Text Categorization: Generative Probabilistic Models

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Overview

- What is text categorization?
- Why text categorization?
- How to do text categorization?
  - **Generative probabilistic models**
  - Discriminative approaches
- How to evaluate categorization results?

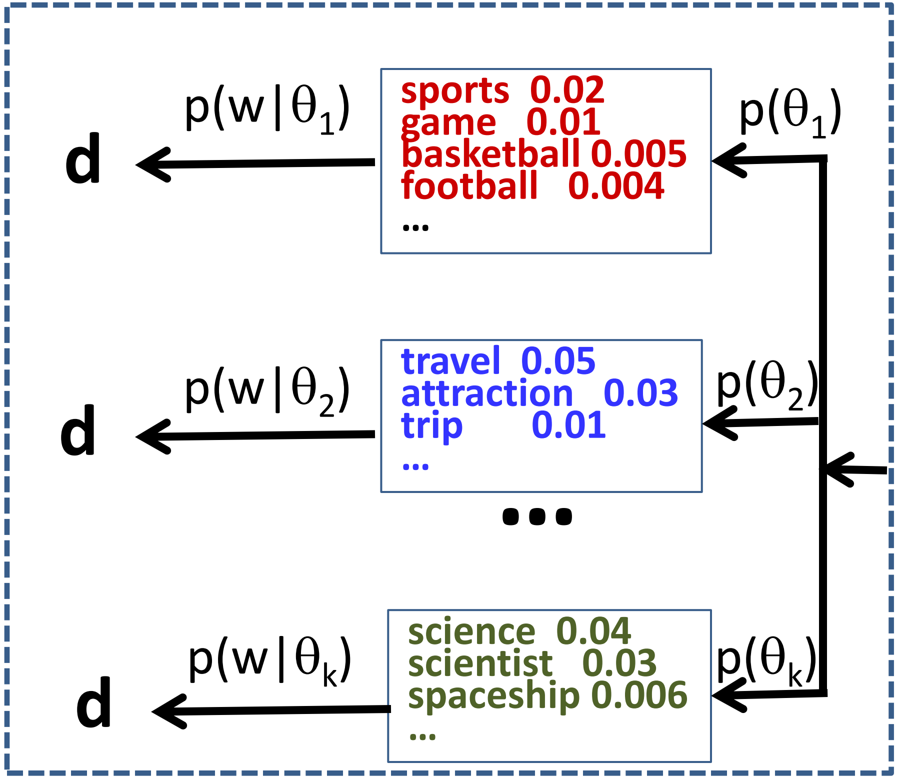
Document Clustering Revisited

Which cluster does d belong to? ➔ Which θ_i was used to generate d?

$\mathbf{d=x_1x_2\ldots x_L}$ where $x_i \in V$

$$\begin{aligned}\text{cluster}(d) &= \arg\max_i p(\theta_i \mid d) \\ &= \arg\max_i p(d \mid \theta_i) p(\theta_i) \\ &= \arg\max_i \prod_{j=1}^{L} p(x_j \mid \theta_i) p(\theta_i) \\ &= \arg\max_i \prod_{w \in V} p(w \mid \theta_i)^{c(w,d)} p(\theta_i)\end{aligned}$$

$$\begin{aligned}p(\theta_i \mid d) &= \frac{p(d \mid \theta_i) p(\theta_i)}{p(d)} \\ &= \frac{p(d \mid \theta_i) p(\theta_i)}{\sum_{j=1}^{k} p(d \mid \theta_j) p(\theta_j)}\end{aligned}$$



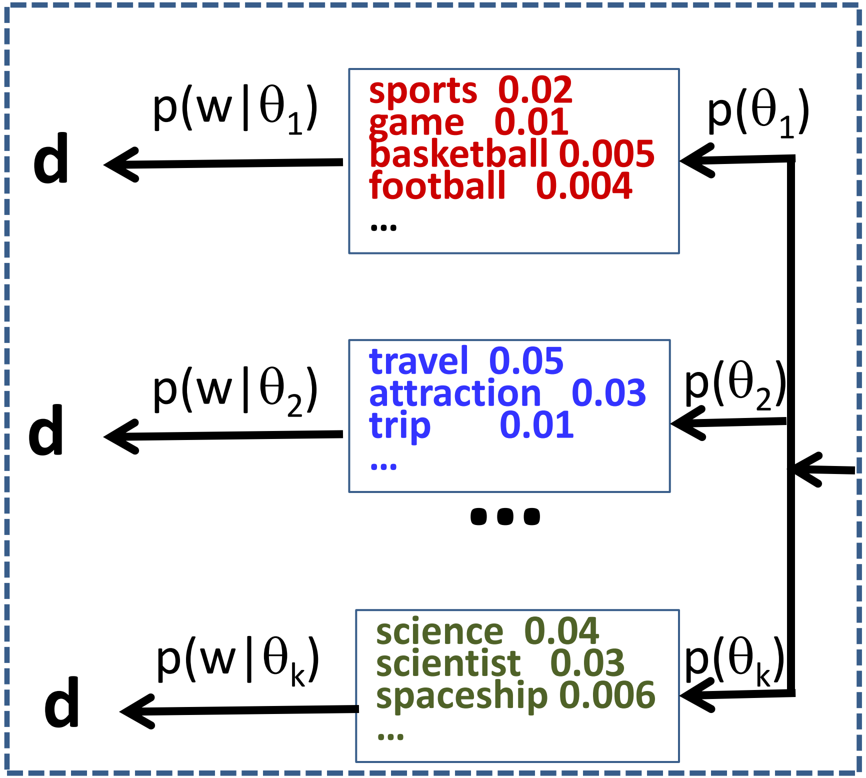
Text Categorization with Naïve Bayes Classifier

$\mathbf{d=x_1x_2\ldots x_L}$ where $x_i \in V$

IF θ_i represents category i accurately, then…

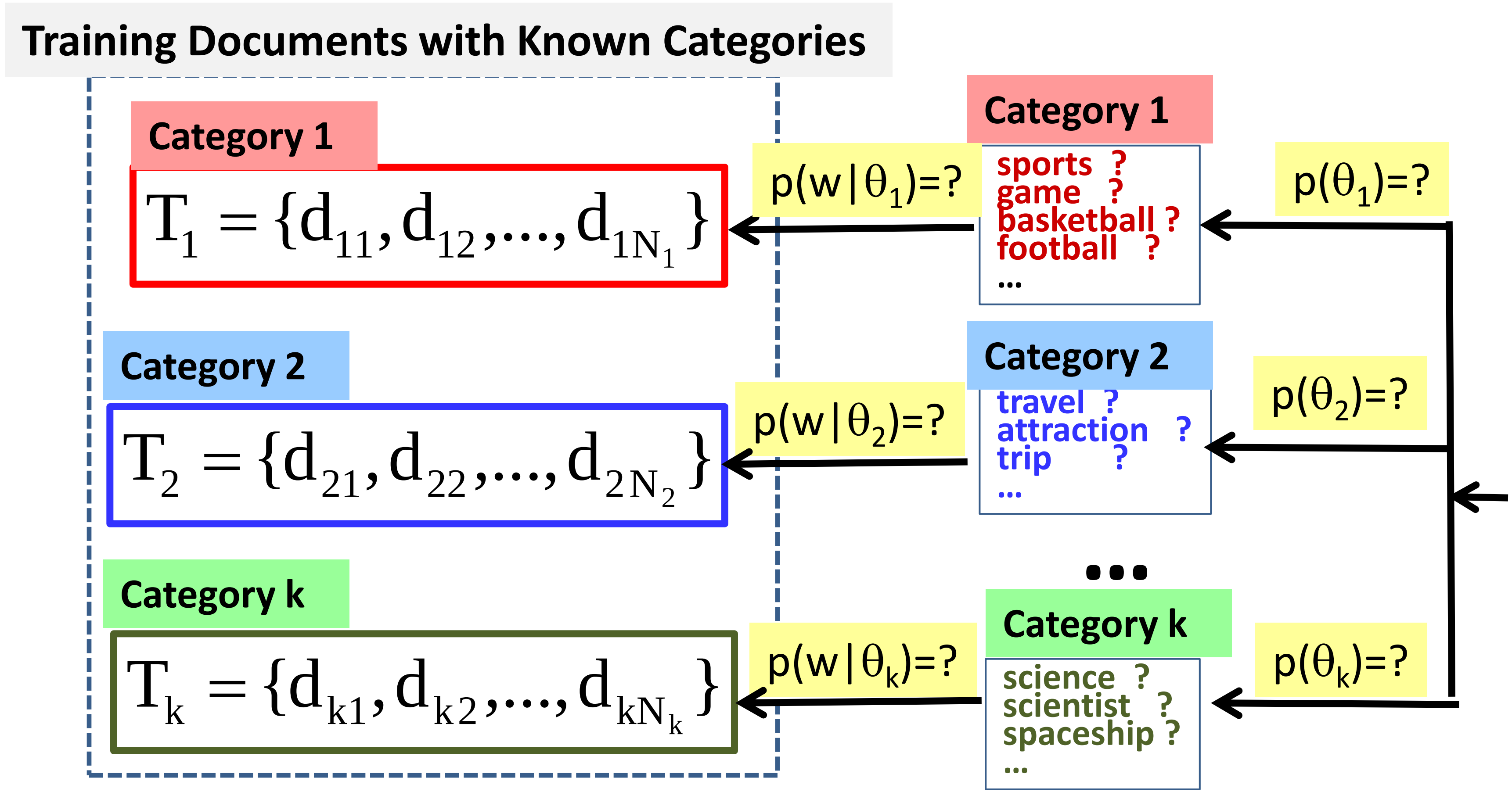
How can we make this happen?

$$\text{category}(d) = \arg\max_i p(\theta_i \mid d)$$
$$= \arg\max_i p(d \mid \theta_i) p(\theta_i)$$
$$= \arg\max_i \prod_{w \in V} p(w \mid \theta_i)^{c(w,d)} p(\theta_i)$$

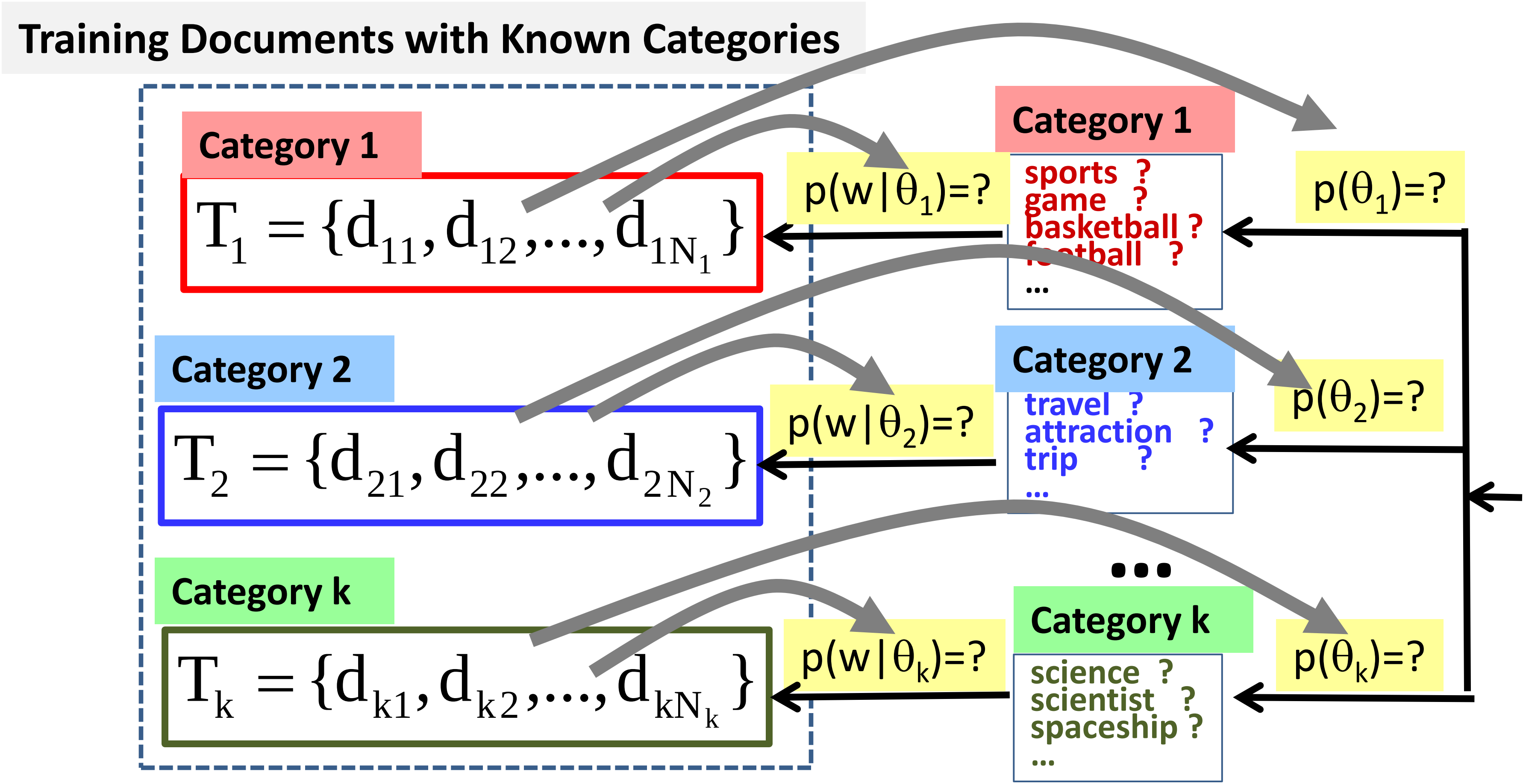


$$\boxed{\text{category}(d) = \arg\max_i \log p(\theta_i) + \sum_{w \in V} c(w,d) \log p(w \mid \theta_i)}$$

Learn from the Training Data



How to Estimate $p(w|\theta_i)$ and $p(\theta_i)$



Naïve Bayes Classifier: $p(\theta_i)$=? and $p(w|\theta_i)$=?

Category 1

$$T_1 = \{d_{11}, d_{12}, \ldots, d_{1N_1}\}$$

Category 2

$$T_2 = \{d_{21}, d_{22}, \ldots, d_{2N_2}\}$$

Category k

$$T_k = \{d_{k1}, d_{k2}, \ldots, d_{kN_k}\}$$

Which category is most popular?

$$p(\theta_i) = \frac{N_i}{\sum_{j=1}^{k} N_j} \propto |T_i|$$

$$p(w|\theta_i) = \frac{\sum_{j=1}^{N_i} c(w, d_{ij})}{\sum_{w' \in V} \sum_{j=1}^{N_i} c(w', d_{ij})} \propto c(w, T_i)$$

Which word is most frequent in category i?

What are the constraints on $p(\theta_i)$ and $p(w|\theta_i)$?

Smoothing in Naïve Bayes

- Why smoothing?
  - Address data sparseness (training data is small ➔ zero prob.)
  - Incorporate prior knowledge
  - Achieve discriminative weighting (i.e., IDF weighting)
- How?

$$p(\theta_i) = \frac{N_i + \delta}{\sum_{j=1}^{k} N_j + k\delta} \qquad \delta \geq 0$$

What if $\delta \to \infty$?

$p(w|\theta_B)$: background LM

$$p(w|\theta_i) = \frac{\sum_{j=1}^{N_i} c(w, d_{ij}) + \mu p(w|\theta_B)}{\sum_{w' \in V} \sum_{j=1}^{N_i} c(w', d_{ij}) + \mu} \qquad \mu \geq 0$$

$p(w|\theta_B)=1/|V|$?

What if $\mu \to \infty$?

Anatomy of Naïve Bayes Classifier

Two categories: θ_1 and θ_2

$$\mathrm{score}(d) = \log \frac{p(\theta_1 \mid d)}{p(\theta_2 \mid d)} = \log \frac{p(\theta_1)\prod_{w\in V} p(w \mid \theta_1)^{c(w,d)}}{p(\theta_2)\prod_{w\in V} p(w \mid \theta_2)^{c(w,d)}}$$

$$= \boxed{\log \frac{p(\theta_1)}{p(\theta_2)}} + \sum_{w\in V} \underline{c(w,d)} \boxed{\log \frac{p(w \mid \theta_1)}{p(w \mid \theta_2)}}$$

Category bias (β_0) doesn't depend on d!

Sum over all words (features $\{f_i\}$)

Feature value: f_i=c(w,d)

Weight on each word (feature) β_i

Generalize

$$d = (f_1, f_2, \ldots, f_M),\ \ f_i \in \Re$$

$$\mathrm{score}(d) = \beta_0 + \sum_{i=1}^{M} f_i \beta_i \qquad \beta_i \in \Re$$

= Logistic Regression!