Naïve Bayes Classification

Introduction

Spam Classification

- Given an email, predict whether it is spam or not

Medical Diagnosis

- Given a list of symptoms, predict whether a patient has disease X or not

Weather

- Based on temperature, humidity, etc… predict if it will rain tomorrow

Naïve Bayes Introduction

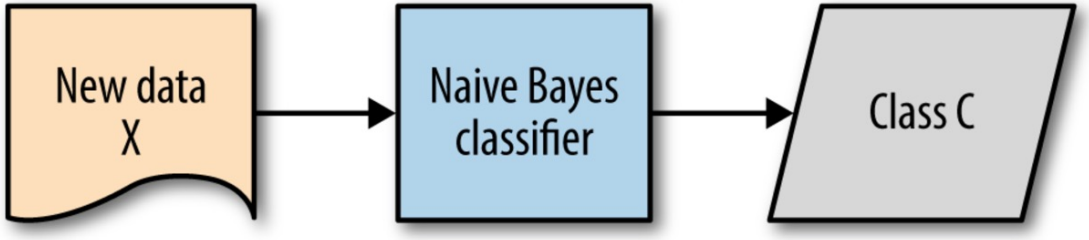
In machine learning, Naïve Bayes classifiers are a family of simple "**probabilistic classifiers**" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

(Wikipedia)

Classification process

New data = (X) = (X_1, X_2, ..., X_m)

Class C is a member of {C_1, C_2, ..., C_k}



Naïve Bayes Introduction

The Naïve Bayes is called "**naïve**" because it makes the assumption that the occurrence of a certain feature is **independent** of the occurrence of other features

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Playing Tennis statistics under various environmental conditions

Naïve Bayes Introduction

Thomas Bayes

Portrait purportedly of Bayes used in a 1936 book,[1] but it is doubtful whether the portrait is actually of him.[2] No earlier portrait or claimed portrait survives.

Born	c. 1701 London, England
Died	7 April 1761 (aged 59) Tunbridge Wells, Kent, England
Residence	Tunbridge Wells, Kent, England
Nationality	British
Alma mater	University of Edinburgh
Known for	Bayes' theorem
	Scientific career
Fields	Statistics

Signature

Bayes -

It refers to the statistician and philosopher Thomas Bayes and the theorem named after him, **Bayes theorem**, which is the base for the Naïve Bayes algorithm.

Bayes Theorem

Bayes theorem is stated as Probability of the event A given B is equal to the probability of the event B given A multiplied by the probability of A upon probability of B

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

- $P(A|B)$: Conditional probability of occurrence of the event A given the event B is true.
- $P(A)$ and $P(B)$: Probabilities of the occurrence of event A and B respectively.
- $P(B|A)$: Conditional probability of occurrence of the event B given the event A is true.

Probability

52 cards

P(Queen) = 1/13

Conditional probability

Draw a Face card,

P(Queen | Face) = ?

- Without Bayes theorem.
- With Bayes Theorem.

Bayes Theorem

Draw a Face card,

P(Queen | Face) = ?

- Without Bayes theorem
- With Bayes Theorem.

$$P(Queen|Face) = \frac{4}{12} = \frac{1}{3}$$

Bayes Theorem

Draw a Face card,

P(Queen | Face) = ?

- Without Bayes theorem
- With Bayes Theorem.

$$P(Queen|Face) = \frac{P(Face|Queen) \times P(Queen)}{P(Face)} = \frac{1 \times \frac{4}{52}}{\frac{12}{52}} = \frac{1}{3}$$

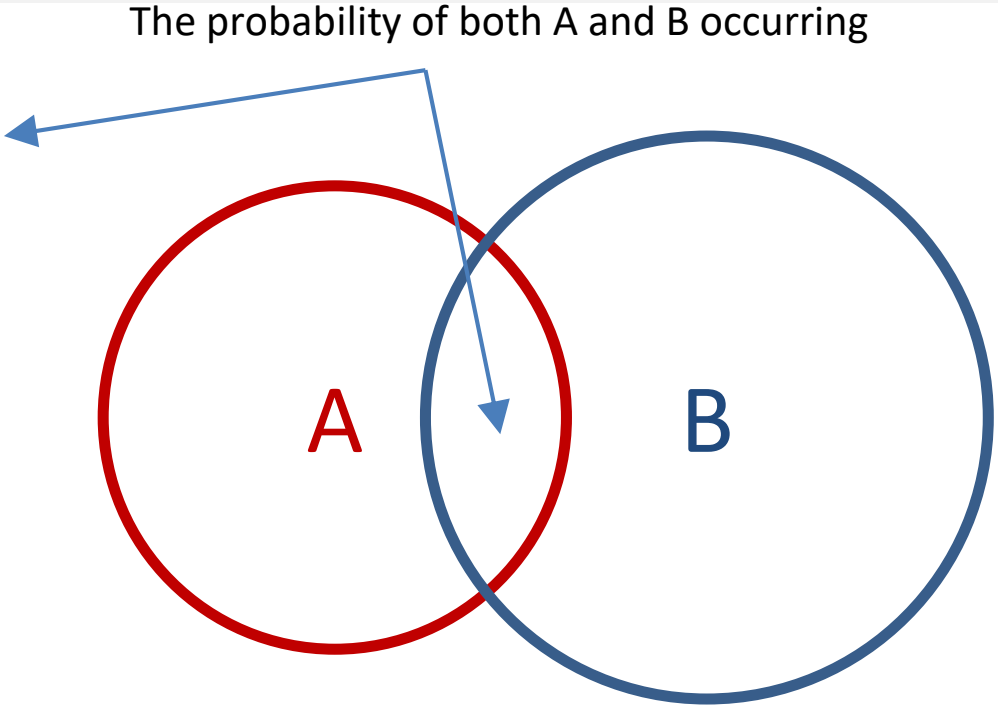
Proof of Bayes Theorem

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

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

$$P(A|B)P(B) = P(B|A)P(A)$$

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



Bayes Theorem for Naïve Bayes Classifier

Problem statement:

- Multiple features: {x_1, x_2, …, x_n}
- Classes: {C_1, C_2, … C_k}

Target: calculate the the conditional probability of a new sample with a feature vector {x_1, x_2, …, x_n} belonging to a particular class C_i.

$$P(C_i|x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n|C_i)P(C_i)}{P(x_1, x_2, \dots, x_n)} \text{ for } 1 \leq i \leq k$$

Bayes Theorem for Naïve Bayes Classifier

$$P(C_i|x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n|C_i)P(C_i)}{P(x_1, x_2, \dots, x_n)} \text{ for } 1 \leq i \leq k$$

where $P(x_1, x_2, \dots, x_n) = P(x_1 \cap x_2 \cap \cdots \cap x_n)$

It will be difficult to collect data $P(x_1, x_2, \dots, x_n|C_i)$ and $P(x_1, x_2, \dots, x_n)$ in real life (e.g., a person is getting fever, also getting cold, also having body temperature 38C, etc.). Thus, we will make an assumption to make equation above implementable in real life. We will use conditionally **independent features** as our assumption.

Independent events / features

Independent event means an event that is **not effected** by previous event, e.g., coin toss.

$$P(A, B) = P(A)P(B)$$

Thus, for conditional probability, the formula will be:

$$P(A, B|C) = P(A|C)P(B|C)$$

Under conditional **independent assumption**, Bayes formula becomes:

$$P(C_i|x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n|C_i)P(C_i)}{P(x_1, x_2, \dots, x_n)}$$

$$= \frac{P(x_1|C_i)P(x_2|C_i)P(x_3|C_i) \dots P(x_n|C_i)P(C_i)}{P(x_1, x_2, \dots, x_n)}$$

$$= \frac{P(C_i)\prod_{m=1}^{n} P(x_m|C_i)}{P(x_1, x_2, \dots, x_n)}$$

Bayes Theorem for Naïve Bayes Classifier

The class is determined:

$$class = \underset{C_i}{\text{argmax}} \frac{P(C_i) \prod_{m=1}^{n} P(x_m|C_i)}{P(x_1, x_2, \ldots, x_n)}$$

and since the denominator for all C_i is the same, Naïve Bayes classifier can be simplified as follows

$$class = \underset{C_i}{\text{argmax}}\, P(C_i) \prod_{m=1}^{n} P(x_m|C_i)$$

Example

Training phase

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

P(Play=Yes)=?

P(Play=No)=?

Outlook	Play=Yes	Play=No
Sunny	?	?
Overcast	?	?
Rain	?	?

Temp	Play=Yes	Play=No
Hot	?	?
Mild	?	?
Cool	?	?

Humid	Play=Yes	Play=No
High	?	?
Normal	?	?

Wind	Play=Yes	Play=No
Strong	?	?
Weak	?	?

Example

Training phase

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

P(Play=Yes)=9/14

P(Play=No)=5/14

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

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

Humid	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

Example

Test phase

Given a new sample, predict its label:

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

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

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

Humid	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(Play=Yes)=9/14
P(Play=No)=5/14

Example

Test phase

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

P(Yes|x')=P(Yes)P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes)

P(No|x')=P(No)P(Sunny|No)P(Cool|No)P(High|No)P(Strong|No)

P(Play=Yes)=9/14
P(Play=No)=5/14

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

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

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

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

Example

Test phase

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

P(Yes|x')=P(Yes)P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes)=**0.0053**

P(No|x')=P(No)P(Sunny|No)P(Cool|No)P(High|No)P(Strong|No)=**0.0206**

P(Yes|x') < P(No|x') → label is "No"

P(Play=Yes)=9/14
P(Play=No)=5/14

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

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

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

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

Gaussian Naïve Bayes

Example: Continuous-valued Features

- Temperature is naturally of continuous value.

  Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

  No: 27.3, 30.1, 17.4, 29.5, 15.1

What is the value of $P(x|y)$ where x is a temperature and y is the class? For instance:

- $P(22.4|Yes) = ?$
- $P(28.5|No) = ?$

Gaussian Naïve Bayes

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

Example: Continuous-valued Features

- Temperature is naturally of continuous value.

  Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

  No: 27.3, 30.1, 17.4, 29.5, 15.1

- Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$

$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

- **Learning Phase**: output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2\times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2\times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

Laplace Smoothing

Training data

Sky	Temp	Humid	Play?
sunny	warm	normal	yes (play)
sunny	warm	high	yes (play)
rainy	cold	high	no (¬play)
sunny	warm	high	yes (play)

Test data

Predict label for X=(**rainy**, **warm**, **normal**)

Laplace Smoothing

Training data

Sky	Temp	Humid	Play?
sunny	warm	normal	yes (play)
sunny	warm	high	yes (play)
rainy	cold	high	no (¬play)
sunny	warm	high	yes (play)

Test data

Predict label for X=(**rainy**, **warm**, **normal**)

But P(rainy|yes) = 0 → P(yes|X)=0 → **Apply Laplace Smoothing**.

Laplace Smoothing

Training data

Sky	Temp	Humid	Play?
sunny	warm	normal	yes (play)
sunny	warm	high	yes (play)
rainy	cold	high	no (¬play)
sunny	warm	high	yes (play)
rainy			yes
sunny			yes

Test data

Predict label for X=(**rainy**, **warm**, **normal**)

P(rainy|yes) = (0+1)/(3+2)=1/5 (Not Zero!)

Laplace Smoothing

Notice that some probabilities estimated by counting might be zero
→Possible overfitting!

$$P(X = x_i | C = c_j) = \frac{m_i}{n_j}$$

Fix by using Laplace smoothing: **adds 1 to each count**

$$P(X = x_i | C = c_j) = \frac{m_i + \mathbf{1}}{n_j + |\boldsymbol{values}(\boldsymbol{X})|}$$

where

- $\boldsymbol{m_i}$ is the count of training samples with value of $\boldsymbol{x_i}$ for attribute $\boldsymbol{X}$ and class label $\boldsymbol{c_j}$
- $\boldsymbol{n_j}$ is the number of training samples having class $\boldsymbol{c_j}$
- $\boldsymbol{|values(X)|}$ is the number of values $\boldsymbol{X}$ can take on.

Log-probability

In practice, we use log-probability to prevent underflow

$$class = \underset{C_i}{\operatorname{argmax}} P(C_i) \prod_{m=1}^{n} P(x_m|C_i)$$

$$= \underset{C_i}{\operatorname{argmax}} \log P(C_i) + \sum_{m=1}^{n} logP(x_m|C_i)$$

Log-probability

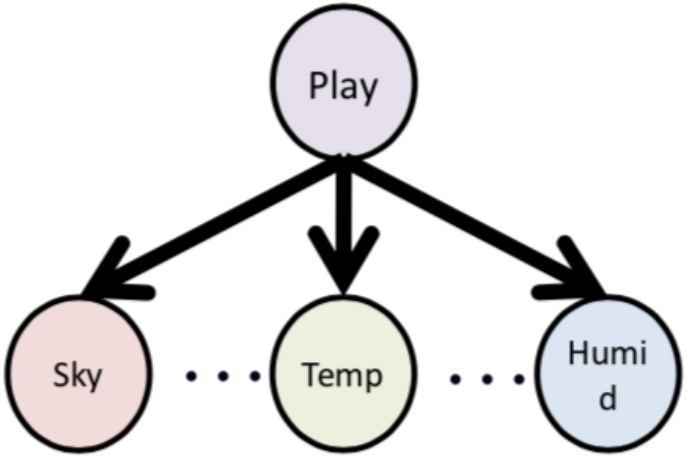
Training data

Sky	Temp	Humid	Play?
sunny	warm	normal	yes (play)
sunny	warm	high	yes (play)
rainy	cold	high	no (¬play)
sunny	warm	high	yes (play)

Test data

Predict label for X=(**rainy**, **warm**, **normal**)

Log-probability with Laplace Smoothing



Play?	P(Play)
yes	3/4
no	1/4

Temp	Play?	P(Temp | Play)
warm	*yes*	4/5
cold	*yes*	1/5
warm	*no*	1/3
cold	*no*	2/3

Sky	Play?	P(Sky | Play)
sunny	*yes*	4/5
rainy	*yes*	1/5
sunny	*no*	1/3
rainy	*no*	2/3

Humid	Play?	P(Humid | Play)
high	*yes*	3/5
norm	*yes*	2/5
high	*no*	2/3
norm	*no*	1/3

Predict label for:

$\mathbf{x}$ = (rainy, warm, normal)

$$P(\text{play} \mid \mathbf{x}) \propto \log P(\text{play}) + \log P(\text{rainy} \mid \text{play}) + \log P(\text{warm} \mid \text{play}) + \log P(\text{normal} \mid \text{play})$$
$$\propto \log 3/4 + \log 1/5 + \log 4/5 + \log 2/5 = -1.319$$

predict PLAY

$$P(\neg\text{play} \mid \mathbf{x}) \propto \log P(\neg\text{play}) + \log P(\text{rainy} \mid \neg\text{play}) + \log P(\text{warm} \mid \neg\text{play}) + \log P(\text{normal} \mid \neg\text{play})$$
$$\propto \log 1/4 + \log 2/3 + \log 1/3 + \log 1/3 = -1.732$$

Naïve Bayes Summary

Advantages:

- Fast to train (single scan through data)
- Fast to classify
- Not sensitive to irrelevant features
- Handles real and discrete data
- Handles streaming data well

Disadvantages:

- Assumes independence of features

Q&A

Thank you