Naive Bayes

J Junaid Mahmud

Probability

💡 Probability helps to predict an event's occurrence out of all the potential outcomes.

$$Probability\ of\ an\ event = \frac{Number\ of\ Favorable\ events}{Total\ Number\ of\ outcomes}$$

.

$$0\ \leq\ Probability\ of\ an\ event\ \leq\ 1$$

Marginal Probability

The probability of an event occurring (**p(A)**) in isolation. It may be thought of as an unconditional probability. It is not conditioned on another event.

- The probability that a card drawn is **red (p(red) = 0.5)**
- The probability that a card drawn is a **4 (p(four) = 1/13)**

Unconditional Probability

- We will only talk about one event. Nothing happened before or will happen after that event.

Joint Probability

💡 Joint Probability is that of event A and event B occurring. It is the probability of the **intersection** of two or more events. The probability of the intersection of A and B may be written and following

$$p\left(A \bigcap B\right)$$

Example: The probability that a card is a four and red = p(four and red) = 2/52 = 1/26

$$P\left(four\ and\ red\right) = \frac{2}{52} = \frac{1}{26}$$

Conditional Probability

💡 Conditional probability is a subset of probability. It reduces the probability of becoming dependent on a single event. You can compute the conditional probability for two or more occurrences.

Take events X and Y, the conditional probability of event Y is defined as the probability that the event occurs when event X is already over. It is written as

$$P\left(Y \mid X\right)$$
$$\ldots$$
$$Here,$$
$$P\left(Y \mid X\right) = \frac{P\left(X\ and\ Y\right)}{P\left(X\right)}$$

Example: Draw a red card, what is the probability that it’s a four

$$P(four \mid red) = \frac{2}{26} = \frac{1}{13}$$

Basics of Bayesian Statistics

Bayes’ Theorem applied to probability distribution

Laplace Smoothing

Advantages

- **Less Complex:** Compares to other classifiers, Naive Bayes is considered a simpler classifier since the parameters are easier to estimate.
- **Scales well:** Compared to logistic regression, Naive Bayes is considered a fast and efficient classifier that is fairly accurate when the conditional independence assumption holds. It also has a low storage requirements.
- **Can handle high-dimensional data:** Use cases, such as document classification, can have a high number of dimensions, which can be difficult for other classifiers to manage.

Disadvantages

- **Subject to Zero Frequency:** Zero frequency occurs when a categorical variable does not exist within the training set. The probability is this case would be zero, and since this classifier multiplies all the conditional probabilities together, this also means that the posterior probability will be zero. To avoid this issue, Laplace smoothing can be leveraged.
- **Unrealistic core assumption:** While the conditional independence assumption overall performs well, the assumption does not always hold, leading to incorrect classifications.

- Difference with discriminative models (LR).
- Naive Bayes is part generative model.
- Naive Bayes is the simplest Bayesian Probabilistic Mode.

Basics of Bayesian Statistics

Diabetes Example

What it does?

Why do we need it?

- For diabetes diagnosis or any test, we can have the following test results.

Result	Predicted	Actual
True Positive	True	True
True Negative	False	False
False Positive	True	False
False Negative	False	True

- False Positive: I **don't have diabetes**, but tested **positive**.
- False Negative: I **have diabetes**, but tested **negative**.

Bayes' Theorem for Point Probabilities

💡 The theorem says that a conditional probability for event B given event A is equal to the conditional probability of event A given event B, multiplied by the marginal probability for event B and divided by the marginal probability for event A.

$$p\ (B\ \mid\ A)\ =\ \frac{p\ (A\ \mid\ B).\ p\ (B)}{p\ (A)}$$

So, for diabetes, we can say, for diabetes:

$$p(diabetes\ \mid\ test+)\ =\ \frac{p\ (test+\ \mid diabetes).\ p(diabetes)}{p(test+\ \mid diabetes).\ p(diabetes) + p(test+\ \mid not\ diabetes).\ p(not\ diabetes)}$$

- After calculation we get:
  - **Posterior probability** (left hand side)

- Here (right hand side):
  - Diabetes is data
  - Test is Observation

- Prior probability
- ⇒ Posterior Probability: It is the estimated probability of being diabetic obtained after observing the data (the positive test).

Test 1

$$
\begin{aligned}
& p(diabetes \mid test+) \\
& = \frac{(.90)\,(0.15)}{(.90)\,(0.15) + (.50)\,(0.85)} \\
& = \frac{0.135}{0.135 + 0.425} \\
& = 0.241
\end{aligned}
$$

- Positive test result
- Result interpretation. Convincing result?
  - No, not good enough.

Test 2

- Updated prior probability of being a diabetic **(p = .241)**

$$
\begin{aligned}
& p(diabetes \mid test+) \\
& = \frac{(.90)\,(0.241)}{(.90)\,(0.241) + (.50)\,(0.759)} \\
& = \frac{0.217}{0.217 + 0.380} \\
& = 0.363
\end{aligned}
$$

Test 3

- Updated prior probability of being a diabetic **(p = .363)**

$$
\begin{aligned}
& p(diabetes \mid test+) \\
& = \frac{(.90)\,(0.363)}{(.90)\,(0.363) + (.50)\,(0.637)} \\
& = \frac{0.327}{0.327 + 0.319} \\
& = 0.506
\end{aligned}
$$

- Still not good enough.

Subsequent positive tests yield the following probabilities:

Test Number	Probability
Test 4	0.649
Test 5	0.769
Test 6	0.857
Test 7	0.915
Test 8	0.951
Test 9	0.972
Test 10	0.984

- The more we test, the difference in result decreases. From **test9 to test10**, only **0.01** probability increased. Why the increment in confidence is getting less?

⇒ **Prior probability** is **increasing**, so the **numerator is increasing**, **denominator decreasing**, so the increment in confidence in getting less.

⇒ From a bayesian perspective, we begin with prior probability for some event, and we update this prior probability with new information to obtain a posterior probability. The posterior probability can then be used as a prior probability in a subsequent analysis. From a Bayesian point of view, this is an appropriate strategy for conducting scientific research.

Bayes’ Theorem applied to probability distribution

- Distribution?
- Uncertainty?
- Exact vs Distribution

Naive Bayes

💡 Naive Bayes classifiers works differently in a way that it operates under a couple of key assumptions, earning it the title of “naive”. It assumes that **predictor** in a Naive Bayes models are **conditionally independent**, or **unrelated** to any of the **other features** in a model.

💡 It also assumes that all features **contribute equally** to the outcome.

While these assumptions are often violated in real-world scenarios (e.g. a subsequent word in an e-mail is dependent upon the word that precedes it, it simplifies a classification problem by making it more computationally tractable.

Example: Apply is red round and 2 cm diameter.

- If the features were independent of one another, we wouldn’t need complex models in deep learning and machine learning.
- But Naive Bayes works with this principal and gives good result.

Formula

$$\boxed{P\left(y \mid X\right) = \frac{P\left(X \mid y\right).\, P\left(y\right)}{P\left(X\right)}}$$

$$X = x_1, x_2, x_3, ..., x_n$$

$$\begin{aligned} P\left(y \mid X\right) &\implies Posterior\ Probability \\ P\left(X \mid y\right) &\implies Conditional\ Probability \\ P\left(y\right) &\implies Marginal\ Probability \\ P\left(X\right) &\implies Marginal\ Probability \end{aligned}$$

Assumption

$$P(y \mid x_1, \ldots, x_n) = \frac{P(x_1 \mid y)\, P(x_2 \mid y) \ldots P(x_n \mid y)\, P(y)}{P(x_1)\, P(x_2) \ldots P(x_n)}$$

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

$$P(x_1, \ldots, x_n) \textit{ is } \textbf{constant} \textit{ given the } \textbf{input}, \textit{ so}$$

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y), \quad \boxed{\textit{Here } \hat{y} \textit{ is the prediction}}$$

Types of Naive Bayes

💡 **Gaussian Naive Bayes (GaussianNB):** This is a variant of the Naive Bayes classifier, which is used with Gaussian Distributions - i.e. normal distributions - and continuous variables

💡 **Multinomial Naive Bayes (MultinomialNB):** This type of Naive Bayes classifier assumes that the features are from multinomial distributions.

💡 **Bernoulli Naive Bayes (BernoulliNB):** This is another variant of the Naive Bayes classifier, which is used with Boolean variables - that is,variables with two values, such as **True and False** or **1 and 0**

Naive Bayes in Action

Spam Classification

	Not Spam	Spam
Dear	8	3
Visit	2	6
Invitation	5	2
Link	2	7
Friend	6	1
Hello	5	4
Discount	0	8
Money	1	7
Click	2	9
Dinner	3	0
Total Words	34	47

$$P(Dear \mid Not\ Spam) = \frac{8}{34}$$

$$P(Visit \mid Not\ Spam) = \frac{2}{34}$$

$$P(Dear \mid Spam) = \frac{3}{47}$$

$$P(Visit \mid Spam) = \frac{3}{47}$$

Input: "Hello friend"

Features: Hello, friend

$$\begin{aligned} p(Not\ spam \mid Hello\ friend) &= Posterior\ Probability\ = ? \\ p(Hello\ friend \mid Not\ spam) &= Conditional\ Probability \\ p(Not\ spam) &= Prior\ Probability \\ p(Hello\ friend) &= Pobability\ of\ input \end{aligned}$$

$$Posterior\ Probability\ = \frac{Conditional\ probability\ *\ Prior\ Probability}{Probability\ of\ input}$$

Denomination Constant

$$p(Not\ spam \mid Hello\ friend)\ = p(Hello\ friend \mid Not\ spam)\ *\ p(Not\ spam)$$
$$\ldots$$
$$Here,\ p(Hello\ friend \mid Not\ spam) = 0 \qquad [There\ is\ no\ data\ given\ for\ "Hello\ friend"]$$

Probability of being Not Spam

$$\begin{aligned} p(Hello\ friend \mid Not\ spam) &= p(Hello \mid Not\ spam)\ *\ p(friend \mid Not\ spam) \\ p(Not\ spam \mid Hello\ friend) &= p(Hello \mid Not\ spam)\ *\ \ p(friend \mid Not\ spam)\ *\ p(Not\ spam) \\ p(Not\ spam \mid Hello\ friend) &= \frac{5}{34} * \frac{6}{34} * \frac{34}{81} = 0.0108 \end{aligned}$$

Probability of being Spam

$$\begin{aligned} p(Hello\ friend \mid spam) &= p(Hello \mid Spam)\ *\ p(friend \mid Spam) \\ p(Spam \mid Hello\ friend) &= p(Hello \mid Spam)\ *\ \ p(friend \mid Spam)\ *\ p(Spam) \\ p(Spam \mid Hello\ friend) &= \frac{4}{47} * \frac{1}{47} * \frac{47}{81} = 0.0493 \end{aligned}$$

Zero Frequency Problem

Input: Dear visit dinner money money money

Probability of being Not Spam

$$\begin{aligned} & p(Not\ spam \mid dear\ visit\ dinner\ money\ \ money\ \ money) \\ =\ & p(dear\ visit\ dinner\ money\ \ money\ money \mid Not\ spam)\ *\ p(Not\ spam) \end{aligned}$$
$$\ldots$$
$$\begin{aligned} \implies p(dear\ visit\ dinner\ money\ \ money\ money \mid Not\ spam)\ &= \frac{8}{34} * \frac{2}{34} * \frac{3}{34} * \left(\frac{1}{34}\right)^3 \\ &= 3.107 * 10^{-8} \end{aligned}$$
$$So,$$
$$\begin{aligned} p(Not\ spam \mid dear\ visit\ dinner\ money\ money\ money)\ &= 3.107 * 10^{-8} * \frac{34}{81} \\ &= 1.864 * 10^{-8} \end{aligned}$$

Probability of being Spam

$$\begin{aligned} &p(Spam \mid dear\ visit\ dinner\ money\ \ money\ \ money) \\ &= \ p(dear\ visit\ dinner\ money\ \ money\ money \mid Spam) \ * \ p(Spam) \\ &\ldots \\ \implies p(dear\ visit\ dinner\ money\ \ money\ money \mid Spam) \ &= \ \frac{3}{47} * \frac{6}{47} * 0 * \left(\frac{7}{47}\right)^3 \\ &= 0 \\ &So, \\ p(Spam \mid dear\ visit\ dinner\ money\ money\ money) \ &= \ 0 * \frac{47}{81} \\ &= 0 \end{aligned}$$

But this should be a spam e-mail as general sense. But the probability shows that its probability of being **spam** is 0. This is basically zero frequency problem.

Laplace Smoothing

💡 It is a technique for smoothing categorical data. A small-sample correction, or pseudo-count, will be incorporated in every probability estimate. Hence, no probability will be zero. This is a way of regularising Naive Bayes.

$$\hat{\theta} = \frac{x_i + \alpha}{N + \alpha d} \qquad\qquad (i = 1, ..., d)$$

$$\begin{aligned} &Here, \\ \hat{\theta} &\implies Final\ value\ of\ laplace\ smoothing \\ x_i &\implies Frequency\ of\ feature \\ N &\implies Total\ Frequency \\ \alpha &\implies Regularisation\ parameter \\ d &\implies Number\ of\ features \end{aligned}$$

Apply Regularisation

Probability of Not Spam

$$\begin{aligned} &p(Not\ spam \mid dear\ visit\ dinner\ money\ \ money\ \ money) \\ &= p(dear\ visit\ dinner\ money\ \ money\ money \mid Not\ spam)\ *\ p(Not\ spam) \\ &... \\ \implies p(dear\ visit\ dinner\ money\ \ money\ money \mid Not\ spam)\ &=\ \frac{8+1}{34+10} * \frac{2+1}{34+10} * \frac{3+1}{34+10} * \left(\frac{1+1}{34+10}\right)^3 \\ &= 1.19 * 10^{-7} \\ &So, \\ p(Not\ spam \mid dear\ visit\ dinner\ money\ money\ money)\ &=\ 1.19 * 10^{-7} * \frac{34}{81} \\ &= .4995 * 10^{-7} \end{aligned}$$

Probability of Spam

$$\begin{aligned} &p(Spam \mid dear\ visit\ dinner\ money\ \ money\ \ money) \\ &= p(dear\ visit\ dinner\ money\ \ money\ money \mid Spam)\ *\ p(Spam) \\ &... \\ &\implies p(dear\ visit\ dinner\ money\ \ money\ money\ |Spam) \\ &= \frac{3+1}{47+10} * \frac{6+1}{47+10} * \frac{0+1}{47+10} * \left(\frac{7+1}{47+10}\right)^3 \\ &= 4.18 * 10^{-7} \\ &So, \\ p(Spam \mid dear\ visit\ dinner\ money\ money\ money)\ &=\ 4.18 * 10^{-7} * \frac{47}{81} \\ &= .2425 * 10^{-6} \end{aligned}$$