Naïve Bayes Classifier

Example Problem

greatlearning

- A Path Lab is performing a Test of disease say “D” with two results “Positive” & “Negative”
- They guarantee that their test result is 99% accurate:
  - If patient has the disease, they will give test positive 99% of the time.
  - If patient don’t have the disease, they will test negative 99% of the time.
- It is given that 3% of all the people have this disease
- New patient goes to Path Lab for test and his test gives “positive” result
- **What is the probability that New Patient actually have the disease?**
- **How doctors can use this information in their inferences?**

Introduction

- Naive Bayes classifier is a straightforward and powerful algorithm for the classification task
- Even if we are working on a data set with millions of records with some attributes, it is suggested to try Naive Bayes approach
- Naive Bayes classifier gives great results when we use it for textual data analysis. Such as Natural Language Processing
- To understand the naive Bayes classifier we need to understand the Bayes theorem. So let's first discuss the Bayes Theorem

Bayes Theorem

greatlearning

- Bayes Theorem works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred.
- Using the conditional probability, we can calculate the probability of an event using its prior knowledge.
- Conditional probability: $P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$
  - Where
    - P(H): The probability of hypothesis H being true. This is known as prior probability
    - P(E): The probability of the evidence (regardless of the evidence)
    - P(E|H): The probability of the evidence given that hypothesis is true
    - P(H|E): The probability of the hypothesis given that the evidence is true

Recall ‘Example Problem’

greatlearning

- A Path Lab is performing a Test of disease say “D” with two results “Positive” & “Negative”
- They guarantee that their test result is 99% accurate:
  - If patient has the disease, they will give test positive 99% of the time.
  - If patient don’t have the disease, they will test negative 99% of the time.
- It is given that 3% of all the people have this disease
- New patient goes to Path Lab for test and his test gives “positive” result
- **What is the probability that New Patient actually have the disease?**
- **How doctors can use this information in their inferences?**

Answering the problem

greatlearning

- We use conditional probability
- Probability of people suffering from Disease D, P(D) = 0.03 = 3%
- Probability that test gives 'positive' result and patient has the disease: P(Positive | Disease) = 0.99 = 99%
- Probability of people not suffering from disease: P(~Disease) = 0.97 = 97%
- Probability that test gives 'positive' result and patient does not have the disease, P(Positive | ~ Disease) = 0.01 = 1%
- The probability that the patient actually have the disease:

  $$P(Disease\ |Postive\ ) = \frac{P(Positive\ |Disease\ ) * P(Disease)}{P(Positive)}$$
- We have all the values of the previous equation except P(Positive)

Answering the problem

greatlearning

- $P(Positive) = P(D \cap Positive\ ) + P((\ {\sim}D) \cap Positive))$
- $P(Positive) = P(Positive \mid D\ ) * P(D) + P(Positive \mid ({\sim}D)\ ) * P({\sim}D)$
- $P(Positive) = 0.99 * 0.03 + 0.01 * 0.97$
- $P(Positive) = 0.0297 + 0.0097$
- $P(Positive) = 0.0394$

Answering the problem

- $P(Disease \mid Positive) = \frac{P(Positive \mid Disease) * P(Disease)}{P(Positive)}$
- $P(Disease \mid Positive) = \frac{0.99 * 0.03}{0.0394}$
- $P(Disease \mid Positive) = 0.753807107$
- **So, approximately 75% chances are there that the patient is actually suffering from disease**

Example Problem 2

- Test Data
  - Approximately 0.1% are infected
  - Test detects all infections
  - Test reports positive for 1% healthy people

Probability of having AIDS if test is positive

- Let A be the event that person has AIDS,
  - $P(A)$ = 0.1%= 0.001, $P(A^c)$ = (100-0.1)% $\equiv$ 0.9999
- Let T be the event that test is positive
  - For healthy people, $P(T|A^c)$ = 1% $\equiv$ 0.01
  - For infected people, $P(T|A)$ = 100% $\equiv$ 1.0
- We want to find $\boldsymbol{P(A|T)}$

Example Problem 2

- Test Data
  - Approximately 0.1% are infected
  - Test detects all infections
  - Test reports positive for 1% healthy people

Probability of having AIDS if test is positive

$$P(A|T) = \frac{P(T|A)P(A)}{P(T)}$$

$$= \frac{P(T|A)P(A)}{P((T \cap A) \cup (T \cap A^c))}$$

$$= \frac{P(T|A)P(A)}{P(T|A)P(A) + P(T|A^c)P(A^c)}$$

$$= \frac{1\ X\ 0.001}{1\ X\ 0.001 + 0.01\ X\ 0.9999} = \mathbf{0.091}$$

Just 9%!!!

Example Problem 2

- Use a follow-up test!
  - Test 2 reports positive for 90% infections
  - Test 2 reports positive for 5% healthy people

Probability of having AIDS if test 1 and test 2 is positive

- Let A be the event that person has AIDS,
  - $P(A)$ = 0.1% $\equiv$ 0.001, $P(A^c)$ = (100-0.1)%= 99.99% $\equiv$ 0.9999
- Let T_1 be the event that test 1 is positive
  - For healthy people, $P(T_1|A^c)$ = 1% $\equiv$ 0.01
  - For infected people, $P(T|A)$ = 100% $\equiv$ 1.0
- Let T_2 be the event that test 2 is positive
  - For healthy people, $P(T_2|A^c)$ = 5 % $\equiv$ 0.05
  - For infected people, $P(T_2|A)$ = 90% $\equiv$ 0.90

Example Problem 2

- Use a follow-up test!
  - Test 2 reports positive for 90% infections
  - Test 2 reports positive for 5% healthy people

Also test $\mathrm{T}_1, \mathrm{T}_2$ *are independent,* $P(T_1 \cap T_2 | A) = P(T_1 | A) P(T_2 | A)$

$$P(A^c | T_1 \cap T_2) = \frac{P(T_1 \cap T_2 | A^c) P(A^c)}{P(T_1 \cap T_2)}$$

$$= \frac{P(T_1 \cap T_2 | A^c) P(A^c)}{P(T_1 \cap T_2 | A) P(A) + P(T_1 \cap T_2 | A^c) P(A^c)}$$

$$= \frac{0.01 \text{ X } 0.05 \text{ X } 0.999}{1\text{X}0.9\text{X}0.005 + 0.01 \text{ X } 0.05 \text{ X } 0.999} = 0.357$$

$P(A | T_1 \cap T_2) = 1 - P(A^c | T_1 \cap T_2)$ = ***0.643***

Naïve Bayes Classifier

- Naive Bayes is a kind of classifier which uses the Bayes Theorem.
- It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class.
- The class with the highest probability is considered as the most likely class.
- This is also known as Maximum A Posteriori (MAP).

- The MAP for a hypothesis is:
  - $MAP(H) = \max\big(P(H|E)\big)$
  - $MAP(H) = \max\big((P(E|H) * P(H)\big) / P(E))$
  - $MAP(H) = \max\big(P(E|H) * P(H)\big)$
  - $P(E)$ is evidence probability, and it is used to normalize the result. Result will not be effected by removing $P(E)$

Naïve Bayes Classifier

- Naive Bayes classifier assumes that all the features are unrelated to each other
- Presence or absence of a feature does not influence the presence or absence of any other feature
- Example:
  - A fruit may be considered to be an apple if it is red, round, and about 4" in diameter.
  - Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple
- In real datasets, we test a hypothesis given multiple evidence(feature).
- So, calculations become complicated.
- To simplify the work, the feature independence approach is used to 'uncouple' multiple evidence and treat each as an independent one.
- $P(H|Multiple\ Evidences) = \frac{P(E_1|H) * P(E_2|H) * \cdots * P(E_n|\ H) * P(H)}{P(Multiple\ Evidences)}$

Let’s understand Naïve Bayes Classification

greatlearning

Let’s Understand Through an Example: Play Badminton Data

Day	Outlook	Temperature	Humidity	Wind	Play Badminton
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Question: For the day <sunny, cool, high, strong>, what’s the play prediction?

The Model Training Phase

For four external factors, we calculate for each we calculate the conditional probability 's table

Day	Outlook	Temperature	Humidity	Wind	Play Badminton
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

P(Play=No) = 5/14

P(Play=Yes) = 9/14

The Model Training Phase

Learning Phase

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14$

$P(\text{Play}=No) = 5/14$

The Model Test Phase

Test Phase

- Given a new instance, predict its label

x'=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

- Look up tables achieved in the learning phrase

P(Outlook=*Sunny*|Play=*Yes*) = 2/9

P(Temperature=*Cool*|Play=*Yes*) = 3/9

P(Huminity=*High*|Play=*Yes*) = 3/9

P(Wind=*Strong*|Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

P(Outlook=*Sunny*|Play=*No*) = 3/5

P(Temperature=*Cool*|Play==*No*) = 1/5

P(Huminity=*High*|Play=*No*) = 4/5

P(Wind=*Strong*|Play=*No*) = 3/5

P(Play=*No*) = 5/14

The Model Test Phase

Test Phase

- Given a new instance, predict its label

  x′=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

- Decision making with the MAP rule

$$P(Yes|\mathbf{x}') \approx [P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes)]P(\mathrm{Play}=Yes)$$

$$= 0.0053$$

$$P(No|\mathbf{x}') \approx [P(Sunny|No)\,P(Cool|No)P(High|No)P(Strong|No)]P(\mathrm{Play}=No)$$

$$= 0.0206$$

Given the fact $P(Yes|\mathbf{x}') < P(No|\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Types of Naïve Bayes Algorithms

- Gaussian Naïve Bayes
- Multinomial Naïve Bayes
- Bernoulli Naïve Bayes

Gaussian Naïve Bayes

- When attribute values are continuous, an assumption is made that the values associated with each class are distributed according to Gaussian i.e., Normal Distribution
- If in our data, an attribute say "x" contains continuous data. We first segment the data by the class and then compute mean μ_y and variance σ_y^2 of each class

$$P(x_i \mid y) = \frac{1}{\sqrt{(2\pi\sigma_y^2)}} \exp\left(-\frac{\left(x_i-\mu_y\right)^2}{2\sigma_y^2}\right)$$

Multinomial Naïve Bayes

greatlearning

- MultiNomial Naive Bayes is preferred to use on data that is multinomially distributed.
- It is one of the standard classic algorithms which is used in text categorization (classification).
- Each event in text classification represents the occurrence of a word in a document.

Bernouli Naïve Bayes

- Bernoulli Naive Bayes is used on the data that is distributed according to multivariate Bernoulli distributions.i.e., multiple features can be there, but each one is assumed to be a binary-valued (Bernoulli, boolean) variable.
- So, it requires features to be binary valued.

Relevant Issues

- Violation of Independence Assumption
  - For many real world tasks,
  - Nevertheless, naïve Bayes works surprisingly well anyway!
- Zero conditional probability Problem
  - If no example contains the feature value
  - In this circumstance, $X_j = a_{jk}, \hat{P}(X_j = a_{jk} \mid C = c_i) = 0$ during test

Avoiding the zero-Probability Problem

- Naïve Bayesian prediction requires each conditional probability be non-zero.
- Otherwise the predicted probability will be zero.

$$P(x|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

- Example: Suppose a dataset with 1000 tuples,

  [income Low] = 0, [income Medium] = 990 and [income High] = 10
- Use Laplacian correction (or Laplacian Estimator)
  - Adding 1 to each case
  - $P(income = Low) = \frac{1}{3+1000} = 1/1003$
  - $P(income = Medium) = 991/1003$
  - $P(income = High) = 11/1003$
- The "corrected" probability estimates are close to their "uncorrected" counterparts

Advantages and Disadvantages

- Advantages
  - Naïve Bayes Algorithm is a fast, highly scalable algorithm
  - Naïve Bayes can be used for Binary and Multiclass classification
  - It provides different types of Naïve Bayes Algorithms like GaussianNB, MultinomialNB, BernoulliNB
  - It is a simple algorithm that depends on doing a bunch of counts
  - It is a popular choice for spam email classification
  - It can be easily train on small dataset
- Disadvantages
  - It considers all the features to be unrelated, so it can not learn the relationship between features.
    - Example: Lets say Remo is going to a party. While selecting cloths for party, Remo is looking at his cupboard. Remo likes to wear a white color shirt. In jeans, he likes to wear a brown jeans. But Remo does not like to wearing a white shirt with brown jeans. Naïve Bayes can learn individual features importance but can not determine the relationship among features

Summary

- Conditional Probability
- Bayes Theorem
- Naïve Bayes Classifier Algorithm
- Advantages and Disadvantages of Naïve Bayes Classifier
- Case Study 1
- Lab 1

Thanks!