Naïve Bayes Model
Lecture 5: Sept 25, 2013

CS886-2 Natural Language Understanding

University of Waterloo

Classification

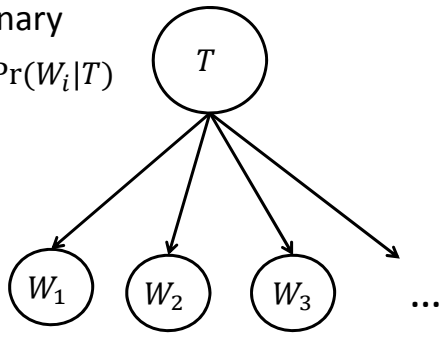
- NLP tasks:
  - Spam filtering
  - Sentiment analysis
  - Text categorization
- Classification techniques
  - Decision trees
  - Support Vector Machines
  - Naïve Bayes

Naïve Bayes Model

Generative model

- Sample topic from $\Pr(T)$
- For each word in dictionary

  Sample $W_i = 1/0$ from $\Pr(W_i|T)$

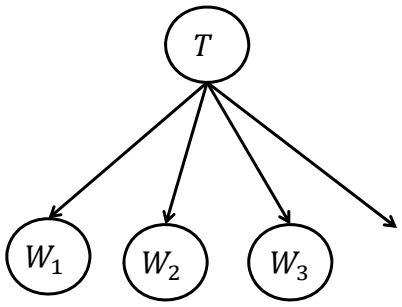


Simple Bayesian Network

- The name Naïve Bayes comes from the fact that it is a very simple Bayesian network.

- Bayesian network:
  - Graphical model:
    - Nodes: random variables
    - Edges: conditional dependencies
  - Conditional distributions: $\Pr(V|parents(V))$

Model Specification



T	$\Pr(T)$
finance	0.2
sports	0.8

W_1	T	$Pr(W_1\|T)$
1	finance	0.1
0	finance	0.9
1	sports	0.8
0	sports	0.2

W_2	T	$Pr(W_2\|T)$
1	finance	0.7
0	finance	0.3
1	sports	0.05
0	sports	0.95

W_3	T	$Pr(W_3\|T)$
1	finance	0.75
0	finance	0.25
1	sports	0.2
0	sports	0.8

Joint distribution

- Joint distribution:

$$\Pr(T, W_1, W_2, W_3, \dots) = \Pr(T) \prod_i \Pr(W_i|T)$$

- Joint assignment: document (bag of words):

d_1: $T = sports, W_1 = 1, W_2 = 1, W_3 = 0, \dots$

d_2: $T = finance, W_1 = 0, W_2 = 1, W_3 = 1, \dots$

…

Classification

- Classification by probabilistic inference:

$$\Pr(T|W_1, W_2, W_3, \ldots, W_n)$$

- Computation:

$$\Pr(T|W_1, W_2, W_3, \ldots, W_n) = \frac{\Pr(T, W_1, W_2, W_3, \ldots, W_n)}{\Pr(W_1, W_2, W_3, \ldots, W_n)}$$
$$= \frac{\Pr(T) \prod_i \Pr(W_i|T)}{\Pr(W_1, W_2, W_3, \ldots, W_n)}$$
$$= \mathrm{k} \Pr(T) \prod_i \Pr(W_i|T)$$

Example

- $d_1 : \Pr(T|W_1 = 1, W_2 = 1, W_3 = 0)$

  $T = sports$: $k(0.8)(0.8)(0.05)(0.8) = 0.88$

  $T = finance$: $k(0.2)(0.1)(0.7)(0.25) = 0.12$

- $d_2 : \Pr(T|W_1 = 0, W_2 = 1, W_3 = 1)$

  $T = sports$: $k(0.8)(0.2)(0.05)(0.2) = 0.17$

  $T = finance$: $k(0.2)(0.9)(0.7)(0.75) = 0.83$

Parameter Estimation

- Idea: set conditional distributions to relative frequency counts

$$\Pr(T = finance) = \frac{\#(T=finance)}{\#documents}$$

$$\Pr(W_i = 1 | T = finance) = \frac{\#(W_i=1 \wedge T=finance)}{\#(T=finance)}$$

Maximum Likelihood

- Parameter estimation by relative frequency follows from the principle of maximum likelihood

- Maximum likelihood:

  $\theta^* = argmax_\theta \Pr(data|\theta)$

  where θ is the set of parameters

Estimation of Topic Distribution

- Let $\theta = \Pr(T = sports)$
- Corpus:

  $\#s$: # of documents about sports

  $\#f$: # of documents about finance

Estimation of Topic Distribution

- 1) Likelihood expression

$$\Pr(data|\theta) = \theta^{\#s}(1-\theta)^{\#f}$$

- 2) log likelihood

$$\log \Pr(data|\theta) = \#s \log \theta + \#f \log(1-\theta)$$

- 3) log likelihood derivative

$$\frac{\partial(\log \Pr(data|\theta))}{\partial\theta} = \frac{\#s}{\theta} - \frac{\#f}{1-\theta}$$

- 4) ML hypothesis

$$\frac{\#s}{\theta} - \frac{\#f}{1-\theta} = 0 \rightarrow \theta = \frac{\#s}{\#s+\#f}$$

Full Parameter Estimation

- Let $\theta = \Pr(T = sports)$
  $\theta_{si} = \Pr(W_i = 1 | T = sports)$
  $\theta_{fi} = \Pr(W_i = 1 | T = finance)$
- Corpus:
  $\#s$: # of documents about sports
  $\#f$: # of documents about finance
  $\#s1i$: # of documents about sports with w_i
  $\#s0i$: # of documents about sports without w_i
  $\#f1i$: # of documents about finance with w_i
  $\#f0i$: # of documents about finance without w_i

Full Parameter Estimation

- 1) Likelihood expression

$$\Pr(data|\theta) = \theta^{\#s}(1-\theta)^{\#f} \prod_i \theta_{si}^{\#s1i}(1-\theta_{si})^{\#s0i}\theta_{fi}^{\#f1i}(1-\theta_{fi})^{\#f0i}$$

…

- 4) ML hypothesis

$$\frac{\#s}{\theta} - \frac{\#f}{1-\theta} = 0 \rightarrow \theta = \frac{\#s}{\#s+\#f}$$

$$\frac{\#s1i}{\theta_{si}} - \frac{\#s0i}{1-\theta_{si}} = 0 \rightarrow \theta_{si} = \frac{\#s1i}{\#s1i+\#s0i}$$

$$\frac{\#f1i}{\theta_{fi}} - \frac{\#f0i}{1-\theta_{fi}} = 0 \rightarrow \theta_{fi} = \frac{\#f1i}{\#f1i+\#f0i}$$

Laplace smoothing

- An important case of overfitting happens when there is no sample for a certain outcome
  - E.g. no document about sports contains w_i

$$\Pr(W_i = 1|T = sports) = \theta_{s1i} = \frac{\#s1i}{\#s1i + \#s0i} = 0$$

  - Zero probabilities are dangerous: they rule out outcomes
- Solution: Laplace (add-one) smoothing
  - Add 1 to all frequencies

$$\Pr(W_i = 1|T = sports) = \theta_{s1i} = \frac{\#s1i + 1}{\#s1i + \#s0i + 2}$$

  - Much better results in practice