Data Science Concepts

Lesson07-Naïve Bayes and Text Classification

Objective

After completing this lesson you will be able to:

- Understand Bayes Theorem
- Explain Naïve Bayes Classifier
- Describe the simplifying assumption of Naïve Bayes Classifier
- Explain the steps in building Naïve Bayes model

Naïve Bayes

One of the ways to assign an individual/object/data to a particular **class** is to use Naïve Bayes

Based on Bayes Theorem which provides away to calculate the probability of a **class** given prior knowledge of the problem.

Bayes Theorem

$$P(c|d) = \frac{P(d|c) * P(c)}{P(d)}$$

- $P(c|d)$ is the probability of class given the data d. This is called the posterior probability.
- $P(d|c)$ is the probability of data d given that the class to which it belongs was true.
- $P(c)$ is the probability of class c being true (regardless of the data). This is called the prior probability of class.
- P(d)is the probability of the data (regardless of the class).

$$Posterior\ Probability = \frac{conditional\ probability * prior\ probability}{evidence}$$

Naïve Bayes Classifier

Compute the posterior probability for a number of different class and select the class with the highest probability.

Objective function is:

$$Maximum\ a\ posterior\ \ class\ [MAP(c)] = \max[P(c|d)]$$

Simplifying assumption of Naïve Bayes

1. Each input value is assumed to be conditionally independent given the outcome variable

$$P(d_1, d_2, d_3|c) = P(d_1|c) * P(d_2|c) * P(d_3|c)$$

How likely it is to observe a particular pattern (d_1, d_2, d_3) given that it belongs to class c?

2. The samples are I.I.D (Independent and Identically Distributed)

It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each class are simplified by making the assumption that inputs are mutually independent.

Representation of Naïve Bayes Model

Class Probabilities: The probabilities of each class in the training dataset.

Conditional Probabilities: The conditional probabilities of each input value given each class value.

Steps in Naïve Bayes Model:

1. Compute the class probability from the data
2. Compute the conditional probability

Example

Renege data with 8995 observation across four variables. Below is a subset of data

DOJ Extended	Gender	Candidate Source	Status
Yes	Female	Agency	Joined
No	Male	Employee Referral	Joined
No	Male	Agency	Joined
No	Male	Employee Referral	Joined
Yes	Male	Employee Referral	Joined
Yes	Male	Employee Referral	Joined
Yes	Male	Employee Referral	Joined
Yes	Female	Direct	Joined
No	Female	Employee Referral	Joined
No	Male	Employee Referral	Joined
No	Male	Employee Referral	Not Joined
No	Male	Employee Referral	Joined
Yes	Male	Agency	Not Joined
No	Male	Direct	Not Joined
No	Male	Employee Referral	Not Joined
No	Male	Direct	Joined
Yes	Male	Agency	Not Joined

Probability Calculation

Calculate the class probability and conditional probability using frequency count

Class Probability	
Joined	Not Joined
0.813007	0.186993

Conditional Probabilities	**Probability**
P(DOJ Extension = Yes|Status = Joined)	0.469164502
P(DOJ Extension = No|Status = Joined)	0.530835498
P(DOJ Extension = Yes|Status = Not Joined)	0.461355529
P(DOJ Extension = No|Status = Not Joined)	0.538644471
P(Gender = Male|Status = Joined)	0.825242718
P(Gender = Female|Status = Joined)	0.174757282
P(Gender = Male|Status = Not Joined)	0.837693222
P(Gender = Female|Status = Not Joined)	0.162306778
P(Candidate Source = Agency|Status = Joined)	0.268015862
P(Candidate Source = Direct|Status = Joined)	0.538356352
P(Candidate Source = Employee Referal|Status = Joined)	0.193627786
P(Candidate Source = Agency|Status = Not Joined)	0.371581451
P(Candidate Source = Direct|Status = Not Joined)	0.513674197
P(Candidate Source = Employee Referal|Status = Not Joined)	0.114744352

Prediction Calculation

Look up the unique combination across three input variables

DOJ Extended	Gender	Candidate Source	Joined	Not Joined	Prediction
Yes	Female	Agency	0.017865506	0.02262147	Not Joined
No	Male	Employee Referral	0.068961031	0.009681516	Joined
No	Male	Agency	0.095454534	0.031352058	Joined
Yes	Male	Employee Referral	0.060949329	0.008292336	Joined
Yes	Female	Direct	0.035885969	0.007192584	Joined
No	Female	Employee Referral	0.014603512	0.001875837	Joined
Yes	Male	Agency	0.084364891	0.026853419	Joined
No	Male	Direct	0.19173699	0.043341085	Joined
Yes	Male	Direct	0.169461518	0.037122166	Joined
Yes	Female	Employee Referral	0.012906917	0.001606677	Joined
No	Female	Direct	0.040603127	0.008397528	Joined
No	Female	Agency	0.020213901	0.0060746	Joined

$$P(Joined|DOJ\ Extended, Gender, Candidate\ Source) = P(DOJ\ Extended|Joined) * P(Gender|Joined) * P(Candidate\ Source|Joined) * P(Joined)$$

Summing up

Tag the prediction for each observation in the dataset

DOJ Extended	Gender	Candidate Source	Status	Prediction
Yes	Female	Agency	Joined	Not Joined
No	Male	Employee Referral	Joined	Joined
No	Male	Agency	Joined	Joined
No	Male	Employee Referral	Joined	Joined
Yes	Male	Employee Referral	Joined	Joined
Yes	Male	Employee Referral	Joined	Joined
Yes	Male	Employee Referral	Joined	Joined
Yes	Female	Direct	Joined	Joined
No	Female	Employee Referral	Joined	Joined
No	Male	Employee Referral	Joined	Joined
No	Male	Employee Referral	Not Joined	Joined
No	Male	Employee Referral	Joined	Joined
Yes	Male	Agency	Not Joined	Joined
No	Male	Direct	Not Joined	Joined
No	Male	Employee Referral	Not Joined	Joined
No	Male	Direct	Joined	Joined
Yes	Male	Agency	Not Joined	Joined

Sentiment analysis and text Classification


FIND OUT IN CINEMAS TOMORROW

LUCKNOW CENTRAL

Follow

Lucknow Central
@LucknowCentral
The official handle of Lucknow Central, dir. by Ranjit Tiwari *ing Farhan Akhtar, Diana Penty, Gippy Grewal, Ronit Roy, Deepak Dobriyal. Releases 15 Sept…

Follow

Farhan Akhtar
@FarOutAkhtar
Writer. Director. Actor. Singer/Songwriter. Founder- @MardOfficial Initiative | @UN_Women Goodwill Ambassador(South Asia)

Shah Rukh Khan @iamsrk · 23h
Replying to @FarOutAkhtar
all the best for the film to the whole team. May #LucknowCentral be ur best one yet. Love to u.
486 1.6K 8.0K

Shivam Mehra @ShivamPaNDole · Sep 13
@FarOutAkhtar #LucknowCentral great movie,beautiful story and tremondous piece of acting on top of it all..absolute thriller👍👍
Solo release


Being Genius @Genius_003 · Sep 9
Watchd #PosterBoys last night. Not lying the movie is supr fun. KRK is an asshole. I think he shud be left alone with Sunny Deol for 2 mins
10 37 117

Mohnish Singh @mohnishmania · Sep 7
Just watched #PosterBoys. One of the best comedy films in recent times. A clean family entertainer with a social message. Superhit!!!
2 54 161

Rohit Jaiswal @rohitjswl01 · Sep 8
#PosterBoys is a Classic exmple of hw comedy films shud b made with realistic approach & Most Important just not Comedy but a Social msg too
3 85 257

Karan Johar @karanjohar · Sep 8
Big love to @thedeol on the release of #PosterBoys today! Preview news is that its a hilarious entertainer!!! Break a leg Bobby!!!! 😘😘❤️

Positive or negative movie review

What people are talking about in Dhoklam dispute

Surasen Goswami @Surasen3 · Sep 3

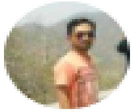
This dolt SuneetChopra shld apologz 2nation 4aligning with neighbor on #Dhoklam Shame! @republic shldnt bring hm on debate #Modi2019Cabinet

1 1 3

Adv Sanwar Khan @AskSanwar · Aug 29

Inconvenient Truth & Convenient Lies Of Modi/BJP, Chinese Troops enter Laddakh after #Dhoklam @drajoykumar @JhaSanjay @AshokTanwar_INC

2 2



rupam ray @rupamray · Aug 30

Best article so far on #dhoklam

> **ORF** @orfonline
>
> #Doklam standoff has revealed chinks in China's armour which was supposed to be impregnable writes Ashok Sajjanhar: goo.gl/zuS7RQ

1

What is the subject of this article


Usefulness of Systematic Review Search Strategies in Finding Child Health Systematic Reviews in MEDLINE

MEDLINE

Contains journal citations and abstracts for biomedical literature from around the world.

Medical Subject Headings (MeSH)

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology

Digital Media and sentiment analysis

- 70% of the sales decisions were made before engaging a sales representative (HP).
- Sales force is becoming irrelevant since customers are engaged through social media.

Classification in unstructured data

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language identification
- Sentiment analysis

Naïve Bayes as one of the algorithms

Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Rule based classification
  - Naïve Bayes
  - Support Vector Machines
  - Hidden Markov Models
  - Heuristics

Classification Methods: Rules

- Rules based on combinations of words present in the documents.

- Spam e-mail classification
  - ("Hello Dear") OR ("You Won a Lottery")

  - ("Soulmate waiting for you")

  - Accuracy can be low

Naïve Bayes for Text Classification

Text Classification

Assume a task of classifying a text as Spam or not spam (ham)

How to convert a text document as a set of features/attributes

- Feature should be important and meaningful (Salient)
- Feature should have enough information to demarcate between different classes (Discriminatory)
- Feature should not be prone to distortion, scaling orientation (Invariant; applicable to image)

Bag of word models – Commonly used model in NLP

1. Create vocabulary – Collection of different words (and its count) which appear in training set.
2. Tokenization – breaking the text corpus into individual elements followed by:
   - Removal of stop words
   - Removal of punctuation characters
   - Stemming
   - Lemmatizing
   - Construction of n-grams

Create Vocabulary

- $D1: Hi\ there, you\ have\ won\ the\ lottery\ prize.$
- $D2: Hi\ there, hope\ this\ mail\ finds\ you\ in\ good\ spirit.$

$$V = \{hi: 2, there\ 2: , you: 2, have: 1, won: 1, the: 1, \\ lottery, : 1\ prize: 1, hope: 1, this: 1, mail: 1, finds: 1, \\ in: 1, good: 1, spirit: 1\}$$

15 different words in the vocabulary can be used to create 15-dimensional feature vector

In general, the vocabulary can be used to construct d-dimensional feature ($|V|$ vectors for the individual documents.   This is called as vectorization of documents.

Bag of word representation for two sample document is:

	hi	there	you	have	won	the	lottery	prize	hope	this	mail	finds	in	good	spirit
t_{D1}	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
t_{D2}	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1
Sum	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1

Each word may be termed as a token. The count can be binary or absolute?

Bernoulli Naïve Bayes or Multinomial Naïve Bayes

Bag of word assumptions

Position of words in the document does not matter

Feature probabilities are independent given the class c

$$P(x_1, x_2, x_3|c) = P(x_1|c) * P(x_2|c) * P(x_3|c)$$

Is training data so well written!!!

Chennai Express from Bollywood Hungama

what an awseome movie....start fr the DDLJ train scene..till the climax...too funny...conversation with singing a song..haha mind blowing... Mina washing powder mina...mina..ting tong...halirous man...SRK you proof once again that you are a Baap of acting and bollwood...

i like the movie v.much. its train secquence srk acting in train, converstion betwn SRK and depica by singing hindi songs it is amaizing.its drama,its music and its climax is v.good ,proud of such a lover rahol..Ce has good entertainment story love to watch it again & again.

wow mind blowing...superb.. truly enjoy to watch this. CE rocksssssssssssssss

The bag of words representation

$\gamma($

Some films are hard to make sense of. Others are just nonsense. Chennai Express, directed by Rohit Shetty, ticks both boxes. More than a quarter of the film is in Tamil, and hence incomprehensible if you're unfamiliar with the language. The rest is a stew of puerile humor, lazy stereotypes, and way-over-the-top acting from a star who appears to be trying too hard.

Chennai Express plays neither to Rohit's strengths nor to Shah Rukh's. It's a strangely sloppy mishmash of cheesy humour, half-hearted romance, half-baked emotion and head-banging action. The film is filled with gigantic men whose size functions as a punch line. Yes, some of it is funny. The locations are beautiful.

$)=c$

Tokenization - Intuition

Y(

~~Some films are hard to make~~ **sense** ~~of. Others are just~~ **nonsense**. ~~Chennai Express, directed by Rohit Shetty, ticks both boxes. More than a quarter of the film is in Tamil, and hence~~ **incomprehensible** ~~if you're~~ **unfamiliar** ~~with the language. The rest is a stew of~~ **puerile humor, lazy stereotypes,** ~~and way-over-the-top acting from a star who appears to be trying too~~ **hard.**

~~Chennai Express plays neither to Rohit's strengths nor to Shah Rukh's. It's a strangely~~ **sloppy mishmash** ~~of~~ **cheesy humour, half-hearted romance, half-baked emotion** ~~and head-banging~~ **action** ~~The film is filled with~~ **gigantic** ~~men whose size functions as a punch line. Yes, some of it is funny. The locations are~~ **beautiful.**

)=c

Vocabulary after Tokenization

$$\Upsilon\left(\begin{array}{ll} \textbf{Vocabulary} & \textbf{Count} \\ \hline \text{Sense} & 1 \\ \text{Nonsense} & 1 \\ \text{Incomprehensible} & 1 \\ \text{Unfamiliar} & 1 \\ \ldots\ldots & \end{array}\right) = c$$

Tokenization

d-dimensional feature vector will have lot of redundancy. Tokenization is pre-processing to remove un-informative and useless texts from the vocabulary

Breaking down text corpus to individual elements that serves as input to NLP algorithm

$D1: Hi\ there, you\ have\ been\ winning\ the\ lottery\ prize\ for\ sometime\ now.$

hi	there	you	have	been	winning	the	lottery	prize	for	sometime	now

Small letters and removed punctuation

The above way of breaking down into individual element/tokens is called creating unigram (each word on its own)

Tokenization–Stop word removal

Words which are common in text corpus and considered un-informative

hi	there	you	have	been	winning	the	lottery	prize	for	sometime	now

Stop words are context specific. The one highlighted in blue above is stop word list while analyzing an English article

How to identify stop words:

- search against a language-specific stop word dictionary.
- create a stop list by sorting all words in the entire text corpus by frequency.

http://www.ranks.nl/stopwords

Stemming

Process of transforming a word into its root form

hi	there	you	have	been	winning	the	lottery	prize	for	sometime	now

“winning” got converted to “win”.

Stemming at times may give incorrect word

“thus” will get converted to “thu”

The original stemming algorithm was developed my Martin F. Porter in 1979 and is hence known as Porter stemmer

Lemmatization

Lemmatization aims to obtain the grammatically correct forms of the words, the so-called lemmas.

D3: A swimmer likes swimming thus he swims

a	swimmer	likes	swimming	thus	he	swims

The one highlighted in blue will be deleted.

Lemmatization is computationally more difficult and expensive than stemming, and in practice, both stemming and lemmatization have little impact on the performance of text classification

N–gram

A token can be defined as a sequence of n items (called n-grams).

$$D1: Hi\ there, you\ have\ been\ winning\ the\ lottery\ prize\ for\ sometime\ now$$

hi	there	you	have	been	winning	the	lottery	prize	for	sometime	now

hi there	you have	been winning	the lottery	prize for	sometime now

Choosing the optimal number n depends on the language as well as the particular application.

Vocabulary–after tokenization

- $D1: Hi\ there, you\ have\ won\ the\ lottery\ prize.$
- $D2: Hi\ there, Hope\ this\ mail\ finds\ you\ in\ good\ spirit.$

$$V = \{hi: 2, won: 1, lottery: 1\ prize: 1, hope: 1, mail: 1, find: 1, good: 1, spirit: 1\}$$

8 different words in the vocabulary can be used to create 8-dimensional feature vector

Vocabulary–after tokenization

Bag of word representation for two sample document is:

	hi	won	lottery	prize	hope	mail	finds	good	spirit	spam
t_{D1}	1	1	1	1	0	0	0	0	0	yes
t_{D2}	1	0	0	0	1	1	1	1	1	no

We can use the raw count of each words in the documents to fill the values for the above feature.

Term frequency document

The term frequency is defined as the number of times a given term t (word/token) appears in a document d.

	hi	won	lottery	prize	hope	mail	finds	good	spirit	spam
t_{D1}	1	1	1	1	0	0	0	0	0	yes
t_{D2}	1	0	0	0	1	1	1	1	1	no

In this case, term frequency is same as binary count

Class conditional probability from term frequency

$$Normalized\ term\ frequnecy = \frac{tf(t,d)}{n_d}$$

- $tf(t,d) = count\ of\ term\ t\ in\ doucment\ d$
- $n_d = the\ total\ number\ terms\ in\ document\ d$

Term frequency can be used to compute maximum likelihood estimate or the class conditional probabilities from the training data

$$P(t_i|c_j) = \frac{\sum tf(t_i, d \in c_j)}{\sum N_{d \in c_j}}$$

- $t_i = a\ word\ or\ token\ from\ feature\ vector\ T\ of\ a\ particular\ document$
- $\sum tf(t_i, d \in c_j) = sum\ of\ count\ of\ term(word)\ t_i\ from\ all\ the\ document\ in\ trainig\ set\ which\ belows\ to\ class\ c_j$
- $\sum N_{d \in c_j} = sum\ of\ all\ term\ frequencies\ in\ the\ training\ dataset\ for\ class\ c_j$

Class conditional probability for the text

The class-conditional probability of encountering the text **T** can be calculated as the product from the likelihoods of the individual words

$$\boldsymbol{P(T|c_j) = P(t_1|c_j) * P(t_2|c_j) * P(t_3|c_j) \ldots . P(t_n|c_j)}$$

$$= \prod_{i=1}^{n} P(t_i|c_j)$$

Naïve assumption of conditional independence: class-conditional probability of encountering the text **T** can be calculated as the product from the likelihoods of the individual words

Problem with maximum likelihood estimate

We have seen no training documents with the word “thalaiva” and classified it in the class super hit movie?

$$P(thaliva|super\ hit) = \frac{count\ of(thaliva,\ \ d \in super\ hit)}{Count\ of\ all\ the\ terms\ in\ d \in superhit} = 0$$

Problem with maximum likelihood estimate

Zero probabili1es cannot be conditioned away, no matter the other evidence

$$C_{max} = Max(P(c_j) * \prod_{i=1}^{n} P(t_i|c_j))$$

Laplace smoothing for Naïve Bayes

$$P(t_i|c_j) = \frac{\sum tf(t_i, d \in c_j) + 1}{\sum(N_{d \in c_j} + 1)} = \frac{\sum tf(t_i, d \in c_j) + 1}{\sum(N_{d \in c_j} + |V|}$$

$$|V| = dimension\ of\ the\ feature\ vector$$

Term frequency–Inverse document frequency

Another way to characterize text is to use weighted term frequency, which is especially useful if stop words have not been removed from the text corpus.

The $Tf - idf$ approach assumes that the importance of a word is inversely proportional to how often it occurs across all documents.

$$Tf - idf = tf_n(t,d) * idf(t)$$

$$tf_n(t,d) = \frac{tf(t,d)}{n_d}$$

$$idf(t) = \log(n_d / n_d(t))$$

- $n_d = total\ number\ of\ documents$
- $n_d(t) =$ number of documents which contains term t
- $tf_n(t,d) = normalized\ term\ frequency$

Putting All Together

	Doc	Word	Class
Training	1	India Bangalore India	I
	2	India India New Delhi	I
	3	India Bangalore	I
	4	China Tokyo Japan	J
Test	5	India India India Tokyo Japan	?

- $P(c) = \frac{N_c}{N}$
- $P(t|c) = \frac{[count\ (t,c)+1]}{count(c)+|V|}$

$$V = \{India: 4, Bangalore: 2, New\ Delhi: 1, Tokyo: 1, Japan: 1, China: 1\}$$

$|V| = 6$

$P(I) = \frac{3}{4}; P(J) = \frac{1}{4}$

Putting All Together

	Doc	Word	Class
Training	1	India Bangalore India	I
	2	India India New Delhi	I
	3	India Bangalore	I
	4	China Tokyo Japan	J
Test	5	India India India Tokyo Japan	?

$$P(India|I) = \frac{6}{14} = \frac{3}{7}$$

$$P(Bangalore|I) = \frac{3}{14}$$

$$P(New\ Delhi|I) = \frac{2}{14}$$

$$P(Tokyo|I) = \frac{1}{14}$$

$$P(Japan|I) = \frac{1}{14}$$

$$P(China|I) = \frac{1}{14}$$

$$P(India|J) = \frac{1}{9}$$

$$P(Bangalore|J) = \frac{1}{9}$$

$$P(New\ Delhi|J) = \frac{1}{9}$$

$$P(Tokyo|J) = \frac{1}{9}$$

$$P(Japan|J) = \frac{1}{9}$$

$$P(China|J) = \frac{1}{9}$$

Putting All Together

	Doc	Word	Class
Training	1	India Bangalore India	I
	2	India India New Delhi	I
	3	India Bangalore	I
	4	China Tokyo Japan	J
Test	5	India India India Tokyo Japan	?

$$P(I|Doc_5) = \frac{3}{4} * \left(\frac{3}{7}\right)^3 * \left(\frac{1}{14}\right) * \left(\frac{1}{14}\right) \approx 0.003$$

$$P(J|Doc_5) = \frac{1}{4} * \left(\frac{1}{9}\right)^3 * \left(\frac{1}{9}\right) * \left(\frac{1}{9}\right) \approx 0.000004$$

$$Thus\ class\ of\ Doc_5 = India$$

Challenges with Text Classification

Extremely difficult to make a model/algorithm/computer understand this language

- All my life I thought Air was free… until I bought a bag of chips.
- Everyone wants your best! Don't let them take it away from you.
- You have been so incredibly helpful and thanks (for nothing)

Summary–Naïve Bayes is not so Naïve

- Very fast with low storage requirements
- Robust to Irrelevant Features
  - Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features
  - Decision Trees suffer from fragmentation in such cases –especially with less data
- Optimal if the independence assumptions hold:
  - If assumed independence is correct, then it is the Bayes Optimal Classifier
- A good dependable baseline for text classification

End of Lesson07–Naïve Bayes and Text Classification