

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

Category 1

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

Category 2

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

Category k

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

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  $p(w | \theta_i)$ ?