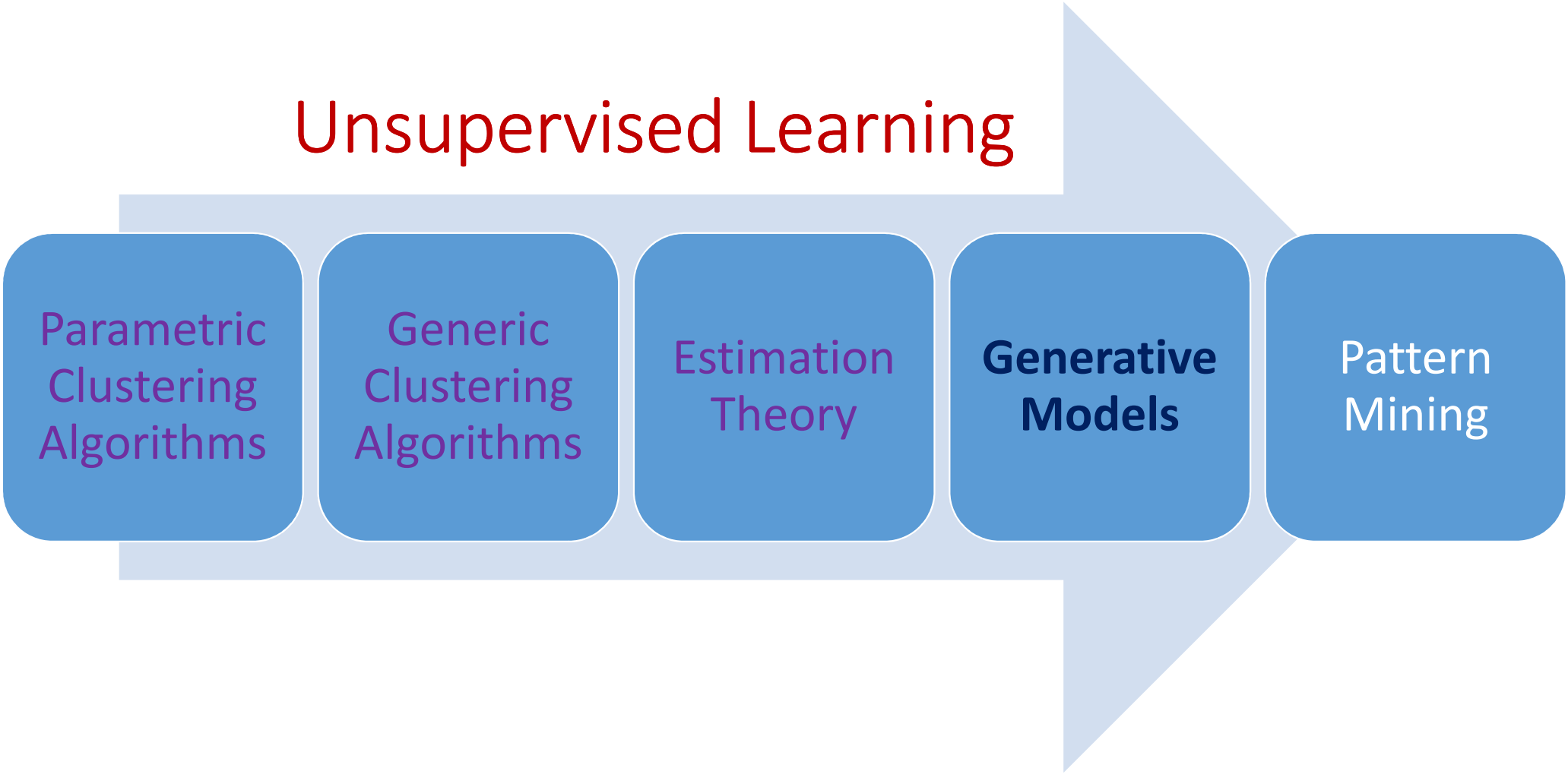
Probabilistic Latent Semantic Analysis

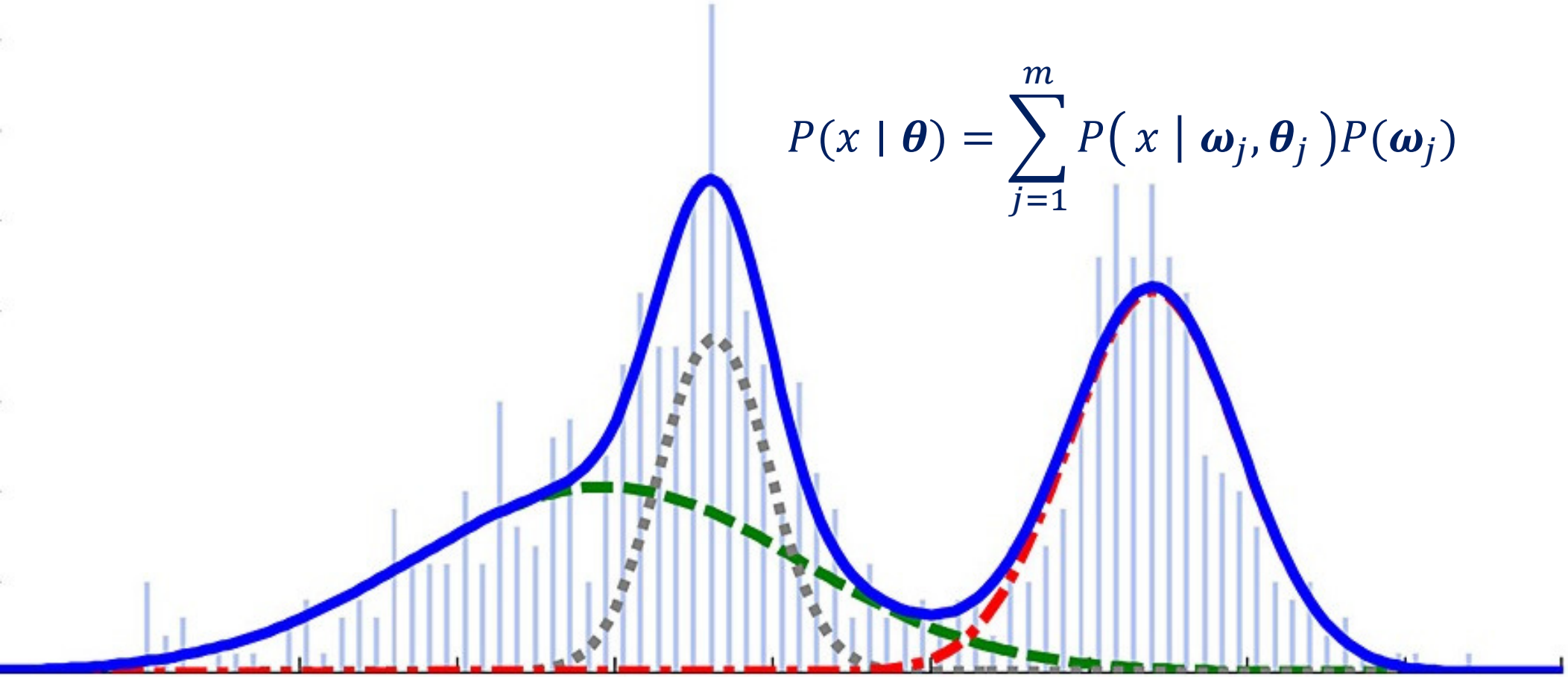
Prithwijit Guha
Dept. of EEE, IIT Guwahati

Unsupervised Learning



Gaussian Mixture Models (GMM)

$$P(x \mid \boldsymbol{\theta}) = \sum_{j=1}^{m} P\left(x \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j\right) P(\boldsymbol{\omega}_j)$$



$$\boldsymbol{D} = \{x_1, x_2, \ldots x_i, \ldots x_{n-1}, x_n\}$$

GMM: Mean & Variance Update

$$\hat{\mu}_r^{(t+1)} = \frac{\sum_{i=1}^{n} x_i P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)}\right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)}\right)}$$

$$\hat{v}_r^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{x_i - \mu_r^{(t)}\right\}^2 P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)}\right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)}\right)}$$

$x \in \mathbb{R}^1$

GMM: Mean & Covariance Update

$$\widehat{\boldsymbol{\mu}}_r^{(t+1)} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}\right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}\right)}$$

$$\widehat{\boldsymbol{C}}_r^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{\boldsymbol{x}_i - \boldsymbol{\mu}_r^{(t)}\right\}\left\{\boldsymbol{x}_i - \boldsymbol{\mu}_r^{(t)}\right\}^T P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}\right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)}\right)}$$

$\boldsymbol{x} \in \mathbb{R}^d$

Expectation-Maximization (EM) Algorithm

Log-Likelihood: $\boldsymbol{L(\theta \mid X)}$

Parameter Vector

Observed Data

Log-Likelihood: $\boldsymbol{L_c(\theta \mid X, Z)}$

Hidden Variables

Expectation-Maximization (EM) Algorithm

E-Step: $\boldsymbol{Q}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = \mathrm{E}\left[\boldsymbol{L}_c(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)}\right]$

M-Step: $\boldsymbol{\theta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\theta}}\, \boldsymbol{Q}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)$

Chennai: A video of government workers in Puducherry hurriedly throwing the body of a COVID-19 positive man into a pit has caused massive outrage, prompting the administration to order a probe into the incident.

COVID

The COVID pandemic has hit the world and the Vedanta Group business. It has incurred losses in oil, gas and mining sectors.

Business

Yuvraj Singh was diagnosed with a cancerous tumor in his left lung following India's World Cup triumph in 2011. He had scored 362 runs and claimed 15 wickets in the tournament and was bestowed with the Player of the Tournament award in the end

Bollywood

Sonu Sood, known as a Bollywood actor, has emerged as the superhero during the COVID-19 pandemic. For thousands of migrants, he is the man who helped them at a time when they were gripped with fear and were walking an uncertain path with no support or even an assurance that all would be well.

Health

Toyota Kirloskar Motor (TKM) has announced one or two percent price increase in India for Toyota Glanza, Yaris, Innova Crysta, and the Fortuner.

Migrant Crisis

Facebook will invest Rs 43,574 crore in Jio Platforms, a unit of Reliance Industries Ltd (RIL), for a 9.99% stake, an allcash deal that will help the oil-toretail conglomerate reduce debt and strengthen the social media company's presence in its largest market, especially for its WhatsApp unit.

Cricket

Word-Document Co-occurrence Matrix

$\mathbb{N} =$

	d_1	d_2	…	d_m	…	d_{M-1}	d_M
w_1							
w_2							
⋮							
w_n				$\eta(d_m, w_n)$			
⋮							
w_{N-1}							
w_N							

Articles, Topics & Words

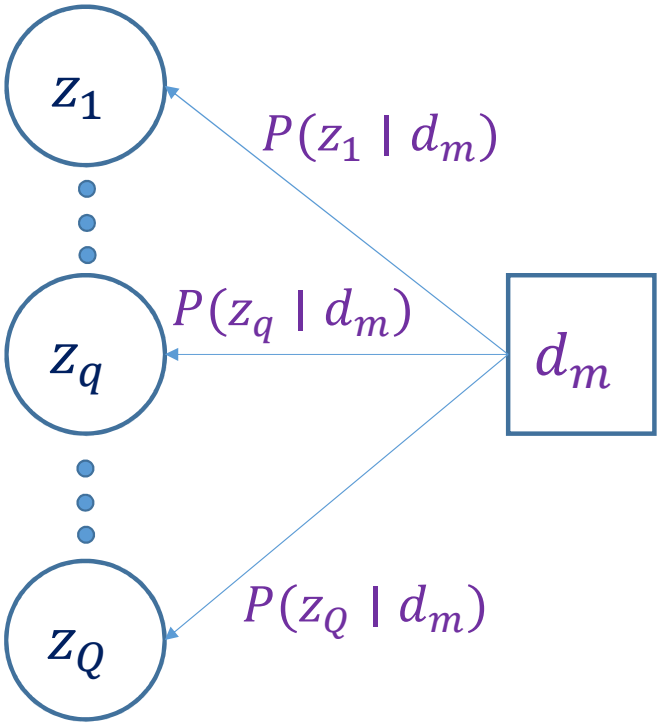
Documents: $d_m; m = 1, \ldots M$

Words: $w_n; n = 1, \ldots N$

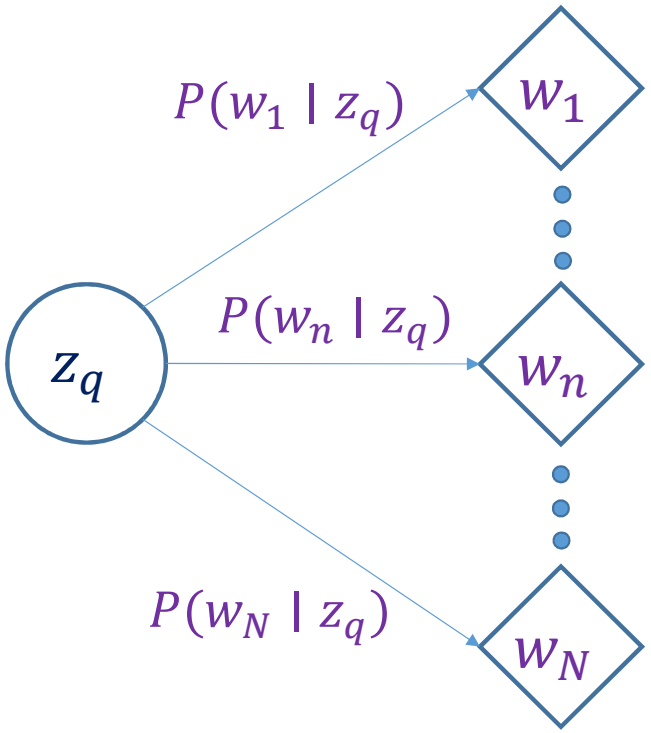
Topics: $z_q; q = 1, \ldots Q$

$Q \ll M$ $\qquad$ $Q \ll N$

Articles, Topics & Words

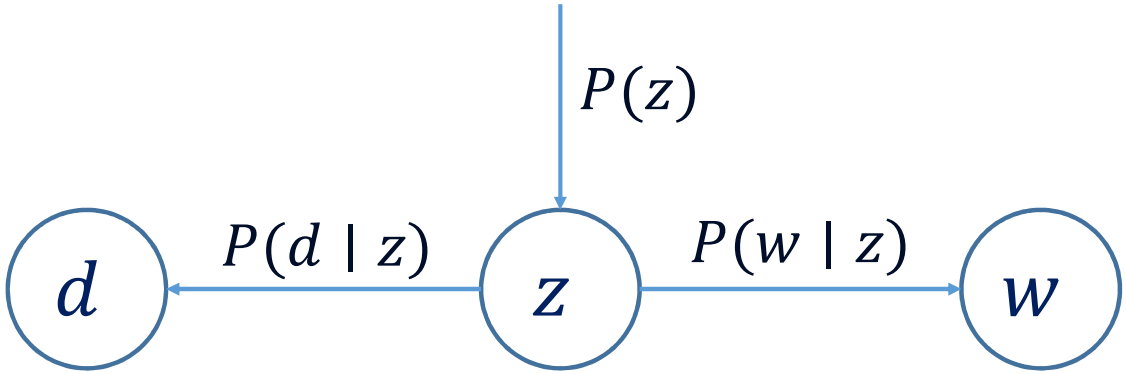


Document as a
Mixture of Topics



Topics Identified
by Words

Aspect Model



$$P(d,w) = \sum_z P(z)P(d \mid z)P(w \mid z)$$

Log-Likelihood Function

$$\boldsymbol{L(z)} = \sum_{m=1}^{M} \sum_{n=1}^{N} \eta(d_m, w_n) \log P(d_m, w_n)$$

E-STEP

Estimating $\boldsymbol{P(z \mid d, w)}$

Document

Word

$$P(z_q \mid d_m, w_n) = \frac{P(z_q)P(d_m \mid z_q)P(w_n \mid z_q)}{\sum_{r=1}^{Q} P(z_r)P(d_m \mid z_r)P(w_n \mid z_r)}$$

Topic

Probability of Occurrence of Topic z_q for Given Word w_n in Document d_m

M-STEP

Estimating $\boldsymbol{P(w \mid z)}$

$$P(w_n \mid z_q) = \frac{\sum_{m=1}^{M} \eta(d_m, w_n) P(z_q \mid d_m, w_n)}{\sum_{m=1}^{M} \sum_{i=1}^{N} \eta(d_m, w_i) P(z_q \mid d_m, w_i)}$$

Word: w_n | Topic: z_q

Probability of Occurrence of Word w_n for Given Topic z_q

Can be Used to Understand the Words that Describe a Topic

M-STEP

Estimating $\boldsymbol{P(d \mid z)}$

$$P(d_m \mid z_q) = \frac{\sum_{n=1}^{N} \eta(d_m, w_n) P(z_q \mid d_m, w_n)}{\sum_{j=1}^{M} \sum_{n=1}^{N} \eta(d_j, w_n) P(z_q \mid d_j, w_n)}$$

Topic

Document

Probability of Occurrence of Word w_n for Given Topic z_q

Can be Used to Identify the Documents that Subscribe to a Topic

M-STEP

Estimating $\boldsymbol{P(z)}$

Topic

$$P(z_q) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\eta(d_m, w_n)P(z_q \mid d_m, w_n)}{\sum_{m=1}^{M}\sum_{n=1}^{N}\eta(d_m, w_n)}$$

The Distribution of Topics

Estimating $\boldsymbol{P(z|d)}$

$$P(z_q \mid d_m) = \sum_{n=1}^{N} P(z_q \mid d_m, w_n)$$

Topic

Document

The Distribution of Topics in a Document

Estimating $\boldsymbol{P(z|w)}$

$$P(z_q \mid w_n) = \sum_{m=1}^{M} P(z_q \mid d_m, w_n)$$

Topic

word

The Distribution of Topics with respect to a Word

Summary

- EM Algorithms
- Aspect Model of PLSA
- Topic Distribution as Dimensionality Reduction
- E-M Equations of PLSA
- Interpreting Different Distributions


भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Thank You