CSCI 4152/6509 — Natural Language Processing *16-Oct-2009*

Lecture 15: Naïve Bayes Model

Room: FASS 2176
Time: 11:35 – 12:25

Previous Lecture

- Fully independent model
  - example,
  - computational tasks,
- Sum-product formula;
- Naive Bayes model: motivation

9 Naïve Bayes model

In the Naïve Bayes model we assume that all variables are independent except one distinguished variable, the **class variable.** This variable is also called the **output variable,** and the other variables may be called the **input variables.**

If we assume that the variable V_1 is the output variable, and the variables V_2, V_3, …, V_n are the input variables, then in the classification problem can be expressed as a conditional probability computation problem, or completion problem of the probability:

$$\mathrm{P}(V_1 = x_1 | V_2 = x_2, V_3 = x_3, \ldots, V_n = x_n)$$

or

$$\mathrm{P}(V_1 | V_2, V_3, \ldots, V_n)$$

for short. After applying Bayes theorem we obtain:

$$\mathrm{P}(V_1 | V_2, V_3, \ldots, V_n) = \frac{\mathrm{P}(V_2, V_3, \ldots, V_n | V_1) \cdot \mathrm{P}(V_1)}{\mathrm{P}(V_2, V_3, \ldots, V_n)}$$

If we assume that variables $V_2, V_3, \ldots, V_n$ are conditionally independent given V_1, then the above equation becomes:

$$\begin{aligned}\mathrm{P}(V_1 | V_2, V_3, \ldots, V_n) &= \frac{\mathrm{P}(V_2, V_3, \ldots, V_n | V_1) \cdot \mathrm{P}(V_1)}{\mathrm{P}(V_2, V_3, \ldots, V_n)} \\ &= \frac{\mathrm{P}(V_2 | V_1) \cdot \mathrm{P}(V_3 | V_1) \cdot \ldots \cdot \mathrm{P}(V_n | V_1) \cdot \mathrm{P}(V_1)}{\mathrm{P}(V_2, V_3, \ldots, V_n)}\end{aligned}$$

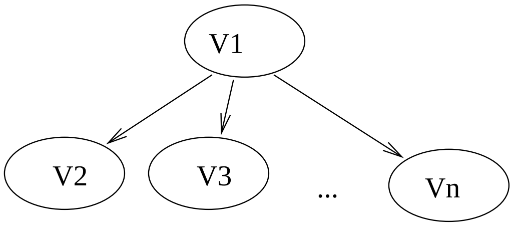
The conditional probabilities $\mathrm{P}(V_i | V_1)$ for $i \in \{2 \ldots n\}$ can be efficiently computed and stored, and they eliminate the sparse data problem.

Another way of deriving the Naïve Bayes formula is the following:

$$\begin{aligned}& \mathrm{P}(V_1 = x_1, \ldots, V_n = x_n) = && (3) \\ = \;& \mathrm{P}(V_1 = x_1)\mathrm{P}(V_2 = x_2 | V_1 = x_1)\mathrm{P}(V_3 = x_3 | V_1 = x_1, V_2 = x_2) \ldots && (4) \\ & \mathrm{P}(V_n = x_n | V_1 = x_1, V_2 = x_2, \ldots, V_{n-1} = x_{n-1}) && (5) \\ \overset{\text{NB}}{\approx} \;& \mathrm{P}(V_1 = x_1)\mathrm{P}(V_2 = x_2 | V_1 = x_1)\mathrm{P}(V_3 = x_3 | V_1 = x_1) \ldots && (6) \\ & \mathrm{P}(V_n = x_n | V_1 = x_1) && (7)\end{aligned}$$

Equality (3,4) holds always, and equality (5,6) is the Naïve Bayes assumption.

Graphical Representation



Example: A Naïve Bayes Model for Spam Detection

In our spam detection example, the Naïve Bayes assumption is:

$$\mathrm{P}(\mathit{Free}, \mathit{Caps}, \mathit{Spam}) = \mathrm{P}(\mathit{Spam}) \cdot \mathrm{P}(\mathit{Free}|\mathit{Spam}) \cdot \mathrm{P}(\mathit{Caps}|\mathit{Spam})$$

Hence, in order to create a Naïve Bayes model from our training data:

Free	*Caps*	*Spam*	Number of messages
Y	Y	Y	20
Y	Y	N	1
Y	N	Y	5
Y	N	N	0
N	Y	Y	20
N	Y	N	3
N	N	Y	2
N	N	N	49
		Total:	100

we calculate the following tables:

Spam	$\mathrm{P}(\mathit{Spam})$
Y	$\frac{20+5+20+2}{100} = 0.47$
N	$\frac{1+0+3+49}{100} = 0.53$

,

Caps	*Spam*	$\mathrm{P}(\mathit{Caps}\vert\mathit{Spam})$
Y	Y	$\frac{20+20}{20+5+20+2} \approx 0.8511$
Y	N	$\frac{1+3}{1+0+3+49} \approx 0.0755$
N	Y	$\frac{5+2}{20+5+20+2} \approx 0.1489$
N	N	$\frac{0+49}{1+0+3+49} \approx 0.9245$

, and

Free	*Spam*	$\mathrm{P}(\mathit{Free}\vert\mathit{Spam})$
Y	Y	$\frac{20+5}{20+5+20+2} \approx 0.5319$
Y	N	$\frac{1+0}{1+0+3+49} \approx 0.0189$
N	Y	$\frac{20+2}{20+5+20+2} \approx 0.4681$
N	N	$\frac{3+49}{1+0+3+49} \approx 0.9811$

.

The probability of a configuration in this model is calculated in the following way:

$$\begin{aligned} &\mathrm{P}(\mathit{Free} = Y, \mathit{Caps} = N, \mathit{Spam} = N) = \\ &= \quad \mathrm{P}(\mathit{Spam} = N) \cdot \mathrm{P}(\mathit{Caps} = N|\mathit{Spam} = N) \cdot \mathrm{P}(\mathit{Free} = Y|\mathit{Spam} = N) \\ &\approx \quad 0.53 \cdot 0.9245 \cdot 0.0189 \approx 0.0093 \end{aligned} \tag{8}$$

9.1 Computational Tasks in the Naïve Bayes Model

We will cover the computational tasks in more details within the Bayesian Network in general.

1. Evaluation

The probability of a complete configuration is calculated using the Naïve Bayes assumption and table lookups. The formula (8) illustrates probability evaluation of a complete configuration: $\mathrm{P}(\textit{Free} = Y, \textit{Caps} = N, \textit{Spam} = N)$ This example illustrates the fact that the Naïve Bayes model is less amenable to the sparse date problem than the joint distribution problem, in which the probability of this same configuration was estimated to be 0.

2. Simulation

Configurations are sampled by first sampling the output variable based on its table, and then the input variables using the corresponding conditional tables.

3. Inference

Marginalization. If the partial configuration includes the output variable, it can be shown that the marginal probability can be calculated using the following formula:

$$\begin{aligned}&\mathrm{P}(V_1 = x_1, \ldots, V_k = x_k) = \\ &\quad \mathrm{P}(V_1 = x_1)\mathrm{P}(V_2 = x_2|V_1 = x_1)\mathrm{P}(V_3 = x_3|V_1 = x_1)\ldots \\ &\quad \mathrm{P}(V_k = x_k|V_1 = x_1)\end{aligned}$$

Conditioning. Example:

$$\mathrm{P}(S = N|F = Y, C = N) \quad = \quad \frac{\mathrm{P}(S = N, F = Y, C = N)}{\mathrm{P}(F = Y, C = N)}$$

Using Naïve Bayes assumption:

$$\begin{aligned}&\mathrm{P}(S = N, F = Y, C = N) = \\ &\quad = \quad \mathrm{P}(S = N)\mathrm{P}(F = Y|S = N)\mathrm{P}(C = N|S = N) \\ &\quad = \quad 0.53 \cdot 0.9245 \cdot 0.0189 \approx 0.093\end{aligned}$$

$$\begin{aligned}&\mathrm{P}(F = Y, C = N) = \text{(by definition)} \\ &\quad = \quad \mathrm{P}(S = Y, F = Y, C = N) + \mathrm{P}(S = N, F = Y, C = N) \\ &\quad \approx \quad \mathrm{P}(S = Y)\mathrm{P}(F = Y|S = Y)\mathrm{P}(C = N|S = Y) + 0.093 \\ &\quad = \quad 0.47 \cdot 0.5319 \cdot 0.1489 + 0.093 \\ &\quad \approx \quad 0.0465\end{aligned}$$

Finally,

$$\mathrm{P}(S = N|F = Y, C = N) \quad = \quad \frac{0.0093}{0.0465} \approx 0.2$$

Completion

Example: $$\operatorname*{arg\,max}_{s \in \{Y,N\}} \mathrm{P}(S = s|F = Y, C = N) \overset{\textit{by definition}}{=} \operatorname*{arg\,max}_{s} \frac{\mathrm{P}(S = s, F = Y, C = N)}{\mathrm{P}(F = Y, C = N)}$$

$\mathrm{P}(F = Y, C = N)$ does not depend on s, hence

$$= \operatorname*{arg\,max}_{s} \mathrm{P}(S = s, F = Y, C = N)$$

and by using Naïve Bayes assumption)

$$= \arg\max_s \underbrace{\mathrm{P}(S=s)\mathrm{P}(F=Y|S=s)\mathrm{P}(C=N|S=s)}_{A(s)}$$

For $s = Y$ $A(s = Y) \approx 0.0465$, and for $s = N$ $A(s = N) \approx 0.0093$; hence

$$\arg\max_s A(s) = Y$$

Learning

Maximum Likelihood Estimation: The parameters are estimated using a corpus.

9.2 Number of Parameters

A Naïve Bayes model with n variables V_1,... V_n is described with tables $\mathrm{P}(V_1)$, $\mathrm{P}(V_2|V_1)$, $\mathrm{P}(V_3|V_1), \ldots, \mathrm{P}(V_n|V_1)$. These tables have constraints since each probability distribution must sum up to 1. If we assume that each variable can take one of m distinct values, then the number of parameters and constraints in required tables are:

	parameters	constraints
table $\mathrm{P}(V_1)$	m	1
table $\mathrm{P}(V_2\|V_1)$	m^2	m
table $\mathrm{P}(V_3\|V_1)$	m^2	m
$\vdots$	$\vdots$	$\vdots$
table $\mathrm{P}(V_n\|V_1)$	m^2	m
sum	$m + (n-1)m^2$	$1 + (n-1)m$

Hence, the number of free parameters is $m + (n-1)m^2 - 1 - (n-1)m = O(m^2 n)$, which is not very large since the joint distribution model requires $O(m^n)$ parameters.

Pros and Cons of the Naïve Bayes model

Pros:
- efficient
- no sparse data problem
- surprisingly good performance (accuracy), e.g., in text classification

Cons:
- can be over-simplifying
- cannot model more than one "output" variable