Strength

* Simple and effective
* makes no assumptions about the data distribution
* fast training phase.

weakness

* does not produce a model limiting the ability to understand how features are related to the class.
* Requires selection of an appropriate k.
* nominal features and missing data requires additional processing.
* slow classification phase.

Probabilistic learning: Understanding Naive Bayes — Conditional probability and Bayes Theorem, Naive Bayes algorithm for classification, the Laplace estimator, Using numeric features with Naive Bayes.

① Understanding Naive Bayes

- Weather estimates are based on probabilistic methods or those concerned with describing uncertainity. They use data on past events to extrapolate future events.
- Mathematician Thomas Bayes, who developed foundational principles to describe the probability of events, and how probabilities should be revised in the light of additional information. These principles formed the foundation for what are now known as Bayesian methods.
- classifiers based on Bayesian methods utilize training data to calculate an observed probability of each outcome based on the evidence provided by feature values. When the classifier is later applied to unlabeled data, it uses the observed probabilities to predict the most likely class for the new feature.

→ Bayesian classifiers used for

* text classification, such as junk email (spam) filtering
* Intrusion or anomaly detection in computer networks.
* Diagnosing medical conditions given a set of observed symptoms.

- It suffices to say that a probability is a number between 0 and 1 (that is between 0 percent and 100 percent) which captures the chance that an event will occur in the light of the available evidence. The lower the probability, the less likely the event is to occur. The probability 0 that the event will definitely not occur, while probability 1 ⇒ the event will occur with 100 percent certainly.

Basic concept of Bayesian methods

- Bayesian probability theory is rooted in the idea that the estimated likelihood of an event or a potential outcome, should be based on the evidence at hand across multiple trials, or opportunities for the event to occur.

Event	Trial
Heads results	Coin flip
Rainy weather	A single day.
Message is Spam	Incoming e-mail msg.
Candidate becomes president	Presidential election
Win the lottery	lottery ticket.

Bayesian methods provide insights into how the probability of these events can be estimated from the observed data.

Understanding probability

- The probability of an event is estimated from the observed data by dividing the no of trials in which the event occured by the total no: of trials.

  Eg: $\frac{3}{10} = 0.30$ or 30%

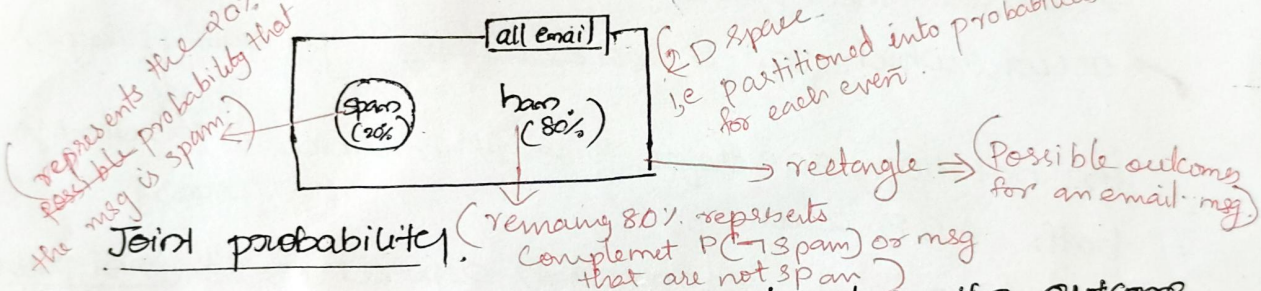
- To denote these probabilities, use notation $P(A)$, signifies the probability of event A Eg: $P(\text{rain}) = 0.30$.
- The probability of all possible outcomes of a trial must always sum to 1, because a trial always results in some outcome happening.

thus, if the trial has two outcomes that cannot occur simultaneously, such as raining versus sunny.

Eg: $P(\text{spam}) = 0.20$ $P(\text{ham}) = 1 - 0.20 = 0.80$

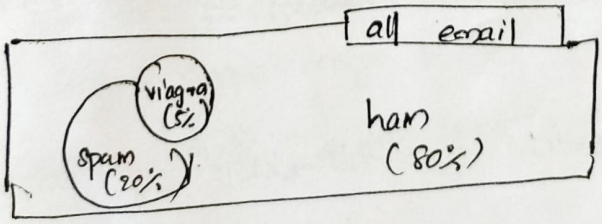
→ this concludes that spam & ham are mutually exclusive & exhaustive events, which implies that they cannot occur at the same time & are the only possible outcomes.

→ Bz an event cannot simultaneously happen and not happen, an event is always mutually exclusive and exhaustive with its complement. The complement of event A is typically denoted as A^c, A'. shorthand notation, $P(\neg A) \equiv P(A')$



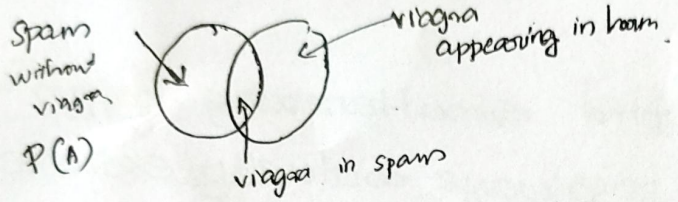
Joint probability:

→ for instance, a second event based on the outcome appears that an e-mail message contains the word viagra. In most cases, this word is likely to appear only in a spam message, its presence in an incoming e-mail is therefore a very strong piece of evidence that the message is spam.



(4) → diagram that the viagra circle does not fill the spam circle, nor is it completely contained the spam circle. This implies that not all spam messages contain the word viagra and not every e-mail with the word viagra is spam.

→ Venn diagram → a visualization,



→ 20% of all msgs were spam & 5% of all msgs contained the word viagra.

→ estimate the probability that both P(spam) and P(viag occur, which can be ~~estimated~~ written as P(spam ∩ viagra)

↓ intersection of events

the notation A ∩ B refers to event in which both A & B occur.

→ calculating P(spam ∩ viagra) depends on the joint probabili of two events. If the two events are totally unrelated they are called independent events., implies that knowing the outcome of one event does not provide any information about the outcome of the other. Eg. outcome of a heads result on a coin flip is independent from whether the weather is rainy or sunny on any given day.

→ dependent events are the basis of predictive modeling.

not contained crossed

→ independent events A & B the probability of both happening can be expressed as $P(A \cap B) = P(A) * P(B)$

Conditional probability with Bayes' theorem

→ the relationships between dependent events can be described using Bayes' theorem.

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

$P(A|B)$ → probability of event A, given that event B occured this is known as conditional probability

By definition, $P(A \cap B) = P(A|B) * P(B)$

Rearranging this formula that $P(A \cap B) = P(B \cap A)$ results in $P(A \cap B) = P(B|A) * P(A)$

formulation of Bayes' theorem.

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

$$P(\text{spam}|\text{viagra}) = \frac{P(\text{viagra}|\text{spam})\,P(\text{spam})}{P(\text{viagra})}$$

likelihood; prior probability

posterior probability; marginal likelihood.

prior probability → probability that any prior msg was spam. This estimate is known as the prior probability

likelihood → the probability that the word viagra was used in previous spam msgs. P(vi

marginal likelihood → probability that viagra appeared in any msg at all

6

→ To calculate these components of Bayes' theorem, it is necessary to construct a frequency table.

Frequency	Viagra: Yes	Viagra: No	Total
Spam	4	16	20
Ham	1	79	80
Total	5	100	100

(rows of likelihood table indicate the conditional probabilities for (yes/no) that email was either Spam or ham)

Likelihood	Viagra: Yes	Viagra: No	Total
Spam	4/20	16/20	20
Ham	1/80	79/80	80
Total	5/100	95/100	100

→ The frequency table can be used to construct a Likelihood table.

$P(\text{Viagra} = \text{Yes} | \text{spam}) = \frac{4}{20} = 0.20 \Rightarrow$ the probability of 20%, that a msg contains the term Viagra, gn that msg is spam.

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

$P(\text{spam} \cap \text{Viagra})$ as $P(\text{Viagra} | \text{Spam}) * P(\text{Spam}) = 4/20 * 20/100$

$= 0.04$

which notes that 4 out of 100 msgs were Spam with the term Viagra. Illustrates the importance of Bayes theorem while calculating conditional probability

To compute the posterior probability,

$P(\text{Spam} | \text{Viagra})$

$P(\text{Viagra} | \text{Spam}) * P(\text{Spam}) / P(\text{Viagra})$ or

$(4/20) * (20/100) / (5/100) = 0.80.$

Therefore the probability is 80% that a message is spam, given that it contains the word Viagra.

Classification with Naive Bayes

spam
ham

The Navie Bayes learner is trained by constructig a likelihood table for the appearance of these four words (labeled W_1, W_2, W_3, and W_4) for 100 e-mails.

Likelihood	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		Total
	Yes	No	Yes	No	Yes	No	Yes	No	
Spam	4/20	16/20	10/20	10/20	0/20	20/20	12/20	8/20	20
ham	1/80	79/80	14/80	66/80	8/80	71/80	23/80	57/80	80
Total	5/100	95/100	24/100	76/100	8/100	91/100	35/100	65/100	100

Problem: Suppose that a msg contains the terms Viagra and Unsubscribe, but does not contain either Money or Groceries.

Using Bayes theorem,

Gn that Viagra = Yes, Money = No, Grocery = No, Unsubscribe = Yes

we can define the problem using the formula.

$$P(\underset{A}{spam} \mid W_1 \cap \neg \underset{B}{W_2} \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg \underset{B}{W_2} \cap \underset{A}{\neg W_3} \cap W_4 \mid spam)\, P(\underset{A}{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

Assuming conditional independence allow us to simplify the formula using the probability rule for independent events, which states that

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

Conditional probability of spam expressed as.

$$P(spam \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \propto P(W_1 \mid spam)\, P(\neg W_2 \mid spam)\, P(\neg W_3 \mid spam)\, P(W_4 \mid spam)\, P(spam)$$

Probability that the message is ham can be expressed

$$P(\underset{A}{ham} \mid W_1 \cap \underset{B}{\neg W_2} \cap \neg W_3 \cap W_4) \propto P(W_1|ham) P(\neg W_2|ham)$$
$$P(\neg W_3|ham)\, P(W_4|ham)\, P(ham)$$

equals symbol has been replaced by the $\propto$ symbol to indicate the fact that the denominator has been omitted.

Using values in the likelihood table.

Overall likelihood of spam is.

$$(4/20)_{W_1} * (10/20)_{\neg W_2} * (20/20) * (12/20) * \overbrace{(20/100)}^{P(spam)} = 0.012$$

likelihood of ham is.

$$(1/80) * (66/80) * (71/80) * (23/80) * \overbrace{(80/100)}^{P(ham)} = 0.002$$

$$0.012/0.002 = 6 \checkmark$$

this msg is 6 times more likely to be Spam than ham.

To convert these numbers into probabilities.

$$0.012/(0.012+0.002) = 0.857.$$

[the probability of spam is equal to the likelihood that the message is spam divided by the likelihood that msg is either spam or ham]

Similarly. the probability of ham is equal to the likelihood that the msg is ham divided by the likelihood that msg is either Spam or ham.

$$0.002/(0.012+0.002) = 0.143$$

be expressed

message is Spam with 85.7 percent probability and ham with 14.3 percent probability. Because these are mutually exclusive and exhaustive events, the probabilities sum to 1.

finally Naive Bayes Classification algorithm we used in the preceding example can be summarized by.

probability of level L for class C, given the evidence provided by features F_1 through F_n, is equal to the product of the probabilities of each piece of evidence conditioned on the class level, the prior probability of the class level, and a scaling factor $1/Z$, which converts the likelihood values into probabilities

$$p(C_L|F_1,\ldots,F_n) = \frac{1}{Z} p(C_L) \prod_{i=1}^{n} p(F_i|C_L)$$

The Laplace estimator

Using the Naive Bayes algorithm as before, we can compute the likelihood of spam as:

$$(4/20) * (10/20) * (0/20) * (12/20) * (20/100) = 0$$

The likelihood of ham is

$$(1/80) * (14/80) * (8/80) * (23/80) * (80/100) = 0.00005$$

Therefore, the probability of spam is:

$$0/(0+0.00005) = 0$$

the probability of ham is

$$0.00005/(0+0.00005) = 1$$

These result suggest that the msg is spam with 0 percent probability and ham with 100% probability.

This prediction make no sense.

Probabilities in the Naive Bayes formula are multiplied in a chain, this 0 percent value causes the posterior probability of spam to be zero, giving the word groceries the ability to effectively nullify and overrule all of the other evidence.

- A solution to this problem involves using something called the Laplace estimator, which is named after the french mathematician Pierre-Simon Laplace. The Laplace estimator essentially adds a small number to each of the counts in the frequency table, which ensures that each feature has a nonzero probability of occuring with each class. Typically, the Laplace estimator is set to 1, which ensures that each class-feature combination is found in the data at least once.

Using a Laplace value of 1, we add one to each numerator in the likelihood function.

The Likelihood of spam is

$(5/24) * (11/24) * (1/24) * (13/24) * (20/100) = 0.0004$

The Likelihood of ham is

$(2/84) * (15/84) * (9/84) * (24/84) * (80/100) = 0.0001$

This means that the probability of spam is 80%, and the probability of ham is 20%.

Using numeric features with Naive Bayes

- Naive Bayes uses frequency tables to learn the data, each feature must be categorical in order to create the combinations of class and feature values comprising of the matrix.

Naive Bayes algorithm does not work with ~~numeric~~ numeric data.

One easy and effective solution is to discretize numeric features, which simply means that the numbers are put into categories known as bins

For this reason, discretization is also sometimes called binning. This method is ideal when there are large amounts of training data, a common condition while working with Naive Bayes

- There are several different ways to discretize a numeric feature. Perhaps the most common is to explore the data for natural categories or cut points in the distribution of data.

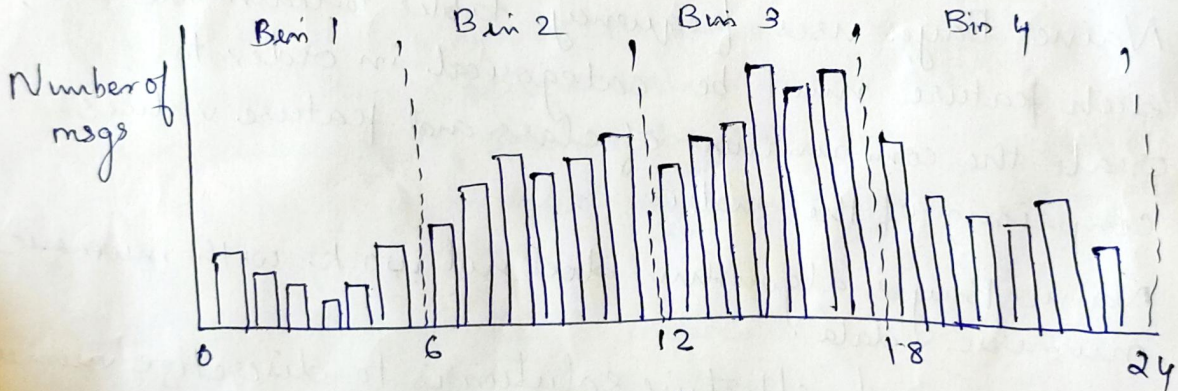
for eg: Suppose that you added a feature to the spam dataset that recorded the time of night or day the e-mail was sent, from 0 to 24 hrs past midnight.

Depicted using a histogram, the time data might look in the following diagram.

⇒ In the early hours of the morning, the msg frequency is low.

⇒ The activity picks up during business hours and tapers off in the evening.

⇒ This seems to create four natural bins of activity, as partitioned by the dashed lines indicating places where the numeric data are divided into levels of a new nominal feature,



Discretizing a numeric feature always results in a reduction of information as the feature's original granularity is reduced to a smaller number of categories.

Too few bins can result in important trends being obscured. Too many bins can result in small counts in the Naive Bayes frequency table, which can increase the algorithm's sensitivity to noisy data.