Naive Bayes

Bayes Theorem problem: https://youtu.be/LadMzl8MaXM
More Bayes Theorem problems:
https://www.math.upenn.edu/~mmerling/math107%20docs/practice%20on%20Bayes%20solutions.pdf
http://gtribello.github.io/mathNET/bayes-theorem-problems.html
http://wwwf.imperial.ac.uk/~ayoung/m2s1/WorkedExamples1.pdf

---

1. What is Conditional probability?

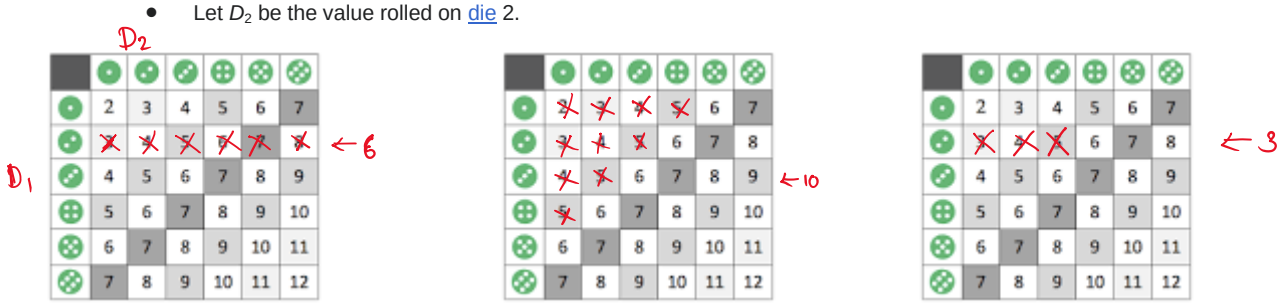
Let, A and B → random variable

$P(A|B) = P(A = a \mid B = b)$ — conditioned on given; a: value, b: value

Example

Suppose that somebody secretly rolls two fair six-sided dice, and we wish to compute the probability that the face-up value of the first one is 2, given the information that their sum is no greater than 5.

- Let D_1 be the value rolled on die 1.
- Let D_2 be the value rolled on die 2.



Probability that $D_1 = 2$
Table 1 shows the sample space of 36 combinations of rolled values of the two dice, each of which occurs with probability 1/36, with the numbers displayed in the red and dark gray cells being $D_1 + D_2$.
$D_1 = 2$ in exactly 6 of the 36 outcomes; thus $P(D_1 = 2) = 6/36 = 1/6$:

Probability that $D_1 + D_2 \leq 5$
Table 2 shows that $D_1 + D_2 \leq 5$ for exactly 10 of the 36 outcomes, thus $P(D_1 + D_2 \leq 5) = 10/36$:

Probability that $D_1 = 2$ given that $D_1 + D_2 \leq 5$
$P(D_1 = 2 \mid D_1 + D_2 \leq 5)$
Table 3 shows that for 3 of these 10 outcomes, $D_1 = 2$.
Thus, the conditional probability $P(D_1 = 2 \mid D_1 + D_2 \leq 5) = 3/10 = 0.3$:
Here, in the earlier notation for the definition of conditional probability, the conditioning event B is that $D_1 + D_2 \leq 5$, and the event A is $D_1 = 2$. We have

conditional probability $\frac{P(A \cap B)}{P(B)} = \frac{(3/36)}{(10/36)} = \frac{3}{10}$

conditional probability :-

probability of A given B (already happened) is equal to probability of A intersection B (means A in the given B) divided by probability of B

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{if } P(B) \neq 0$$

2. Define Independent vs Mutually exclusive events?

A and B are said to be independent

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

ex:-
- Getting value of 6 in die 1 $D_1 = 6$
- Getting value of 3 in die 2 $D_2 = 3$

$$P(D_1 = 6 \mid D_2 = 3) = P(D_1 = 6)$$

$$P(D_2 = 3 \mid D_1 = 6) = P(D_2 = 3)$$

Mutually Exclusive :-

If $P(A|B) = P(B|A) = 0$

Then A & B are said to be mutually exclusive

$$\frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(A)} = 0$$

3. Explain Bayes Theorem with example?

Baye's Theorem (Thomas Bayes 1700's)

Likelihood Probability | Prior Probability

Theorem :

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad \text{if } P(B) \neq 0$$

Posterior Probability | Evidence Probability

Proof :

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

we know → $A \cap B = B \cap A$

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

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

$$P(B \cap A) = P(B|A) * P(A)$$

Replacing it in initial eqn

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ ← Bayes-Theorem.

Example :-

Defective item rate

A factory produces items using three machines—A, B, and C—which account for 20%, 30%, and 50% of its output respectively. Of the items produced by machine A, 5% are defective; similarly, 3% of machine B's items and 1% of machine C's are defective. If a randomly selected item is defective, what is the probability it was produced by machine C?

This problem can also be solved using Bayes' theorem: Let X_i denote the event that a randomly chosen item was made by the i^{th} machine (for i = A,B,C). Let Y denote the event that a randomly chosen item is defective. Then, we are given the following information:

$P(X_A) = 0.2$ $P(X_B) = 0.3$ $P(X_C) = 0.5$

If the item was made by the first machine, then the probability that it is defective is 0.05; that is, $P(Y \mid X_A) = 0.05$. Overall, we have

$P(Y|X_A) = 0.05$ $P(Y|X_B) = 0.03$ $P(Y|X_C) = 0.01$

To answer the original question, we first find $P(Y)$. That can be done in the following way:

$$P(Y) = P(Y|X_A)P(X_A) + P(Y|X_B)P(X_B) + P(Y|X_C)P(X_C)$$

Hence, 2.4% of the total output is defective.

$$P(Y) = \sum_i P(Y|X_i)P(X_i) = (0.05)(0.2) + (0.03)(0.3) + (0.01)(0.5) = 0.024.$$

We are given that Y has occurred, and we want to calculate the conditional probability of X_C. By Bayes' theorem,

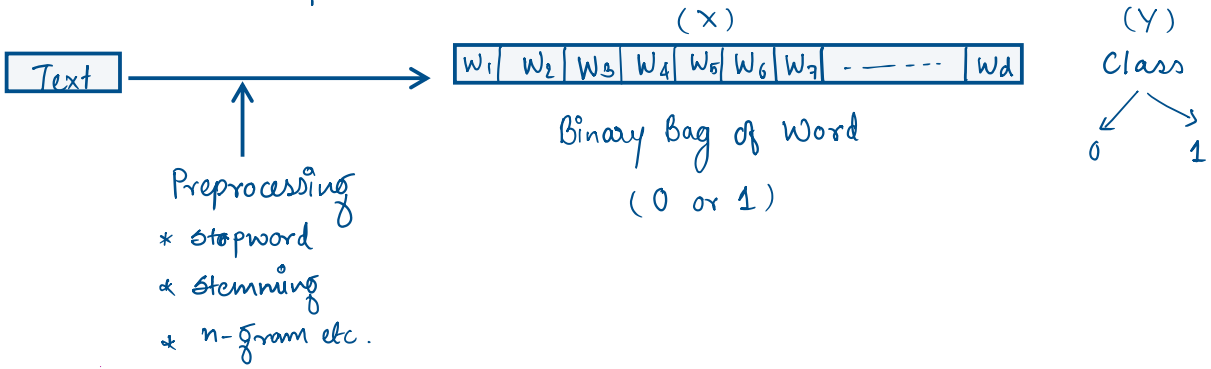
$$P(X_C|Y) = \frac{P(Y|X_C)P(X_C)}{P(Y)} = \frac{(0.01)(0.5)}{0.024} = 5/24$$

$\frac{1 \times 50}{2.4} \times \frac{10}{10}$

Given that the item is defective, the probability that it was made by machine C is 5/24. Although machine C produces half of the total output, it produces a much smaller fraction of the defective items. Hence the knowledge that the item selected was defective enables us to replace the prior probability $P(X_C) = 1/2$ by the smaller posterior probability $P(X_C \mid Y) = 5/24$.

4. How to apply Naive Bayes on Text data?

for text data naive bayes is a simple and most used as a baseline model.



$$P(Y=1|text) \propto P(Y=1|w_1, w_2, w_3 \cdots w_d)$$

$$P(Y=1|text) \propto P(Y=1) * P(w_1|Y=1) * (w_2|Y=1) \cdots * (w_d|Y=1)$$

posterior class prior likelihood

$$P(y=1 \mid \text{text}) \propto P(y=1) * \prod_{i=1}^{d} P(w_i \mid y=1)$$

similarly

$$P(y=0 \mid \text{text}) \propto P(y=0) * \prod_{i=1}^{d} P(w_i \mid y=0)$$

Prior : $P(y=1) = \dfrac{\#\text{ training pts with } y=1}{\text{Total \# of training points}}$

$$P(y=0) = \frac{\#\text{ training pts with } y=0}{\text{Total \# training pts}}$$

Likelihood

$$P(w_i \mid y=1) = \frac{\#\text{ data points with } w_i \text{ and with } y=1}{\#\text{ data points with } y=1}$$

$$P(w_i \mid y=0) = \frac{\#\text{ data points with } w_i \text{ and } y=0}{\#\text{ data points with } y=0}$$

5. What is Laplace/Additive Smoothing?

Problem :- There is a situation where word (w') is not in training data but it is in testing data.

Have to handle ? $P(w' \mid y=1)$ or $P(w' \mid y=0)$?

Ignoring it or dropping it is same as $P(w' \mid y=1) = 1$ or $P(w' \mid y=0) = 1$

$$P(w' \mid y=1) = \frac{P(w', y=1)}{P(y=1)} \leftarrow \#\text{ pts such that } w' \text{ occurs \& } y=1$$

$$= \frac{0}{n} = 0$$

It is dangerous as it will be multiplies

So, it will make $P(Y=1 \mid \text{text})$ becomes zero

similarly $P(y=0 \mid \text{text})$ becomes zero

solution: Laplace Smoothing (Additive Smoothing)

case 1:

$$P(w'|y=1) = \frac{0+\alpha}{n_1 + \alpha k}$$

normally $\alpha = 1$ (typically any smaller no.)

k ← no. of distinct value w' can take here 2 (ie. 0 & 1)

$$P(w'|y=1) = \frac{0+\alpha}{100 + \alpha \cdot 2}$$ ← let $n_1 = 100$ ∴ $\frac{1}{100+2} = \frac{1}{102} \neq 0$

case 2:

let $\alpha = 10{,}000$ & $n_1 = 100$

$$P(w'|y=1) = \frac{0+10000}{100+2\times 10000} = \frac{10000}{20100} \approx \frac{1}{2}$$

when α is large, we will assume w_i can be probable equally $y=1$ or $y=2$

$$\because P(w'|y=1) = P(w'|y=0) = \frac{1}{2} \approx 0.5$$

Additive → because we are adding new variable

Smoothing → as $\alpha \uparrow$; moving my likelihood probabilities towards uniform dist.

Laplace Smoothining should be done during training & testing.

6. Explain Log-probabilities for numerical stability?

$$P(y=1 | w_1, w_2 \cdots w_d) \propto P(y=1) * P(w_1|y=1) * P(w_2|y=1) \cdots P(w_d|y=1)$$

If $d = 100$ ∵ d is very large (Then we are multiplying 100 numbers) all which are probability (0 to 1)

we will have numerical stability problem

numerical underflow:- Python → double precision is upto 16 significant value after that python does rounding.

solution: we can use Log Probabilities

ex:- $0.2 \times 0.1 \times 0.2$

$= 0.0004$ & $\log(0.0004) = -2.23979$

log is a monotonic function and doesn't change order ie $x \uparrow$; $\log(x) \uparrow$

$$\log P(y=1 \mid w_1, w_2, w_3 \ldots w_d) = \log(P(y=1)) + \sum_{i=1}^{d} \log(P(w_i \mid y=1))$$

* log converts multiplication to addition
* converts exponential to multiplication ($\log(a^b) = b \log a$)

7. In Naive bayes how to handle Bias and Variance trade-off?

(Alpha) α in Laplace smoothing

case 1: when $\alpha = 0$

small change in D_{train} results in large change in model (overfitting)

case 2: when α is very large ex: $\alpha = 10000$ & w' occurs 2 times ; $n = 1000$

$$P(w' \mid y=1) = \frac{2 + 10000}{1000 + 2 \times 10000} \approx \frac{1}{2} = 0.5$$

$$P(Y_q = 1 \mid x_q) \approx P(Y_q = 0 \mid x_q) \approx \frac{1}{2}$$ (underfitting)

in this case if $n_1 > n_2$ then A will be declared as n_1 (majority) class
(n_1: $Y=1$, n_2: $Y=0$)

∴ Bias variance tradeoff depends on α (Alpha) in Naive Bayes.

8. What is Imbalanced data?

n — n_1 (+ve), n_2 (-ve) where $n_1 >> n_2$ or $n_2 >> n_1$

$$P(y=1 \mid \text{text}) = \underbrace{P(y=1)}_{\text{class prior}} * \prod_{i=1}^{d} (w_i \mid y=1)$$

Assume: 90% of pts are +ve (n_1)

10% of pts are −ve (n_2)

In class prior

$P(Y=1) = n_1/n = 0.9$ & $P(Y=0) = n_2/n = 0.1$

If we have imbalanced data

(1.) class prior → majority / dominant class has advantage.

sol^n 1. upsampling } $n_1 \approx n_2$ $P(Y=0) = P(Y=1) = 1/2$

2. down sampling

3. Modify N.B to account for class imbalance.

↳ we use different α for minority & majority class

9. What is Outliers and how to handle outliers?

1. outlier during test time:

$x_q = w_1, w_2, w_3, w_4, w^*$ } w^* was not present during train

Laplace Smoothing will take care of outlier during test time

(2) what happens to outlier during training?

$\{w_1, w_2, w_3 \cdots w_d\}$ set of words in D_{train}

Assume w_8 occurs very few times in +ve & -ve class

↑ Outlier

Solution:

(1) Set $\tau = 10$ (threshold)

If word (w_j) occurs less than 10 times (τ), then just drop or ignore that word from training data

(2) Laplace Smoothing with right (α) value.

10. How to handle Missing values?

(1) Text data → There is usually no case for missing data

(2) categorical feature → consider Nan as category ∵ $f_i \in \{a_1, a_2, a_3, NaN\}$

(3) Numerical feature → standard imputation
or model imputation

11. How to Handling Numerical features (Gaussian NB)

$f_1, f_2, f_3 \cdots f_d \longrightarrow$ real or numerical feature

$P(y=1 \mid x_{i1}, x_{i2}, x_{i3} \cdots x_{id})$

$P(y=1) \prod_{i=1}^{d} (x_{ij} \mid y=1)$

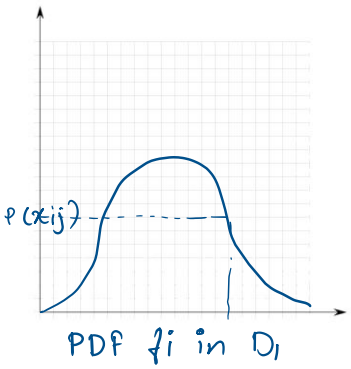
$P(y=1)$ — Class prior — $\frac{n_1}{n_1+n_2}$

$(x_{ij} \mid y=1)$ — likelihood

$D_{train} \begin{cases} D_1 \text{ (+ve only)} \\ D_2 \text{ (-ve only)} \end{cases}$

$P(x_{ij} \mid y=1) \longrightarrow P(x_{ij} \mid D_1)$

$P(x_{ij} \mid y=0) \longrightarrow P(x_{ij} \mid D_2)$



plot PDF for D_1 & D_2

obtain $P(x_{ij} \mid D_1)$ from PDF of f_i in D_1.

Assume:- f_i in D_1 (+ve) $\longrightarrow$ Gaussian distribution $N(\mu_j^{y=1}, \sigma_j^{y=1})$

similarly f_i in $D_2 \longrightarrow$ Gaussian distribution $N(\mu_j^{y=0}, \sigma_j^{y=0})$

Assumption: probability is coming from Gaussian distribution

① such model is called as "Gaussian Naive Bayes"

② Binary $P(x_i \mid y=1)$ $\begin{matrix} 0 \\ 1 \end{matrix}$ is called as Bernoulli Naive Bayes (uniform distribution)

③ Multinomial Naive Bayes → likelihood probability are multinomial

④ Naive Bayes → conditional independence.

12. Define Multiclass classification.?

Naive Bayes can do multiclass classification inherently :

C → class

$P(y=0 \mid w_1, w_2 \cdots w_d)$
$P(y=1 \mid w_1, w_2, \cdots w_d)$
$P(y=2 \mid w_1, w_2, \cdots w_d)$
$\vdots$
$P(y=C \mid w_1, w_2 \cdots w_d)$

} compare probabilities

Note Naive bayes cannot be use distance / similarity Matrix

it is a probability based method and it requires actual feature value

Other important points :-

① Large dimensionality : Naive Bayes is extremely used for text-based classification which has high dimensionality

② conditional independance ← assumption

If True ← works very well

else ← works reasonably well

③ categorical feature → N.B is mostly used for categorical features
→ not much for numerical feature

④ Interpretability and feature importance is very good.

⑤ runtime & Traintime (space) complexity is low (can be used for low latency)

⑥ Easily overfit if you donot do laplace Smoothing