Natural Language Processing
CSCI 4152/6509 — Lecture 13
Naïve Bayes Model

Instructors: Vlado Keselj
Time and date: 16:05 – 17:25, 17-Oct-2023
Location: Rowe 1011

Previous Lecture

- P0 discussion: P-02
- Probabilistic modeling:
  - random variables, random models
  - full and partial model configurations
  - computational tasks in probabilistic modeling
- Joint distribution model
  - Spam example
- Fully independent model
- Naïve Bayes classification model
  - Assumption, definition
  - Graphical representation

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:

$$\mathrm{P}(V_1 = x_1, \ldots, V_n = x_n) = \tag{1}$$

$$= \quad \mathrm{P}(V_1 = x_1)\mathrm{P}(V_2 = x_2|V_1 = x_1)\mathrm{P}(V_3 = x_3|V_1 = x_1, V_2 = x_2) \ldots \tag{2}$$

$$\mathrm{P}(V_n = x_n|V_1 = x_1, V_2 = x_2, \ldots, V_{n-1} = x_{n-1}) \tag{3}$$

$$\overset{\text{NB}}{\approx} \quad \mathrm{P}(V_1 = x_1)\mathrm{P}(V_2 = x_2|V_1 = x_1)\mathrm{P}(V_3 = x_3|V_1 = x_1) \ldots \tag{4}$$

$$\mathrm{P}(V_n = x_n|V_1 = x_1) \tag{5}$$

Summary of the Naïve Bayes Model

Naive Bayes assumption

$$P(V2,V3,\ldots Vn|V1) = P(V2|V1)\,P(V3|V1)\,\ldots\,P(Vn|V1)$$

text features — class variable

Second way of expression Naive Bayes Assumption:

$$P(V1,V2,V3,\ldots,Vn) = P(V1)\,P(V2,V3,\ldots,Vn|V1) =$$
$$= P(V1)\,P(V2|V1)\,P(V3|V1)\,\ldots\,P(Vn|V1)$$

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*	$\mathrm{P}(\mathit{Caps} \mid \mathit{Spam})$
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*	$\mathrm{P}(\mathit{Free} \mid \mathit{Spam})$
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:

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

3.b) Conditioning: Example

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

Using Naïve Bayes assumption:

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

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

Finally,

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

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:

	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
⋮	⋮	⋮
table $\mathrm{P}(V_n\|V_1)$	m^2	m
sum	$m + (n-1)m^2$	$1 + (n-1)m$

Total: $O(m^2 n)$

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)

N-gram Model

- Before we introduce this model, introduce *language modeling*
- *Language Modeling:* Estimating probability of arbitrary NL sentence: P(sentence)
- Example: Speech recognition

$$\begin{aligned}
\arg\max_{\text{sentence}} \mathrm{P}(\text{sentence}|\text{sound}) &= \arg\max_{\text{sentence}} \frac{\mathrm{P}(\text{sentence}, \text{sound})}{\mathrm{P}(\text{sound})} \\
&= \arg\max_{\text{sentence}} \mathrm{P}(\text{sentence}, \text{sound}) \\
&= \arg\max_{\text{sentence}} \mathrm{P}(\text{sound}|\text{sentence})\mathrm{P}(\text{sentence})
\end{aligned}$$

- Acoustic model and Language model

Language Modeling

- Task of estimating probability of arbitrary utterance in a language
- Alternative task: Predicting the next token in a sequence: e.g., the next word or words, in a sentence, or next character or characters
- N-gram model: a "natural" model for this task

N-gram Model Assumption

$$\mathrm{P}(w_1 w_2 \dots w_n) = \mathrm{P}(w_1 | \cdot\cdot)\mathrm{P}(w_2 | w_1 \cdot)\mathrm{P}(w_3 | w_2 w_1) \dots \mathrm{P}(w_n | w_{n-1} w_{n-2})$$

N-gram Model: Notes

- Reading: Chapter 4 of [JM]
- Use of log probabilities
  - similarly as in the Naïve Bayes model for text
- Graphical representation