Supervised Learning Classification

Agenda

- Bayes theorem
- Business Problem
- Naïve Bayes

Naïve Bayes

Business problem: label the email as spam or ham

It is can be helpful if an algorithm is can label received emails as important (ham) emails or junk (spam) emails for an user.

Such model can ease the effort of the user by directly showing them the important emails and filter out the junk

Visiting Basics

Probability

- Probability is how likely an event is to occur

$$P = \frac{\text{No. of ways an event can occur}}{\text{Total possible events}}$$

- The probability of an event always lies in between 0 and 1

- 0 indicates impossibility of the event and 1 indicates a certain event

Probability

Question:

There are 40 candidates in a team with equal calibre. Out of which 25 are men and 15 are women. A person is randomly chosen to be the team leader. What is the probability that the person is a woman?

Probability

Solution:

Number of ways event can occur: 15

Total number of outcomes: 40

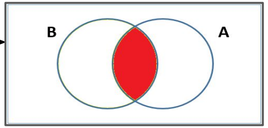
Therefore the probability: 15/40 = 0.375

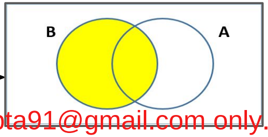
Conditional probability

- The **conditional probability of** an event A given B is the probability that the event A will occur given that an event B has already occurred

- Denoted by P(A|B)

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





Conditional probability

Question:

A pair of fair dice is rolled. If the sum of numbers that appear is 6, find the probability that one of the dice shows 2?

Conditional probability

Solution:

Let A: the event of getting the sum as 6

The ways A can occur: {(1,5), (2,4), (3,3), (4,2), (5,1)}

Let B: the event that number 2 appears on the dice

The ways B can occur: {(1,2), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,2), (4,2), (5,2), (6,2)}

First Dice

Second Dice	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Thus, the event that sum of the die is 6 and number 2 appears on the dice is A ∩ B.

A ∩ B = {(2,4), (4,2)}

The total number of samples is 36.

Conditional probability

Solution:

The required probability is

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

$$P\left(\text{getting a number as 2} \mid \text{getting the sum as 6}\right) = \frac{\text{P (getting the sum as 6 and a number as 2)}}{\text{P(getting the sum as 6)}}$$

$$P\left(\text{getting a number as 2} \mid \text{getting the sum as 6}\right) = \frac{\frac{2}{36}}{\frac{5}{36}} = \frac{2}{5} = 0.4$$

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Multiplication theorem

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

$$\Rightarrow P(A \cap B) = P(A \mid B).\, P(B)$$

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

$$\Rightarrow P(A \cap B) = P(B \mid A).\, P(A)$$

Thus, $P(A \cap B) = P(A \mid B).\, P(B) = P(B \mid A).\, P(A)$

Bayes theorem

- Conditional probability is the likelihood of an event given that another event has occurred

- Bayes theorem provides a way to updated the probability based on the new information

- It is completely based on the conditional probability

- Known as Bayes' Rule or Bayes law

Bayes theorem - formula

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

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

Where, A and B are events

- P(A | B) the likelihood of event A occurring given that B is true
- P(B | A) the likelihood of event B occurring given that A true
- P(A), P(B): The independent probabilities of A and B

Bayes theorem - formula

For the naïve bayes classification the formula is as

conditional probability of t given that the predictor x, i.e. posterior probability

probability of the class label, it is the prior probability

$$P(t \mid x) = \frac{P(t).P(x|t)}{P(x)}$$

conditional probability of x given that its class label is t, i.e. likelihood

probability of the value taken by the predictor variable, i.e. evidence

Posterior Probability

In context with a classification problem the posterior probability is the conditional probability of a class label taking value t given that the predictor takes value x

Example:

Consider the example of labelling an email as spam or ham. The conditional probability that it is a spam message given the word appears in it, i.e. P(spam | word) is the posterior probability

Prior probability

Prior probability is the probability of an event computed from the data at hand

Example:

Consider the example of labelling an email as spam or ham. The probability the email is spam, i.e. P(spam) is the prior probability
Likewise P(ham) is also a prior probability

Likelihood

In context with a classification problem the Likelihood is the conditional probability of a predictor taking value x given that its class label is t

Example:

Consider the example of labelling an email as spam or ham. The conditional probability that the word appears in a spam, i.e. P(word | spam) is the likelihood

Evidence

- It is the probability that the predictor takes value x
- Also known as marginal probability

Example:

Consider the example of labelling an email as spam or ham. The probability that the word appears in a message, i.e. P(word) is the evidence

Naïve bayes classification

- A naïve bayes classifier uses the the Bayes' theorem for classification
- It is an eager learning algorithm. Since it does not wait for test data to learn, it can classify the new instance faster

Assumptions

Assumption 1: The predictors are independent of each other.

Example:

Consider the example of labelling an email as spam or ham.
The probability of the word *Good* appearing in the email is independent of the *Money*.

Thus $P(Good \cap Money) = P(Good) . P(Money)$ … since events are independent

Assumptions

Assumption 2: All the predictors have an equal effect on the outcome.

Example:

Consider the example of labelling an email as spam or ham. The appearance of a particular word in the email does not have more importance in deciding whether it is a spam or ham

Eg: The word *Friendship* does not have more importance to say whether it's a spam/ham email.

Bayes theorem - classification problem

We have the Bayes theorem as

$$P(t \mid x) = \frac{P(t).P(x|t)}{P(x)}$$

For $X=(x_1, x_2, \ldots, x_n)$, applying the chain rule, we have

$$P(t \mid x_1, x_2, \ldots, x_n) = \frac{P(t).P(x_1|t).P(x_2|t)\ldots P(x_n|t)}{P(x_1)P(x_2)\ldots.P(x_n)}$$

Since the denominator does not change for the values taken by the predictor as assumed in the second assumption. The denominator can be removed.

Bayes theorem - classification problem

We get

$$P(t \mid x_1, x_2, ..., x_n) \propto P(t).\,P(x_1 \mid t).\,P(x_2 \mid t)...P(x_n \mid t)$$

For convenience, write it as

$$P(t \mid X) = P(t).\,P(x_1 \mid t).\,P(x_2 \mid t)...P(x_n \mid t)$$

Bayes theorem - classification problem

$$P(t \mid X) = P(t).\, P(x_1 \mid t).\, P(x_2 \mid t) \ldots P(x_n \mid t)$$

The class label with maximum probability gets assigned to the instance (x_1, x_2, …, x_n)

Naïve Bayes: Procedure

Obtain the frequency of the predictors

↓

Compute the likelihood of the predictors and obtain the prior probabilities based on the train data

↓

For an instance, compute the posterior probabilities for each the class labels

↓

Assign the most probable class label

Obtain the frequency of the predictors

↓

Compute the likelihood of the predictors and obtain the prior probabilities based on the train data

↓

For an instance, compute the posterior probabilities for each the class labels

↓

Assign the most probable class label

Business problem: label the email as spam or ham

- We shall consider the problem of labelling the received emails as spam or ham

- Choose a few words you find in emails

Work

Snacks

Good

Lonely

Horoscope

Money

Spam-ham example

1. Consider the frequency of these words used in spam and ham emails as shown below

	Spam	Ham
Good	2	10
Lonely	2	1
Horoscope	20	5
Work	5	12
Snacks	0	5
Money	21	7

Spam-ham example

2

	Spam	Ham
Good	2	10
Lonely	2	1
Horoscope	20	5
Work	5	12
Snacks	0	5
Money	21	7
Total	50	40

Obtain the likelihoods

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Likelihood

The probability that the word *Good* appears in a spam email, ie P(*Good* | Ham) is 0.25.

This is the Likelihood.

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Spam-ham example

3 Obtain the prior probability

From the data we have 15% of the emails are spam and the remaining are ham

Thus the prior probabilities are

P(Spam) = 0.15 and P(Ham) = 0.85

Spam-ham example

4

Consider the word sequence *Good Work*, does it belong to a spam message?

Our instance is *Good Work*.

For our instance, compute the posterior probabilities for each the class labels - spam or ham

Spam-ham example

4 Compute the posterior probabilities for each the class labels - Spam or Ham

For Spam,

$P(\text{Spam}|\ \textit{Good}, \textit{Work}) = P(\text{Spam})\ .\ P(\textit{Good}\ |\ \text{Spam}).\ P(\textit{Work}\ |\ \text{Spam})$
$= (0.15)\ .\ (0.04)\ .\ (0.1)$
$= 0.0006$

For Ham,

$P(\text{Ham}|\ \textit{Good}, \textit{Work}) = P(\text{Ham})\ .\ P(\textit{Good}\ |\ \text{Ham}).\ P\ (\textit{Work}\ |\ \text{Ham})$
$= (0.85)\ .\ (0.25)\ .\ (0.30)$
$= 0.063$

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Spam-ham example

5 Assign the most probable class label

For Spam,

$P(Spam| \textit{Good, Work}) = 0.0006$

For Ham,

$P(Ham| \textit{Good, Work}) = 0.063$

Since 0.063 > 0.0006, we assign the class label as Ham to the instance *Good Work*.

PLEASE!! NOTE

For Spam,

P(Spam| *Good, Work*) = 0.0006

For Ham,

P(Ham| *Good, Work*) = 0.063

$$\frac{\text{P(Spam| Good, Work)}}{\text{P(Spam| Good, Work)+P(Ham| Good, Work)}} = \frac{0.0006}{0.0006+0.063} = 0.009$$

$$\frac{\text{P(Ham| Good, Work)}}{\text{P(Spam| Good, Work)+P(Ham| Good, Work)}} = \frac{0.063}{0.0006+0.063} = 0.991$$

The sum is 1

Obtain the class label

Question:

With help of the previous data. Label the email containing word Horoscope 1 time, Money 2 times and Snack 1 time

The prior probabilities are:

$P(Spam) = 0.15$ and $P(Ham) = 0.85$

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Obtain the class label

Solution:

For Ham,

P(Ham| *Horoscope, Money, Money, Snack*)

= P(Ham) . P (*Horoscope* | Ham) . P(*Money* | Ham). P(*Money* | Ham) . P(*Snack* | Ham)

= (0.85) . (0.125) . (0.175) . (0.175) . (0.125)

= 0.0004

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Obtain the class label

Solution:

For Spam,

P(Spam| *Horoscope, Money, Money, Snack*)

= P(Spam) . P (*Horoscope* | Spam) . P(*Money* | Spam). P(*Money* | Spam) . P(*Snack* | Spam)

= (0.15) . (0.4) . (0.42) . (0.42) . (0.00)

… Here is a problem

= 0.0

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Obtain the class label

Solution:

For Spam,

P(Spam| *Horoscope, Money, Money, Snack*)

= P(Spam) . P (*Horoscope* | Spam) . P(*Money* | Spam). P(*Money* | Spam) . P(*Snack* | Spam)

= (0.15) . (0.4) . (0.42) . (0.42) . (0.00)

= 0.000

… No matter which other word(s) is seen along with Snack, the email will never be classified as Spam. Since the frequency for Snack is 0.

Laplace smoothing method

- To solve the zero probability problem we use Laplace smoothing method
- Add α to every count so the count is never zero
- $\alpha > 0$. Generally, $\alpha = 1$
- Consider the α for the divisor as well

Obtain the class label

Solution:

	Spam	Ham
Good	2	10
Lonely	2	1
Horoscope	20	5
Work	5	12
Snacks	0	5
Money	21	7
Total	50	40

Add $\alpha = 1$, to each count

	Spam	Ham
Good	3	11
Lonely	3	2
Horoscope	21	6
Work	6	13
Snacks	1	6
Money	22	8
Total	56	46

Obtain the class label

Solution:

	Spam	Ham
Good	3	11
Lonely	3	2
Horoscope	21	6
Work	6	13
Snacks	1	6
Money	22	8
Total	56	46

Obtain the new likelihoods

	Spam	Ham
Good	3/56 = 0.05	11/46 = 0.24
Lonely	3/56 = 0.05	2/46 = 0.04
Horoscope	21/56 = 0.37	6/46 = 0.13
Work	6/56 = 0.11	13/46 = 0.28
Snacks	1/56 = 0.02	6/46 = 0.13
Money	22/56= 0.40	8/46 = 0.18

Obtain the class label

Solution:

For Ham,

P(Ham| *Horoscope, Money, Money, Snack*)

= P(Ham) . P (*Horoscope* | Ham) . P(*Money* | Ham). P(*Money* | Ham) . P(*Snack* | Ham)

= (0.85) . (0.13) . (0.18) . (0.18) . (0.13)

= 0.0004

	Spam	Ham
Good	3/56 = 0.05	11/46 = 0.24
Lonely	3/56 = 0.05	2/46 = 0.04
Horoscope	21/56 = 0.37	6/46 = 0.13
Work	6/56 = 0.11	13/46 = 0.28
Snacks	1/56 = 0.02	6/46 = 0.13
Money	22/56= 0.40	8/46 = 0.18

Obtain the class label

Solution:

For Spam,

P(Spam| *Horoscope, Money, Money, Snack*)

= P(Spam) . P (*Horoscope* | Spam) . P(*Money* | Spam). P(*Money* | Spam) . P(*Snack* | Spam)

= (0.15) . (0.37) . (0.4) . (0.4) . (0.02)

= 0.0017

… The problem is solved using the Laplace smoothing method

	Spam	Ham
Good	3/56 = 0.05	11/46 = 0.24
Lonely	3/56 = 0.05	2/46 = 0.04
Horoscope	$\boxed{21/56 = 0.37}$	6/46 = 0.13
Work	6/56 = 0.11	13/46 = 0.28
Snacks	$\boxed{1/56 = 0.02}$	6/46 = 0.13
Money	$\boxed{22/56 = 0.40}$	8/46 = 0.18

Obtain the class label

Assign the most probable class label

For Spam,

P(Spam| *Horoscope, Money, Money, Snack*) = 0.0017

For Ham,

P(Ham| *Horoscope, Money, Money, Snack*) = 0.0004

Since 0.0017 > 0.0004, we assign the class label as Spam to the instance *Horoscope, Money, Money, Snack*.

did you know?

Naïve Bayes Classifier available in the scikit learn library

- Gaussian Naïve Bayes:
  - It is used when predictors are continuous
  - Assumes that the predictors follow normal distribution
  - The Gaussian Naïve Bayes Classifier used the normal distribution for classification

- Multinomial Naïve Bayes
  - Used for document classification problem - classify whether a document is a sports, history, or science article
  - The predictors are the frequency of the words present in the article

- Bernoulli Naïve Bayes
  - This is similar to the multinomial naive bayes but the predictors are binary valued (boolean)

Applications of Naïve Bayes

- Spam Filtering
- Sentiment Analysis
- Recommendation System

Naïve Bayes: advantages

- Easy to implement in the case of text analytics problems
- Used for multiple class prediction problems
- Performs better for categorical data than numeric data

Naïve Bayes: disadvantages

- Fails to find relationship among features
- May not perform when the data has more number of predictor
- The assumption of independence among features may not always hold good

Thank You