Machine Learning

Lesson 9

Probability & Bayes Learning

Lecturer: Dr. Msagha J Mbogholi, PhD

Flashback from Lesson 8

- Classification tasks can be further subdivided into four distinct groups: binary classification, multiclass classification, multilabel classification, and imbalanced classification.
- In binary classification each of the classes has to be assigned a label; the normal case is assigned a value 0, while the abnormal case is assigned the value 1.
- Systems which involve classification of more than two labels are referred to as multiclass systems.
- Reducing a multiclass problem to a binary problem is done using one vs all (OvA a.k.a. OvR) method.
- Multilabel classification refers to those problems involving two more labels for each example provided; they are solved using 3 techniques: Problem transformation, adapted algorithm and Ensemble approaches.
- Imbalanced classification involves an instance space where the examples are not equally distributed.

Content

- Introduction to Probability
- The Bayes theorem
- The classic golf game example
- Naïve Bayes classifiers
- Naïve Bayes application areas
- Advantages & disadvantages

Part 1

Introduction to Probability

1.1 Introduction

- There is a lot of mathematics and calculation in machine learning.
- In particular the areas of probability and statistics are used more than others; in fact I daresay most of the time these two topics go hand in hand.
- In this lesson we use a lot of probability theorem to explain the core of the lesson.
- However, for most students the word probability in itself is enough to cause a scare, and a quick 'jump' to the next lesson (I will survive without this one).
- Please do not do that; the lesson begins with the basics of probability and goes step by step till we get to apply it in machine learning.

1.2 Probability theory

- Probability is simply the likelihood of something happening; that something is usually an event.
- This likelihood is measured as a number that will range from 0 (it won't happen) to 1 (it will definitely happen).
- What about values in between like 0.2, 0.5, 0.75, and so on?
- These values indicate that the event can occur with a certain frequency if tried enough times.
- This is done by multiplying the frequency (an integer) with the probability value; this will give an indication of how many times the event will happen on average if it is tried that number of times (the frequency).
- Let us demonstrate this with an example.

1.2 Probability theory (cont'd)

- If an event has a probability (p) = 0.3, and we try it 50 times (this is the frequency) then it can happen 0.3 * 50 = 15 times out of the 50 times.
- Let us demonstrate this with 2 real life examples:
- We all watch football, cricket, basketball or some game that you like. If I say a team has a p = 0.5, then it means that in a league of say 38 games it will most likely win 19 games. I have used most likely since remember that probability is the likelihood of an event happening.
- Consider a stack of cards; there are normally 52 cards consisting of spades, diamonds, hearts and clubs. These are 4 distinct groups; so my probability of drawing a queen is 4/52 since there are 4 queens in the whole deck; this works out to 0.076. further this means that if I do this 52 times (frequency) then I will most likely draw all the 4 queens.

1.2 Probability theory (cont'd)

- So using this theory you can calculate the probability of anything happening simply by observation; you can observe how many times it happens over a certain number of observations of interest; for example in bank fraud (an area of immense interest) we can count the number of frauds over a certain number of transactions, or the number of people who got covid 19 in a certain population perhaps over a year, or cumulatively over 5 years.
- But based on the above examples this can take a lot of time and resources to collect, right?
- Hence the need to know about sampling. A sample is used to represent a population (did I not say earlier that probability and statistics usually go hand in hand?).
- By using a sample that is a small subset of the larger population of interest we can calculate probabilities which will be fairly accurate; thus it can be fairly stated that the results of the sample are representative of the population.

1.2 Probability theory (cont'd)

- This method can be used for both qualitative as well as quantitative measures; for example if we wished to know which car is the most popular in the country (don't always believe the motto "the car in front is always a Toyota") we will just pick a sample of car dealers from all over the country and get the numbers and models sold in the last quarter; based on this sample we can tell the number of units sold (quantitative) as well as which model sold the most (qualitative).
- We can extend this method in other areas such as polling; during election time (in most countries) there are organizations that will tell you that this party or that one will win the election based on a sample of the population; some will predict the candidates who are more likely to win based on the same samples (unfortunately this isn't an exact science in many countries as some voters will either not be truthful or will change their minds in the run up to the actual election).
- Sampling can also be used to check on the spread of a disease that appears to be affecting the population mortally, and so on. There are many applications of sampling.

1.3 Probability operation rules

- Addition (summation) – this can only happen in instances where the events are mutually exclusive, meaning they can not happen together (as opposed to mutually inclusive meaning they happen together). Let us use the deck of cards to demonstrate this; in the deck we have 4 groups of cards (spades, clubs, hearts and diamonds). This means that the probability of drawing a diamonds card is ¼ or 0.25. Similarly the probability of drawing a spades card is also ¼ or 0.25. Thus the probability of drawing a diamonds and spades is: p = 0.25 + 0.25 = 0.50
- Subtraction (difference/complement) – this is used in case where you wish to find the probability of an event different from the one that has already been computed; essentially this means that if we know the probability of something happening we can find the probability of it not happening. We do this since we know that when p = 1 it represents probability of the event happening. So the probability of it not happening is (1 – probability of it happening). For example the probability of drawing a diamonds card is 0.25; therefore, the probability of not drawing a diamonds card is: p = (1 – 0.25) = 0.75

1.3 Probability operation rules (cont'd)

- Multiplication – this is used to calculate the intersection of events that do not influence each other. The six sided dice is normally used to demonstrate this: to get the maximum score when playing a game using dice is a six. So how about when you are using 2 dice (like in monopoly game). What is the probability of getting 2 sixes from a throw? The answer is 1/6 (from the first dice) multiplied by 1/6 (from the second dice). This gives 1/36 or 0.028 (very low I know, but if you're lucky well….)
- Using the dice we can now use the learnt rules to calculate more complex situations:
- The probability of having 2 sixes : p = 1/6 * 1/6 = 1/36
- The probability of having a six on one dice and something other than a six on the second dice: p = 1/6 * (1 – 1/6) = 1/6 * 5/6 = 5/36
- The probability of having a six on the second dice and something other than a six on the first dice: p = 1/6 * (1 – 1/6) = 1/6 * 5/6 = 5/36

1.3 Probability operation rules (cont'd)

- How about the probability of getting at least one six from two thrown dice?
- The solution to this will apply all the rules we have learnt so far: there are two mutually exclusive events that apply here.
- There is (probability of getting two sixes on both dice) AND (probability of getting a six on the first dice OR a six on the second dice). This works out to:
- p = (1/6 * 1/6) + ((1/6 * 5/6) * (1/6 * 5/6)) = 0.305
- By understanding these rules it is much easier to know which rule to use and when; also it helps to understand the application of probability to machine learning as a whole (and it's a lot) and to the main topic of this lesson.

Part 2

The Bayes theorem

2.1 Introduction

- In lesson 8 we learnt about the different types of classification. We further learnt that classification is a supervised learning form of machine learning.
- When classifying unseen data the classifier will use the training it has received to classify the unseen instances.
- It will then classify the data as belonging to one of the classes. Further probabilistic classification can also be used.
- Probabilistic classification is a sub branch of classification; remember the email spam example we looked at? We stated that an email is either spam or not based on what the learner had learnt from the training data.
- Probabilistic classification algorithms will use an inferential way to determine the class of a given example.
- By using probability these types of algorithms will determine the best class for an example.

2.1 Introduction (cont'd)

- They do this by calculating the probability that the example belongs to each of the classes; then they pick the highest probability and assign the example to that class.
- Thus a classifier can be either probabilistic or non-probabilistic.
- The probabilistic classifiers are preferred in bigger machine learning tasks partly due to the fact that they can calculate the confidence value (probability) that can be related to a given class label.
- The Bayes algorithm developed from the Bayes theorem is such a classifier.

2.2 Bayes theorem

- Before describing the theorem let us describe some terms first.
- I am sure we're in agreement that probability in its general form is not an exact science. You can not apply the same probability in different situations.
- Consider the situation of a football league; we have determined that the probability of the team winning is 0.5; essentially this means that out of 10 games they will probably win 5.
- But this is dependent on so many factors; the weather (can be sunny, rainy, cold, warm, and so on), which players actually play in the 90 minutes, who the referee is, and so on….so many different factors.
- How do we account for these?

2.2 Bayes theorem (cont'd)

- Priori probability – this is a general probability. It is based on the assumption that it can be applied to every situation; for example if you throw a dice (in an ideal situation) then the priori probability of getting a head is 1/6 .
- Posteriori probability – this is a priori probability after something happened to change the original one. This means that something happened which changed the priori and thus it isn't valid anymore. For example, consider the priori that Real Madrid's probability of winning the national league is 0.5. This doesn't take into consideration the weather, who plays, the referee, conditions of the different stadiums they play in, and so on. Due to this a posteriori probability will be different from the priori one.
- Arising from this we can introduce the term conditional probability.
- A conditional probability denoted P(y | x) is the probability of y happening given that x has happened. Again for example the probability of Real Madrid winning the league (y) given that it has been a very rainy season (x); this will change the probability to another different figure (lower or higher).

2.2 Bayes theorem (cont'd)

- Based on observations a priori probability can be made more accurate; this makes the observations of the examples a very powerful tool to use in doing proper classification.
- The naïve Bayes algorithm is based on the evidence of the observations and it is based the theorem by Thomas Bayes which is named after him.
- A correct prediction is made based on the circumstances surrounding it.
- Let us now introduce the theorem and the notations associated with it.

2.2 Bayes theorem

- Bayes theorem states as follows:

$$P(B|E) = P(E|B) * P(B) / P(E)$$

- Where,
- P(B|E) is the probability of B (a belief or hypothesis) given evidence (E) .
- P(E|B) is the probability of the (observed) evidence given the hypothesis, in itself a conditional probability.
- P(B) is the priori probability of the belief
- P(E) is the priori probability of the (observed)evidence.
- Let us demystify the formula with a simple example.

2.2 Bayes theorem

- Consider the situation of an undergraduate student (U) transitioning to postgraduate (P); they could have had a first class degree or a second class degree. The chances of a first class transitioning is 50%.
- We add evidence that the student did either a computer science degree or an engineering degree; we determine that the probability of having done a computer science degree is 40% of the population.
- Observing only those who had first class degrees we discover 60% probability of them having done computer science degrees.
- Considering the priori probability of 50% we recognize a difference here.
- Thus we can determine an algorithm to find out whether a postgraduate student did a computer science or engineering degree.

2.2 Bayes theorem

- P(B|E) – the probability of the hypothesis given the evidence; in this case the probability that a student had a first class honors and the evidence is they did a computer science degree. Knowing the probability given the evidence will help us classify the grade (first class or second class) a student got.
- P(E|B) – the probability of having done computer science given the student had first class honours. This is a conditional probability which is 60%, which translates to 0.6
- P(B) – the general probability of having had a first class degree, which is 50% or 0.5
- P(E) – the general probability of having done a computer science degree which is 40% or 0.4
- Substituting these into the equation we have:

2.2 Bayes theorem

- P(B|E) = P(E|B) * P (B) / P(E)
- P(B|E) = (0.6 * 0.5) / 0.4 = 0.75 or 75%
- This is a very high probability and thus it means that most probably our student had a first class honors .
- To understand this further what we have shown is that even though the first class honors and second class honors are generally shared out equally among postgraduate students (priori probability) we find that the priori is influenced by the choice of undergraduate degree the student undertook, in this case computer science.

Part 3

The Golf game

3.1 The setting

- The Bayes theorem can not be discussed without visiting the example of the golf game.
- This example will be found across most literature and websites.
- I call it the classic golf game in naïve Bayes since almost every single book that discusses this algorithm will use it to explain the working of the algorithm.
- So whereas one reference has been used for this lesson nobody knows who originally came up with this!
- As such it has earned its place here.
- In this setting the table advises a player whether they can play golf based on certain weather conditions, i.e. outlook, temperature, humidity, and windy. Each of these variables give only binary results, i.e., yes or no. Table 1 shows the dataset depicting conditions and corresponding answers.

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Table 1. The golf game conditions (Gandhi, 2018)

3.2 Bayes application

- It is important to note that naïve Bayes classifiers is the name used by all those that use Bayes theorem in classification.
- They are called naïve since they assume that each input variable is independent.
- Further each input variable contributes equally to the outcome.
- The independence means that no input variable depends on any other.
- In table 1 the rows represent the examples (from 0 to 13)
- The columns (minus the last right one) represents the features/attributes/independent variables. These are outlook, temperature, humidity and windy.
- The extreme right column represents the output/outcome, which is whether to play golf (yes) or not (no).

3.3 Using the formula

- Let us interpret just two examples in table 1: rows 1 and 3.
- Row 1 – (rainy, hot , high, true) → No
- Row 3 – (sunny, mild, high, false) → Yes
- Row 1 can be read as: when its rainy, hot, high humidity and windy, don't play golf
- Row 3 can be read as: when its sunny, mild, high humidity and not windy, then you can play golf.
- Thus each example (row) can be read and interpreted in this way.
- Recall that we are using supervised learning and so this is our training dataset.
- We can now use the formula based on this understanding.

3.3 Using the formula

- P (B|E) = P(E|B) * P(B) / P(E)
- In our training dataset we substitute y for B (the outcome), and X for E (the observed values); thus X constitutes all the observed values in the dataset.
- We can now rewrite the formula going forward as:
- P (y|X) = P(X|y) * P(y) / P(X)
- In our example our outcome (y) is to play golf
- X represents the features (attributes/rows in the table) .
- X can be written in the form X = (x_1, x_2, x_3, ...x_n), the terms in parantheses represent the individual features; thus x_1 = outlook , x_2 = temperature, x_3 = humidity, x_4 = windy

3.2 Using the formula

- Thus substituting for X in the formula we have,
- $P(y \mid x_1, x_2, x_3, \ldots x_n) = \frac{P(x_1|y).\ P(x_2|y).\ P(x_3|y).\ P(x_4|y)\ldots.\ P(x_n|y).\ P(y)}{P(x_1).P(x_2)\ldots.P(x_n)}$

The values for each in the dataset can now be substituted accordingly.

Looking further we find that for each the value of the denominator will remain the same (constant).

We can thus modify our equation and introduce proportionality:

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

Fig 1. Proportionality formula (Gandhi, 2018)

3.3 Using the formula

- In this case the outcome is binary, the outcome is either yes or no.
- In cases where the outcome involves more than one variable (in which case it is called multivariate) then the output with the maximum probability is used; this is called the argmax function and is represented thus:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Fig 2 Argmax function (geekforgeeks.com)

- Using this formula the values can be computed.

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		**P(Yes)/P(No)**
Yes	9	9/14
No	5	5/14
Total	14	100%

Table 2 Probabilities calculation (geekforgeeks.com)

3.4 Implementation

- Table 2 shows values of probabilities that have already been computed.
- The table for example shows the following:
- The probability to play golf if the outlook is overcast is 4/9
- The probability to play golf if humidity is normal is 6/9
- The probability to play golf if the temperature is hot is 2/9
- Probability of playing golf : P (y) = P(play golf = yes) = 9/14
- Probability of playing golf: P (play golf = no) = 5/14
- These can be tested with unseen data to verify the accuracy of the classifier.

Part 4

Naive Bayes Classifiers

Introduction

- As discussed earlier the term naive Bayes algorithm does not refer to a specific algorithm; rather a family of algorithms that implement the Bayes theory.
- Normally the choice will depend on the nature of the task, the number of outcomes expected (whether binary or multivariate), the features, and so on.
- Therefore at times one algorithm will suffice, while in other times a combination will be required. Nonetheless regardless, the algorithms fall under 3 groups:
    - Multinomial naive Bayes
    - Gaussian naive Bayes
    - Bernoulli naive Bayes

4.1 Multinomial naive Bayes

- As the name implies this type is used in multinomial distributions.
- It is used where the outcome (independent variables) are in more than two categories, that is to say, more than 2 labels.
- This is unlike the golf dataset example where the outcome of playing golf was either yes or no.
- Suppose the outcome was yes, no and maybe (which is a probability in itself right); then a multinomial naive Bayes would have to be used.
- Due to this nature this type of naive Bayes is suitable in multiclass classification scenarios.
- Its most popular use is in documentation. It can be used to classify the document as belonging to one of diverse groups; this is done by looking at the type of text used in the document, the images and so on.
- Thus documents can be classified as either confidential, technology, arts, and so on.
- The classifier can also be used on websites where news articles are classified according to their content.

4.2 Gaussian naive Bayes

- A gaussian distribution was described in an earlier lesson.
- This type of naive Bayes is used for continuous variables. The algorithm assumes a normal distribution.
- In this case the continuous values of every attribute are assumed to be distributed as a normal distribution.
- The Gaussian distribution presents something of an easier task as you will only need to calculate mean and standard deviation of your training dataset.
- This means that these two (mean and standard deviation) will have to be stored for each feature (input variable) in the training dataset.

4.3 Bernoulli naive Bayes

- Remember the Bernoulli distribution? Yes.
- These are those distributions where the output is an "either – or".
- What this means is that you use this classifier where there are only two outputs; a yes and a no.
- Remember the playing golf example we have discussed in this lesson? That is a good example of a bernoulli distribution. The player will either play golf or not.
- Ideally these types of classifiers will be used in situations like the spam mail filter (spam or ham), text classification (is the document confidential or not), and so on.

Part 5

Naive Bayes application areas

Application areas

- Filing documents – the classifier here is used for document organization. It is used to organize and classify documents as to where they belong in say, an archive. This is even more effective in online systems and organizations that utilize cloud technology. In such scenarios instead of manually organizing documents (like most of us do with our google storage) a classifier can be used to store documents in a more logical and fashionable manner. This makes it easier to find them and retrieve them as the need arises. This is especially effective for news sites and other similar informational sites.
- Spam filters – the classifier can be used to sort and classify spam (as well as priority emails), and move them to the respective folders, making life much easier for the user. Thus users actually never get to see those spam emails and manually move them to the trash folder. If you have a gmail account those guys have made life even simpler. There is now a filter that separates your email into three folders: personal, social and promotional. So when I don't want to read any promotional email I just focus on the personal folders for more urgent and important matters. You never know what those guys will come up with next.

Application areas

- They are also used in sentiment analysis to determine the feelings of certain target groups. In this scenario an analysis can be done on positive and negative sentiments within the target group.
- They are also used in collaborative filtering (discussed in an earlier lesson) and recommender systems. These systems will determine whether a user will like a certain resource, like books or movies, or not.
- Ranking and classifying – they are used to rank pages, index scores according to relevancy, and thus classify data categorically.
- In biomedical informatics naive bayes has been used to predict Alzheimers disease from genome wide data.

Part 6

Advantages and disadvantages

Advantages

- Naïve Bayes is quite simple and easy to implement; a look at the examples covered in this lesson affirms this.
- Using a dataset such as the playing golf dataset is enough to determine our classifier; thus naïve Bayes does not require much training data.
- From the different types of naïve Bayes, it can be seen that this classifier can handle both discrete and continuous data.
- Like cloud computing technology, naïve Bayes is highly scalable with predictors and data points.
- Naïve Bayes is fast and therefore can be used to make real time predictions.
- When it comes to irrelevant features naïve Bayes is not sensitive.

Disadvantages

- As described earlier the algorithm is called naïve since it assumes the independence of features; however, this is rarely so in the real world. Thus its use in real world cases is limited.
- The algorithm assigns zero probability to any categorical variable in the unseen data that wasn't there in the training dataset, a problem called 'zero frequency'.
- It can be wrong in some cases, therefore this should be taken into account.

Summary

- By using a sample that is a small subset of the larger population of interest we can calculate probabilities which will be fairly accurate; thus it can be fairly stated that the results of the sample are representative of the population
- Probabilistic classification algorithms will use an inferential way to determine the class of a given example
- A conditional probability denoted P(y | x) is the probability of y happening given that x has happened.
- Bayes theorem states as follows: P (B|E) = P(E|B) * P(B) / P(E)
- It is important to note that naïve Bayes classifiers is the name used by all those that use Bayes theorem in classification. They are called naïve since they assume that each input variable is independent. Further each input variable contributes equally to the outcome. The independence means that no input variable depends on any other.
- All naïve Bayes algorithms fall under 3 groups: Multinomial naive Bayes, Gaussian naive Bayes and Bernoulli naive Bayes

References

- Aggarwal, C. C. (2014). *Data Classification: Algorithms and Applications*. Chapman & Hall/CRC.
- Brownlee, J. (2020, August 14). *Naive Bayes for machine learning*. Machine Learning Mastery. Retrieved June 4, 2022, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/
- Gandhi, R. (2018, May 17). *Naive Bayes classifier*. Medium. Retrieved June 4, 2022, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
- Mining, E. (2019). *Machine Learning for Beginners: A Complete and Phased Beginner's Guide to Learning and Understanding Machine Learning and Artificial Intelligence* . Independent.
- Mueller, J., & Massaron, L. (2016). *Machine learning for dummies*. John Wiley & Sons, Inc.

References

- *Naive Bayes classifiers*. GeeksforGeeks. (2022, February 2). Retrieved June 4, 2022, from https://www.geeksforgeeks.org/naive-bayes-classifiers/
- Shah, R. (n.d.). *Naïve Bayes algorithm's advantages and disadvantages: Data Science and Machine Learning*. Kaggle. Retrieved June 5, 2022, from https://www.kaggle.com/getting-started/225022
- Wei, W., Visweswaran, S., & Cooper, G. F. (2011). The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association : JAMIA*, *18*(4), 370–375. https://doi.org/10.1136/amiajnl-2011-000101
- Vadapilli, P. (2021, December 14). *Naive Bayes classifier: Pros & Cons, applications & types explained*. upGrad blog. Retrieved June 4, 2022, from https://www.upgrad.com/blog/naive-bayes-classifier/