) \prod_n \beta_{k w_{dn}}$$


z
w
N
M

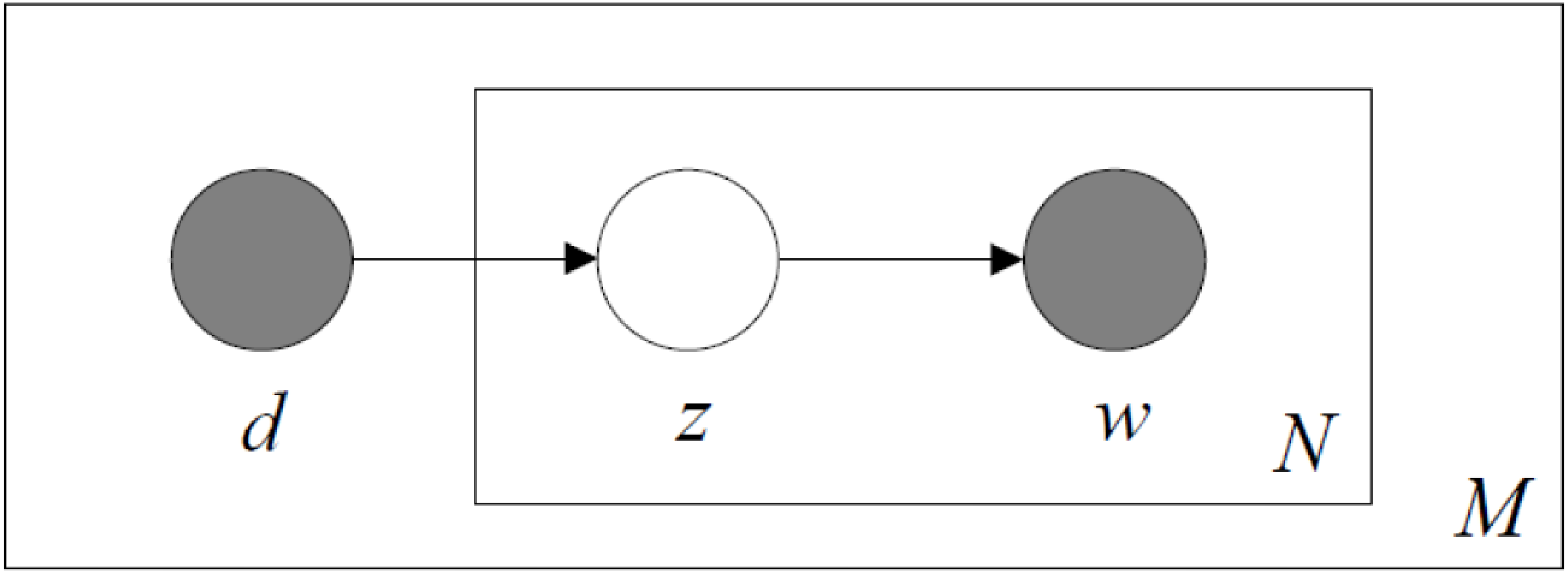
Generative Model for pLSA

- Describe how a document is generated probabilistically
  - For each position in d, $n = 1, \ldots, N_d$
    - Generate the topic for the position as $z_n \sim Multinoulli(\boldsymbol{\theta}_d), i.e., p(z_n = k) = \theta_{dk}$

      (Note, 1 trial multinomial, i.e., categorical distribution)
    - Generate the word for the position as

      $w_n \sim Multinoulli(\boldsymbol{\beta}_{z_n}), i.e., p(w_n = w) = \beta_{z_n w}$

Graphical Model



Note: Sometimes, people add parameters such as θ and β into the graphical model

- Two documents, two topics
  - Vocabulary: {data, mining, frequent, pattern, web, information, retrieval}
  - At some iteration of EM algorithm, E-step

word (w)	word count in Document 1 ($c(w, d_1)$)	$p(z=1\|w, d_1)$
data	5	0.8
mining	4	0.8
frequent	3	0.6
pattern	2	0.8
web	2	0.5
information	1	0.2

word (w)	word count in Document 2 ($c(w, d_2)$)	$p(z=1\|w, d_2)$
information	5	0.2
retrieval	4	0.2
web	3	0.1
mining	3	0.5
frequent	2	0.6
data	2	0.5

- M-step

$$\beta_{11} = \frac{0.8 * 5 + 0.5 * 2}{11.8 + 5.8} = 5/17.6$$

$$\beta_{12} = \frac{0.8 * 4 + 0.5 * 3}{11.8 + 5.8} = 4.7/17.6$$

$$\beta_{13} = 3/17.6$$

$$\beta_{14} = 1.6/17.6$$

$$\beta_{15} = 1.3/17.6$$

$$\beta_{16} = 1.2/17.6$$

$$\beta_{17} = 0.8/17.6$$

$$\theta_{11} = \frac{11.8}{17}$$

$$\theta_{12} = \frac{5.2}{17}$$

- Q2: In pLSA, For the same word in different positions in a document, do they have the same conditional probability $p(z|w,d)$?

- Q2: In pLSA, For the same word in different positions in a document, do they have the same conditional probability $p(z|w,d)$?

word (w)	word count in Document 1 ($c(w,d_1)$)	$p(z=1\|w,d_1)$
data	5	0.8
mining	4	0.8
frequent	3	0.6
pattern	2	0.8
web	2	0.5
information	1	0.2