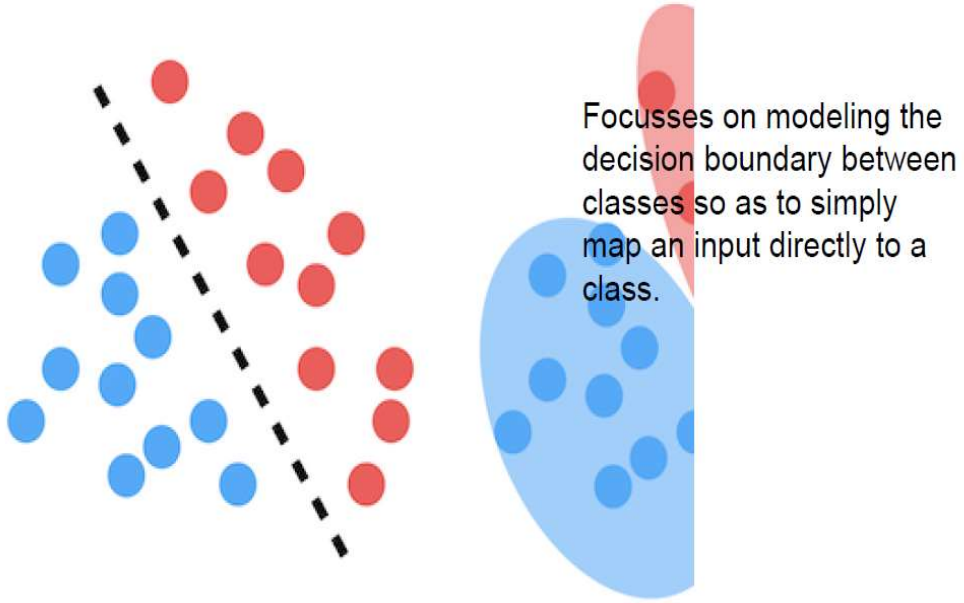
Naïve Bayes classifier

Naïve Bayes classifier

- Introduction to Naïve Bayes Classifier
- Basics of Probability
- Understanding Naïve Bayes Classifier
- Classify SMS message to be spam or not

Background

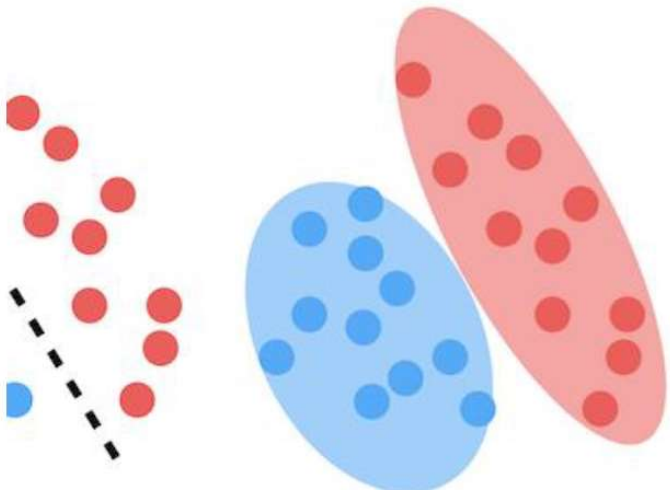
- Classification algorithms that differentiates between classes on the basis of definite decision boundaries.
- Classification algorithms that learn boundaries between classes.
- Classification algorithms that constructs decision boundaries that separates classes are called **discriminative models**.



Background

- What if we differentiate between two classes by analyzing probability distribution of data…

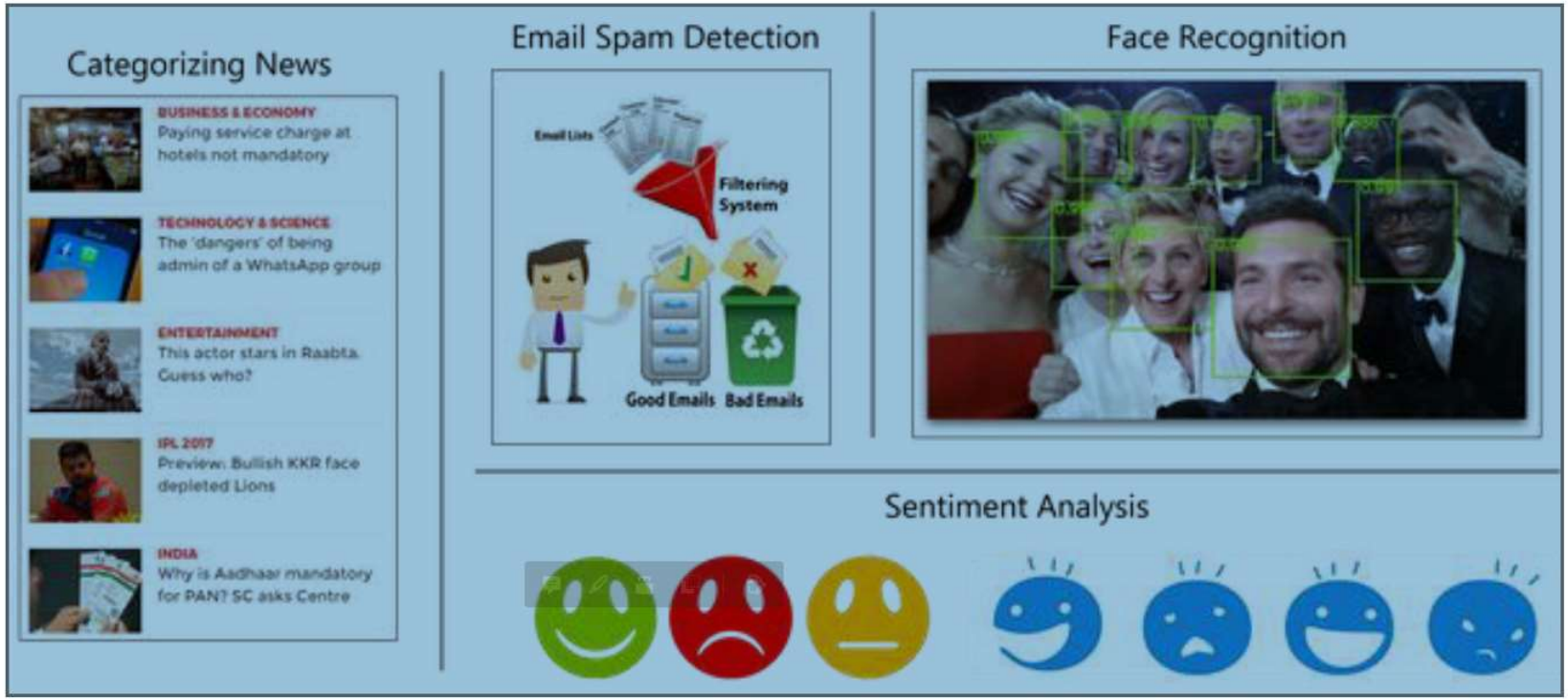
Let us build a model A that characterizes the red dots and a model B that characterizes the blue dots using probability.
Any new data point can be assigned to either class based on how likely it is to belong to either A or B. These are called generative models.



What is Naïve Bayes?

- Naive Bayes classifier is an algorithm that learns the probability that an object with certain features, belong to a particular group or class.

Where is Naïve Bayes used?



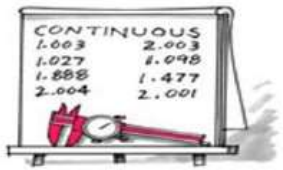
Advantages of Naïve Bayes

Simple and Easy to implement

Fast

Needs lesser training data

Scalable



Handles both continuous and discrete data

Basics of Probability

- What is Probability?

Estimate How **likely** something is to happen.

Rain is likely to fall

Obtaining a movie ticket

Will Arsenal win?

Stocks are likely to rise or fall

Boy or girl?

What is Probability?

We draw on previous experience to determine how likely something is to occur

40%

60%

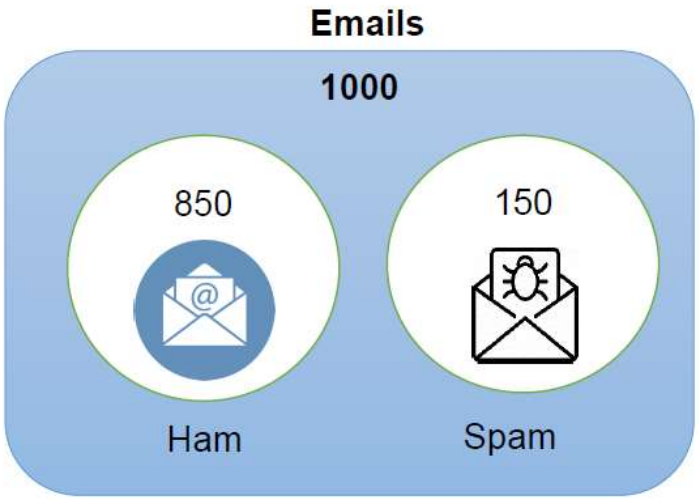
0.7

70%

0.5

Probability explained through an example

John's email id has a total of 1000 emails, 850 emails are genuine whereas 150 mails are spam.



Probability of a spam mail appearing in John's account is:

$$P(\text{spam}) = \frac{\text{Number of spam emails}}{\text{Total number of emails}} = \frac{150}{1000}$$

$$= 0.15 = 15\%$$

John’s emails have multiple occurrences of the word ‘Lottery’. Let’s analyze them closely..

Analyze Emails with word “lottery”

Scrolling through John’s mails we see the word ’lottery’ appearing frequently.

From: Information Desk <info@euroonlinelottery.com>
Subject: EU / Commonwealth Lottery Promotions

Your email address was selected to claim the sum of $ 500,000.00 in the 2011 European lottery.
To claim your prize, please contact our agent in Lagos, Nigeria.
Contact person: Mr. Marshall Ellis e-mail: marshallellis11@live.com
Phone: +2348036954742
Congratulations!
Vincent Kilkenny (Coordinator)

The National Lottery

Congratulations! Winning Notification!

UK online international Lottery Award Prize of £820,731.00 (Eight Hundred and Twenty Thousand, Seven Hundred and Thirty One British Pounds).

Email Account Owner,

Congratulations!! We are happy to announce that you have won an online lottery prize in our international lottery promotion.

Your active e-mail address attached to computer generated **ticket number: BH 05607545 7152** with reference number **UK/KA2C110EN5** has won UK Lottery 2nd category award prize.

Contact our Fiduciary agents immediately to commence release of your lottery prize by providing details as listed below.
1. Full Name:
2. Email Address:
3. Age/Occupation:
4. Reference Number/Ticket Number
5. Phone Number:
6. Country:
7. Date of draw

UK Lottery Fiduciary Agents:
Mr. Peter Wills
Foreign Service Manager
Watford Regional Centre
Tolpits Lane, Watford WD18 9RN .United Kingdom
E-mail: agent.claim@natlott.com
wiliam_boyd@yahoo.co.uk

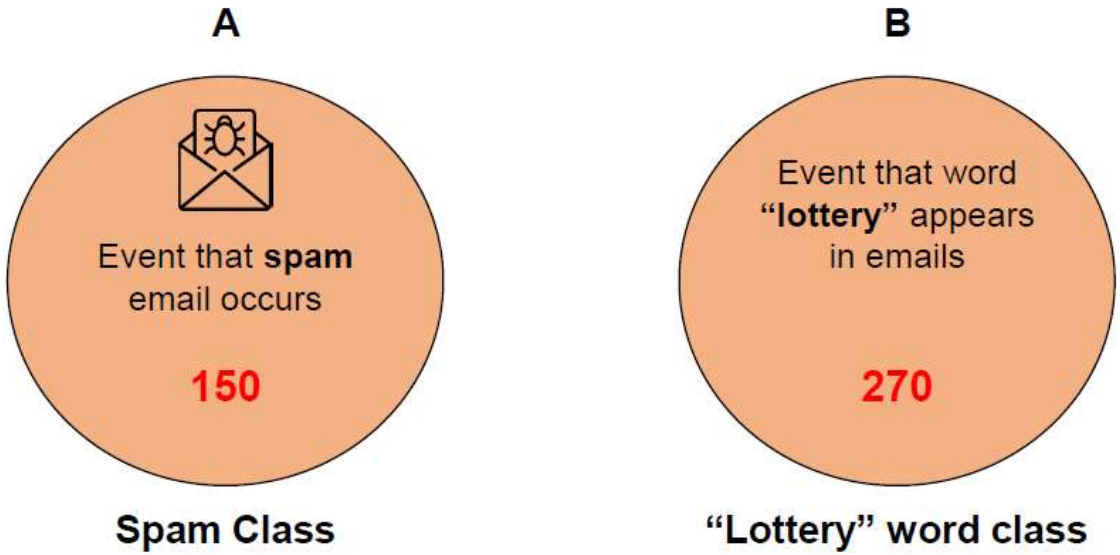
Thank you and Congratulation once again!

Yours faithfully,
Angela M. Johnson.
(Online coordinator)
The UK Lottery International Promotion Inc.

"Please do not reply back to the senders address or the from email address, this notification is sent automatically via computer virtual notification to winning email addresses and a response will not be attended by Human but computer" contact the fiduciary agents as above "

Let us consider two simple events..

Let us consider two simple events in Emails



John's mail id consists of:

- **150** spam emails
- **270** emails containing the word "Lottery"

Appearance of "lottery" in spam and genuine emails

Frequency table of "lottery" word occurring in emails

	Emails containing "lottery"	Emails not containing "lottery"	Total number of emails
Number of Spam emails	140	10	**150**
Number of Ham emails	130	720	**850**
Total number of emails	**270**	**730**	**1000**

Compute probability of word 'lottery' appearing in emails

27% of emails have the word "**lottery**".

P(lottery)=270/1000=0.27

	Emails containing "lottery"	Emails not containing "lottery"	Total number of emails
Number of Spam emails	140	10	**150**
Number of Ham emails	130	720	**850**
Total number of emails	**270**	**730**	**1000**

Let us explore different types of probabilities…

Types of Probabilities: Joint Probability

Types of Probabilities: Joint Probability

Joint probability represents **probability of two different events occurring together** at the same point in time.

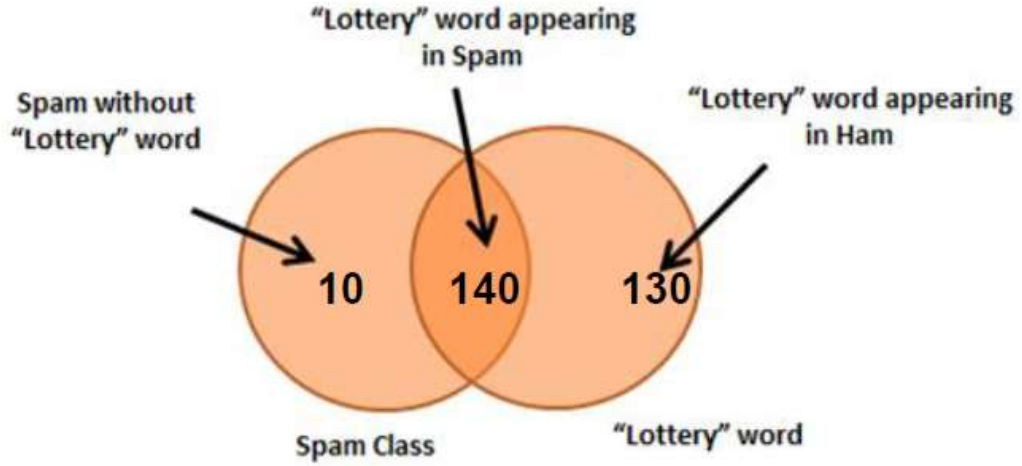
We have two events

A: **Event** that **Spam email** occurs and
B: **Event** that **word "lottery"** appears in an email

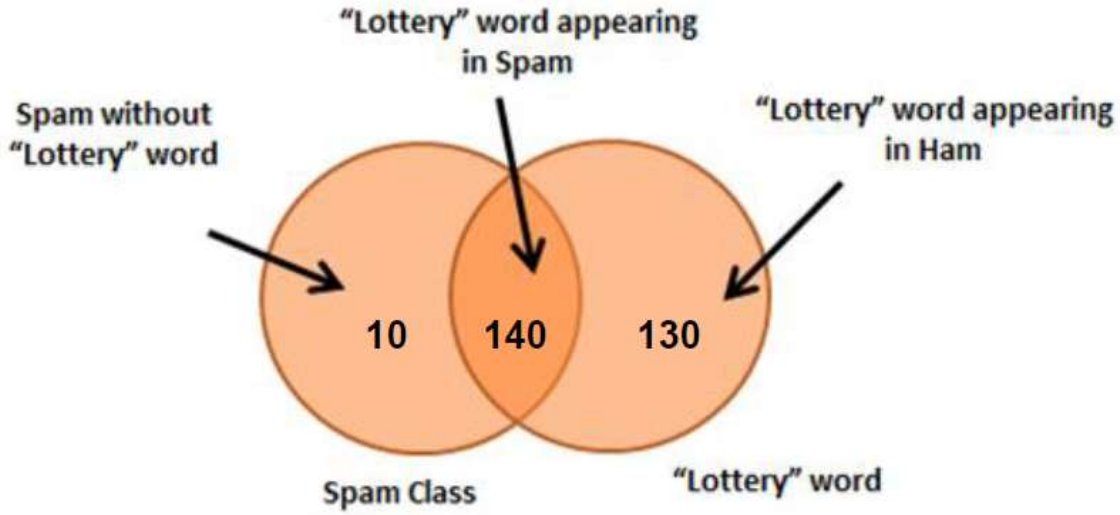
And let us assume that they are Independent events.

Venn Diagram for representing count of events

	Emails containing "lottery"	Emails not containing "lottery"	Total number of emails
Number of Spam emails	140	10	**150**
Number of Ham emails	130	720	**850**
Total number of emails	**270**	**730**	**1000**



Let us compute joint probability of word 'lottery' appearing in spam



$$P(Spam, Lottery) = 140/1000 = 0.14$$

14% of total emails are spam and contain the word lottery.

Types of Probabilities: Marginal Probability

Types of Probabilities: Marginal Probability

Probability of a single event occurring, irrespective of any other event is called marginal probability.

A: Spam email occurs

P(A)=150/1000=0.15

***B*: Word lottery appears in email**

P(B)=270/1000=0.27

Frequency table of "lottery" word occurring in email

	Emails containing "lottery"	Emails not containing "lottery"	Total number of emails
Number of Spam emails	140	10	150
Number of Ham emails	130	720	850
Total number of emails	270	730	1000

Types of Probabilities: Conditional Probability

Types of Probabilities: Conditional Probability

Suppose

A: Event that **a spam email occurs** and
B: Event that **word lottery** appears in an email

are dependent events.

How do we Predict whether email is spam given the word lottery?

Conditional Probability is a simple way to calculate probability of uncertain events given some prior information.

Probability of **an event** given that **another event** has **already occurred** is called **conditional probability**.

Conditional Probability

- For example, suppose you go out for lunch at the same place and time every Friday and you are served lunch within 15 minutes with probability 0.9. However, given that you notice that the restaurant is exceptionally busy, the probability of being served lunch within 15 minutes may reduce to 0.7. This is the conditional probability of being served lunch within 15 minutes given that the restaurant is exceptionally busy.
- The usual notation for "event A occurs given that event B has occurred" is "A | B" (A given B). The symbol | is a vertical line and does not imply division.
- P(A | B) denotes the probability that event A will occur given that event B has occurred already.

Conditional Probability

- A rule that can be used to determine a conditional probability from unconditional probabilities is:

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

- where:
- P(A | B) = the (conditional) probability that event A will occur given that event B has occurred already.
- P(A∩B) = the (unconditional) probability that event A and event B both occur.
- P(B) = the (unconditional) probability that event B occurs.

Naive Bayes classifier

- Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.
- Bayesian classification is based on Bayes' Theorem.
- It is based on simplifying assumpations that the attribute values are *conditionally independent,*
- A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

Naive Bayes classifier

- For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. *A naive Bayes classifier considers all these features to contribute independently to the probability that this fruit is an apple*, whether or not they're in fact related to each other or to the existence of the other features.

- This reduces significantly computation cost since calculating each one of the $P(a_i|v_j)$ requires only a frequency count over the tuples in the training data with class value equal to v_j .

Bayes Theorem : Basics

- Let **X** be a data sample : class label is unknown
- Let H be a *hypothesis* that X belongs to a specified class C
- For classification problems, we want to determine P(H|**X**), the probability that the hypothesis holds given the observed data sample **X**
- P(H) (*prior probability*), the initial probability
  - E.g., **X** will buy computer, regardless of age, income or any other information, for that matter.
- P(H|X) (*posteriori probability*), the probability of observing the sample **X**, given that the hypothesis holds
  - For example, suppose our world of data tuples is confined to customers described by the attributes *age* and *income*, respectively,
  - and that ***X*** is a 35-year-old customer with an income of $40,000.
  - Suppose that *H* is the hypothesis that our customer will buy a computer.
  - Then P(H|X) reflects the probability that customer ***X*** will buy a computer given that we know the customer's age and income.

Bayesian Theorem

- Given data **X**, *posteriori probability of a hypothesis* H, P(H|**X**), follows the Bayes theorem

$$P(H \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid H)P(H)}{P(\mathbf{X})}$$

- *P*(***X***|*H*) is the descriptor posterior probability of ***X*** conditioned on *H*. That is, it is the probability that a customer, ***X***, is 35 years old and earns $40,000, given that we know the customer will buy a computer.

- Predicts **X** belongs to Ci if the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|X)$ for all the *k* classes.
- Practical difficulty: require initial knowledge of many probabilities.

Bayesian Theorem

- Assume target function $f : X \rightarrow Y$ (A function *f* with domain *X* and codomain *Y*). The elements of *X* are called arguments of *f*. For each argument *x*, the corresponding unique *y* in the codomain is called the function value at *x* or the *image* of *x* under *f*.
- If, each instance X is describes by attributes <a1, a2, a3, ….an>.
- Most probable value of f(X) is: v_{MAP}
- Using *Bayes Theorem* we can write the expression as :

$$v_{MAP} = \arg \max_{v_j \in V} P\left(v_j \middle| a_1, a_2, ..., a_n\right)$$

$$= \arg \max_{v_j \in V} \frac{P\left(a_1, a_2, ..., a_n \middle| v_j\right) P\left(v_j\right)}{P\left(a_1, a_2, ..., a_n\right)}$$

$$= \arg \max_{v_j \in V} P\left(a_1, a_2, ..., a_n \middle| v_j\right) P\left(v_j\right)$$

$$v_{MAP} = \arg \max_{v_j \in V} P\left(v_j\right) \prod_i P\left(a_i \middle| v_j\right)$$

- The *denominator* does not depend on the choice of v_j and thus, it can be omitted from the arg max argument.

Bayesian Theorem

- In mathematics, **argmax** stands for the **argument of the maximum**, that is to say, the set of points of the given argument for which the given function attains its maximum value.

Towards Naïve Bayesian Classifier

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n- dimensional attribute vector **X** = (x_1, x_2,..., x_n), showing *n* measurements made on the tuple from *n* attributes.
- Suppose there are *m* classes C_1, C_2, …, C_m.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

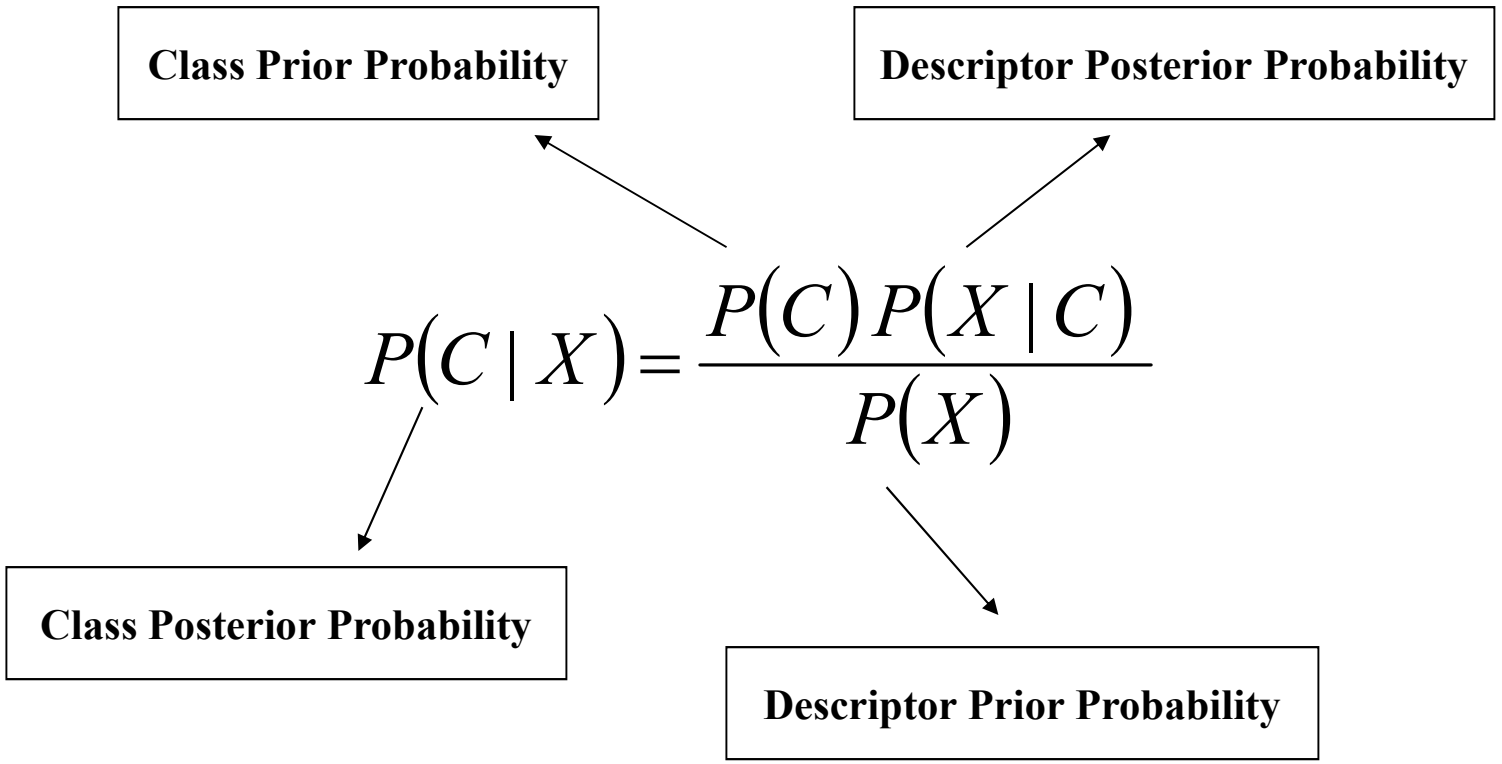
$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized

Bayesian Classifier – Basic Equation



Naïve Bayesian Classifier: Training Dataset

Class:
C1:buys_computer = 'yes'
C2:buys_computer = 'no'

Data sample
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

age	income	student	credit_rating	_comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayesian Classifier: An Example

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357

- Compute $P(X|C_i)$ for each class
  P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
  P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
  P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
  P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019
$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

Training Data

Outlook	Temp	Humidity	Windy	Play?
sunny	hot	high	FALSE	No
sunny	hot	high	TRUE	No
overcast	hot	high	FALSE	Yes
rainy	mild	high	FALSE	Yes
rainy	cool	normal	FALSE	Yes
rainy	cool	Normal	TRUE	No
overcast	cool	Normal	TRUE	Yes
sunny	mild	High	FALSE	No
sunny	cool	Normal	FALSE	Yes
rainy	mild	Normal	FALSE	Yes
sunny	mild	normal	TRUE	Yes
overcast	mild	High	TRUE	Yes
overcast	hot	Normal	FALSE	Yes
rainy	mild	high	TRUE	No

P(yes) = 9/14
P(no) = 5/14

Bayesian Classifier – Probabilities for the weather data

Frequency Tables

Outlook	No	Yes
Sunny	3	2
Overcast	0	4
Rainy	2	3

Temp.	No	Yes
Hot	2	2
Mild	2	4
Cool	1	3

Humidity	No	Yes
High	4	3
Normal	1	6

Windy	No	Yes
False	2	6
True	3	3

Outlook	No	Yes
Sunny	3/5	2/9
Overcast	0/5	4/9
Rainy	2/5	3/9

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

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

Windy	No	Yes
False	2/5	6/9
True	3/5	3/9

Likelihood Tables

Bayesian Classifier – Predicting a new day

	Outlook	Temp.	Humidity	Windy	Play
X →	sunny	cool	high	true	NO

Class?

P(**yes**|**X**) = p(sunny|yes) x p(cool|yes) x p(high|yes) x p(true|yes) x **p(yes)**

= 2/9 x 3/9 x 3/9 x 3/9 x **9/14** = 0.0053 => 0.0053/(0.0053+0.0206) = 0.205

P(**no**|**X**) = p(sunny|no) x p(cool|no) x p(high|no) x p(true|no) x **p(no)**

= 3/5 x 1/5 x 4/5 x 3/5 x **5/14** = 0.0206=0.0206/(0.0053+0.0206) = 0.795

Outlook	No	Yes
Sunny	3/5	2/9
Overcast	0/5	4/9
Rainy	2/5	3/9

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

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

Windy	No	Yes
False	2/5	6/9
True	3/5	3/9

Bayesian Classifier – zero frequency problem

- What if a descriptor value doesn't occur with every class value

  P(outlook=overcast|No)=0

- Remedy: add 1 to the count for every descriptor-class combination (Laplace Estimator)

Outlook	No	Yes
Sunny	3+1	2+1
Overcast	0+1	4+1
Rainy	2+1	3+1

Temp.	No	Yes
Hot	2+1	2+1
Mild	2+1	4+1
Cool	1+1	3+1

Humidity	No	Yes
High	4+1	3+1
Normal	1+1	6+1

Windy	No	Yes
False	2+1	6+1
True	3+1	3+1

Bayesian Classifier – General Equation

$$P(C_k \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C_k)\, P(C_k)}{P(\mathbf{X})}$$

Likelihood: $P(\mathbf{X} \mid C_k)$

Continues variable: $$P(x \mid C) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

Bayesian Classifier – Dealing with numeric attributes

- Usual assumption: attributes have a *normal* or *Gaussian* probability distribution (given the class)
- The *probability density function* for the normal distribution is defined by two parameters:
  - The *sample mean* μ: $$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$
  - The *standard deviation* $$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$
  - The density function *f(x)*: $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

EXAMPLE-I

Department	*status*	*age*	*salary*
Sales	senior	31. . .35	41K.. .45K
Sales	junior	26. . .30	26K.. .30K
Sales	junior	31. . .35	31K.. .35K
systems	junior	21. . .25	31K.. .35K
systems	senior	31. . .35	66K.. .70K
systems	junior	26. . .30	31K.. .35K
systems	senior	41. . .45	66K.. .70K
marketing	senior	26. . .30	46K.. .50K
marketing	junior	31. . .35	41K.. .45K
secretary	senior	46. . .50	41K.. .45K
secretary	junior	26. . .30	26K.. .30K

- Define Bayesian Classification .Given a **data tuple** having the values *"systems", "26. . . 30",* and "41K.. .45K*"* for the attributes *department, age*, and *salary*, respectively, what would be a naive Bayesian classification of the status for the **given data tuple** ?

Example- continuous attributes

- Consider the training dataset as shown in below table. Let ***Play*** be the class label attribute. There are two distinct classes, namely, ***yes*** and ***no*** and two numeric attributes namely "temp" and "humidity".

Outlook	Temp	Humidity	Windy	Play?
sunny	85	85	FALSE	**No**
sunny	80	90	TRUE	**No**
overcast	83	86	FALSE	**Yes**
rainy	70	96	FALSE	**Yes**
rainy	68	80	FALSE	**Yes**
rainy	65	70	TRUE	**No**
overcast	64	65	TRUE	**Yes**
sunny	72	95	FALSE	**No**
sunny	69	70	FALSE	**Yes**
rainy	75	80	FALSE	**Yes**
sunny	75	70	TRUE	**Yes**
overcast	72	90	TRUE	**Yes**
overcast	81	75	FALSE	**Yes**
rainy	71	91	TRUE	**No**

- Given a **data tuple** having the values *"sunny", 66, 89 and "true"* for the attributes outlook*, temp.*, humidity and windy respectively, what would be a naive Bayesian classification of the *Play* for the given tuple?

Example- continuous attributes

The numeric weather data with summary statistics

Outlook			temperature			humidity			windy		play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std dev	6.2	7.9	std dev	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											