Natural Language Processing CSCI 4152/6509 — Lecture 11 Naïve Bayes Model; P0 Topics Discussion (3)

Instructors: Vlado Keselj

Time and date: 16:05 – 17:25, 16-Oct-2024

Location: Carleton Tupper Building Theatre C

Previous Lectures

- Joint distribution model
    - Spam example
- Fully independent model
- P0 Discussion (2): P-02

3.c Completion in Fully Independent Model

$$\begin{aligned}
y^*_{k+1},\ldots,y^*_n &= \arg\max_{y_{k+1},\ldots,y_n} \mathrm{P}(V_{k+1}=y_{k+1},\ldots,V_n=y_n|V_1=x_1,\ldots,V_k=x_k) \\
&= \arg\max_{y_{k+1},\ldots,y_n} \mathrm{P}(V_{k+1}=y_{k+1})\cdots\mathrm{P}(V_n=y_n) \\
&= \left[\arg\max_{y_{k+1}} \mathrm{P}(V_{k+1}=y_{k+1})\right]\cdots\left[\arg\max_{y_n} \mathrm{P}(V_n=y_n)\right]
\end{aligned}$$

Joint Distribution Model vs. Fully Independent Model

- Fully Independent Model addresses some issues of the Joint Distribution Model
- Efficient and small number of parameters
- However: too strong assumption, no structure
- Too trivial to be usable
- Better method: Structured probability models
    - compromise between no dependence and too much dependence

Naïve Bayes Classification Model

- Fully independent model is not useful in classification: class variable should be dependent on other variables
- A solution: make class variable dependent, but everything else independent
- Let V_1 be the class variable
- V_2, V_3, ..., V_n are input variables (features)
- Classification can be expressed as

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

Naïve Bayes Independence Assumption

- After applying Bayes theorem we obtain:

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

- We assume that $V_2, V_3, \ldots, V_n$ are conditionally independent given V_1: **Naïve Bayes Independence Assumption (1):**

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

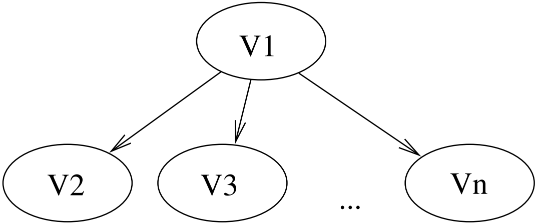
- or as an equivalent formula for **Naïve Bayes Independence Assumption (2):**

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

Graphical Representation: Naïve Bayes Model

Assumption:

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



Naïve Bayes Classification

- The classification formula becomes

$$\arg\max_{x_1} \frac{\mathrm{P}(V_2|V_1)\cdot\mathrm{P}(V_3|V_1)\cdot\ldots\cdot\mathrm{P}(V_n|V_1)\cdot\mathrm{P}(V_1)}{\mathrm{P}(V_2,V_3,\ldots,V_n)} =$$

$$\arg\max_{x_1} \mathrm{P}(V_2|V_1)\cdot\mathrm{P}(V_3|V_1)\cdot\ldots\cdot\mathrm{P}(V_n|V_1)\cdot\mathrm{P}(V_1)$$

- To calculate marginal probability in the denominator we use

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

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

Another Derivation of Naïve Bayes Assumption

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

$$
\begin{align}
& \mathrm{P}(V_1 = x_1, \ldots, V_n = x_n) = \tag{1}\\
= \quad & \mathrm{P}(V_1 = x_1)\mathrm{P}(V_2 = x_2 | V_1 = x_1)\mathrm{P}(V_3 = x_3 | V_1 = x_1, V_2 = x_2) \ldots \tag{2}\\
& \mathrm{P}(V_n = x_n | V_1 = x_1, V_2 = x_2, \ldots, V_{n-1} = x_{n-1}) \tag{3}\\
\overset{\text{NB}}{\approx} \quad & \mathrm{P}(V_1 = x_1)\mathrm{P}(V_2 = x_2 | V_1 = x_1)\mathrm{P}(V_3 = x_3 | V_1 = x_1) \ldots \tag{4}\\
& \mathrm{P}(V_n = x_n | V_1 = x_1) \tag{5}
\end{align}
$$

Summary of the Naïve Bayes Model

$$\underbrace{P(V2,V3,\ldots Vn}_{\text{text features}}|\underset{\text{class variable}}{V1}) \overset{\text{Naive Bayes assumption}}{=} P(V2|V1)\ P(V3|V1)\ \ldots\ P(Vn|V1)$$

Second way of expression Naive Bayes Assumption:

$$P(V1,V2,V3,\ldots,Vn) = P(V1)\ P(V2,V3,\ldots,Vn|V1) \overset{\downarrow}{=}$$
$$= P(V1)\ P(V2|V1)\ P(V3|V1)\ \ldots\ P(Vn|V1)$$

Naive Bayes Model is a set of tables

V1	P(V1)

V1	V2	P(V2|V1)

V1	Vn	P(Vn|V1)

(CPT –– Conditional Probability Tables)

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

Naïve Bayes Model Parameters

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*	$\mathrm{P}(\mathit{Caps}\|\mathit{Spam})$
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*	$\mathrm{P}(\mathit{Free}\|\mathit{Spam})$
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	$\frac{3+49}{1+0+3+49} \approx 0.9811$

.

Computational Tasks in the Naïve Bayes Model:

1. Evaluation

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

$$\begin{aligned} &\mathrm{P}(\mathit{Free}=Y, \mathit{Caps}=N, \mathit{Spam}=N) = \qquad (6)\\ &= \quad \mathrm{P}(\mathit{Spam}=N)\cdot\mathrm{P}(\mathit{Caps}=N|\mathit{Spam}=N)\cdot\mathrm{P}(\mathit{Free}=Y|\mathit{Spam}=N)\\ &\approx \quad 0.53\cdot 0.9245\cdot 0.0189 \approx 0.0093 \end{aligned}$$

No sparse data problem, when compared with previous Joint Distribution model.

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

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

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

3.b) Conditioning: Example

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

Using Naïve Bayes assumption:

$$\begin{aligned}
&\mathrm{P}(S=N,F=Y,C=N) = \\
&\quad = \mathrm{P}(S=N)\mathrm{P}(F=Y|S=N)\mathrm{P}(C=N|S=N) \\
&\quad = 0.53 \cdot 0.9245 \cdot 0.0189 \approx 0.0093
\end{aligned}$$

$$\begin{aligned}
&\mathrm{P}(F=Y,C=N) = \text{(by definition)} \\
&\quad = \mathrm{P}(S=Y,F=Y,C=N) + \mathrm{P}(S=N,F=Y,C=N) \\
&\quad \approx \mathrm{P}(S=Y)\mathrm{P}(F=Y|S=Y)\mathrm{P}(C=N|S=Y) + 0.0093 \\
&\quad = 0.47 \cdot 0.5319 \cdot 0.1489 + 0.0093 \\
&\quad \approx 0.0465
\end{aligned}$$

Finally,

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

3.c) Completion in the NB Model

- Classification is the completion task:

$$\arg\max_{s\in\{Y,N\}} \mathrm{P}(S=s|F=Y,C=N)$$

- It works out that we calculate:

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

and

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

and choose the larger value.

Naïve Bayes Model: Learning

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

Number of Parameters

A Naïve Bayes model with n variables V_1,… V_n is described with tables $\mathrm{P}(V_1)$, $\mathrm{P}(V_2|V_1)$, $\mathrm{P}(V_3|V_1)$, …, $\mathrm{P}(V_n|V_1)$. Number of

		parameters	constraints
parameters:	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	$m+(n-1)m^2$	$1+(n-1)m$

Total: $O(m^2n)$

Pros and Cons of the Naïve Bayes Model

- Pros
  - efficient
  - no sparse data problem
  - surprisingly good classification performance (accuracy); e.g. in text classification
- Cons
  - can be over-simplifying (too strong assumption)
  - cannot model more than one “output” variable; i.e., hidden variable

Additional Notes on Naïve Bayes Model

- Text classification: how do we choose features?
- Two options:
  - Bernoulli Naïve Bayes — binary variables for each word
  - Multinomial Naïve Bayes — variable for each word position
- Zero-probability problem
  - Smoothing using $+1$ or similar addition (Laplace smoothing)

P0 Topics Discussion (3)

- Discussion of individual projects as proposed in P0 submissions
- Projects discussed: P-07, P-08, P-09, P-11, P-12, P-13