Naïve Bayes for Sentiment Analysis

- The PDF Notes are part of a YouTube tutorial: https://youtu.be/YVC-L_lLQb4
- The PDF slides, notes, and the video tutorial provided are for informational purposes only. Although the author has made efforts to ensure the quality of the slides, notes, and video tutorial, their accuracy or completeness is not guaranteed. It is advisable to cross-verify information with other sources and exercise your own judgment when making decisions.
- This tutorial is intended for personal use only. Therefore, please refrain from distributing it through any means.

1. Conditional Probability: Basic Format

$$p(A|B) = \frac{P(AB)}{P(B)}$$

Suppose you roll a fair six-sided die. The sample space, S, is $\{1, 2, 3, 4, 5, 6\}$.

Now, let's define two events:

- Event A: Getting an even number (2, 4, or 6)
- Event B: Getting a number greater than 3 (4, 5, or 6)

Now, we can calculate conditional probabilities.

- P(A) = 3/6 = 1/2.
- P(B) = 3/6 = 1/2.

Now, let's calculate the conditional probability of Event A given that Event B has occurred, denoted as P(A|B).

$$p(A|B) = 2/3$$

For $p(A|B)$, you can also calculate in another way:

$$p(A|B) = \frac{P(AB)}{P(B)}$$

- $P(AB) =$ P(A ∩ B) $= \frac{2}{6}$
- $P(B) = \frac{1}{2}$

Thus,

$$p(A|B) = \frac{2/6}{1/2} = \frac{2}{3}$$

2. Conditional Probability: Expanded Format

$$p(A|B) = \frac{p(AB)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

- Event A: Getting an even number (2, 4, or 6)
- Event B: Getting a number greater than 3 (4, 5, or 6)

$$p(B|A) = \frac{2}{3}$$

$$p(A) = \frac{1}{2}$$

$$p(B) = \frac{1}{2}$$

$$p(A|B) = \frac{p(B|A)P(A)}{p(B)} = \frac{\frac{2}{3} * \frac{1}{2}}{\frac{1}{2}} = \frac{2}{3}$$

3. Naïve Bayes for Sentiment Analysis

$$p(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- B = Sentence of "Great Product"
- A = P (Positive) vs. N (Negative)

$$p(B|A) = p("I\ like\ the\ product"|P)$$
$$p(B|A) = p("I\ like\ the\ product"|N)$$

We can then extend it as follows.

$$p(P|"I\ like\ the\ product") = \frac{p("I\ like\ the\ product"\ |P)\ p(P)}{p("I\ like\ the\ product")}$$

$$p(N|"I\ like\ the\ product") = \frac{p("I\ like\ the\ product"\ |N)\ p(N)}{p("I\ like\ the\ product")}$$

The decision is to compare which has a greater probability.

$$p(P|"I\ like\ the\ product") > p(N|"I\ like\ the\ product"): Positive$$
$$p(N|"I\ like\ the\ product") > p(P|"I\ like\ the\ product"): Negative$$

To make such a comparison, the probability $P("I\ like\ the\ product")$ does not impact such comparison. Thus, we can drop it and use the **proportional** ($\propto$) to represent it.

$$p(P|"I\ like\ the\ product") \propto p("I\ like\ the\ product"\ |P) * p(P)$$

$$p(N|"I\ like\ the\ product") \propto p("I\ like\ the\ product"\ |N) *\ p(N)$$

Here,

- $p(P)$ and $p(N)$ are the priors.
- $p("I\ like\ the\ product"\ |P)$ and $p("I\ like\ the\ product"\ |N)$ are the likelihood.
- $p(P|"I\ like\ the\ product")$ and $p(N|"I\ like\ the\ product")$ are the posteriors.

Further, we can extend as follows.

$$p(P|"I\ like\ the\ product") \propto p(P) * p("I"\ |P) * p("like"\ |P) * p("the"\ |P) *\ p("Product"\ |P)$$

$$p(N|"I\ like\ the\ product") \propto P(N) * P("I"\ |N) * p("like"\ |N) * p("the"\ |N) *\ p("Product"\ |N)$$

Note that, in the actual calculation, you might not need to include all the words in a sentence in the calculations of posteriors.

How to calculate $p("I"\ |P)\ and\ p("I"\ |N)$?

$$p("I"\ |P) = \frac{N_{"I"}}{N_{words\ in\ P}}$$

$$p("I"\ |N) = \frac{N_{"I"}}{N_{words\ in\ N}}$$

4. Naïve Bayes for Sentiment Analysis with Laplace smoothing.

$$p("I" \mid P) = \frac{N_{"I"}}{N_{words\ in\ P}} \quad p("I" \mid N) = \frac{N_{"I"}}{N_{words\ in\ N}}$$

We can simplify them into the following:

$$p(w_i|c) = \frac{N_{w_i}}{N_{words\ in\ C}}$$

Some words (w_i) in the test data might appear in one class (c) but not another class within the training data. For instance, "I" might not exist in any of the negative sentiment category. If that occurs, p("I" | c=negative) = 0. Then, $\prod_{i=1}^{n} p(w_i|c = \text{negative})$ will be 0.

To solve the problem, you need to use the **Laplace smoothing**.

$$p(w_i|c) = \frac{N_{w_i} + k}{N_{words\ in\ C} + k|V|}$$

where,

- k can be any number, but typically it is 1.
- $|V|$ represents all unique words across different classes.

5. Example for Naïve Bayes for Sentiment Analysis

	Reviews	Sentiment Class
Training	*Great product*	P
	A good product	P
	I like it	P
	Difficult to use	N
	complicated product	N
Test	*I like the product*	

Step 1: Calculate priors.

$$p(P) = \frac{3}{5}$$

$$p(N) = \frac{2}{5}$$

Step 2: Create a frequency table.

	Word	**P**	**N**
1	Great	1	
2	product	1+1=2	1
3	a	1	
4	good	1	
5	I	1	
6	Like	1	
7	it	1	
8	Difficult		1
9	To		1
10	use		1
11	complicated		1
	Total	8	5

Step 3: Calculate conditional probability:

• p("I" | P) = $\frac{1+1}{8+1*11} = \frac{2}{19}$ • p("like" | P) = $\frac{1+1}{8+1*11} = \frac{2}{19}$ • p("product" | P) = $\frac{2+1}{8+1*11} = \frac{3}{19}$	• p("I" | N) = $\frac{0+1}{5+1*11} = \frac{1}{16}$ • p("like" | N) = $\frac{0+1}{5+1*11} = \frac{1}{16}$ • p("product" | N) = $\frac{1+1}{5+1*11} = \frac{2}{16}$

Step 4: Classification:

$p(P|"I\ like\ the\ product")$:

- $p(P) *$ p("I" \| P) $*$ p("like" \| P) $*$ p("product" \| P) $= \frac{3}{5} * \frac{2*2*3}{19^3} = 1.0 \times 10^{-3}$

$p(N|"I\ like\ the\ product")$:

- $p(N) *$ p("I" \| N) $*$ p("like" \| N) * p("product" \|N) = $\frac{2}{5} * \frac{1*1*2}{16^3} = 0.2{\times}10^{-3}$