Pattern Classification and recognition
Statistical Decision Making Unit-2
Dr. H C Vijayalakshmi

References

1. Pattern Classification by Richard O. Duda, Peter E. Hart, David G. Stork
2. Pattern Recognition and Image Analysis by Earl Gose, Richard Johasonbaugh and Steve Jost
3. CSE IIT Kgphttps://cse.iitkgp.ac.in › course › BayesClassifierPosteriori probability (Bayes rule)
4. Luca Chech and Jolanda Malamud Supervisor: Thomas Parr
5. Dr. Debasis Mahantha CS40003 Lecture#8, Department of Computer Science, IITKG

Classification (Revision)

It is the task of assigning a class label to an input pattern. The class label indicates one of a given set of classes. The classification is carried out with the help of a model obtained using a learning procedure. There are two categories of classification. **supervised learning** and **unsupervised learning.**

- **Supervised learning** makes use of a set of examples which already have the class labels assigned to them.

- **Unsupervised learning** attempts to find inherent structures in the data.

- **Semi-supervised learning** makes use of a small number of labeled data and a large number of unlabeled data to learn the classifier.

Learning - Continued

- The classifier to be designed is built using input samples which is a mixture of all the classes.
- The classifier learns how to discriminate between samples of different classes.
- If the Learning is offline i.e. Supervised method then, the classifier is first given a set of training samples and the optimal decision boundary found, and then the classification is done.
- Supervised Learning refers to the process of designing a pattern classifier by using a Training set of patterns to assign class labels.
- If the learning involves no teacher and no training samples (Unsupervised). The input samples are the test samples itself. The classifier learns from the samples and classifies them at the same time.

Supervised Learning

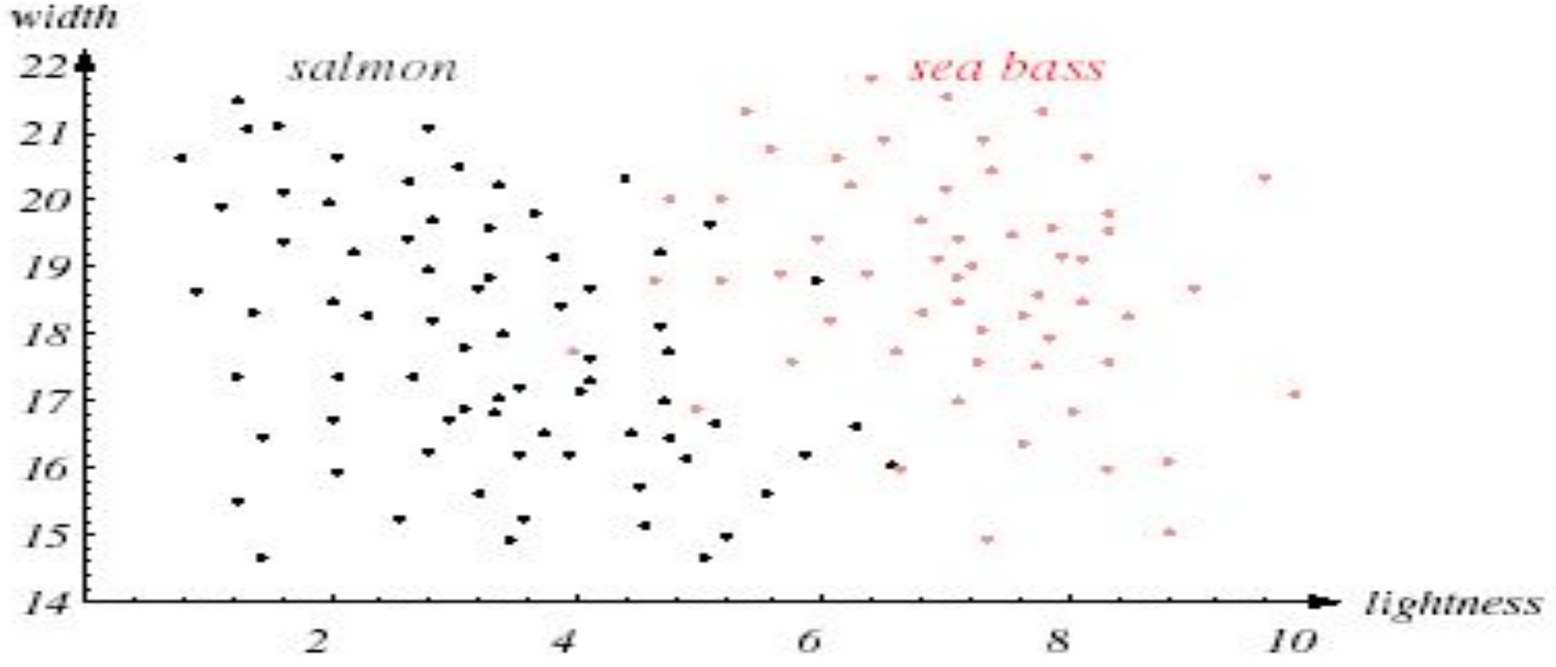


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Parametric decision making

This refers to the situation in which we assume the general form of probability distribution function or density function for each class.

- Parametric Methods **uses a fixed number of parameters to build the model**.

- Parametric methods are assumed to be a normal distribution.

- Parameters for using the normal distribution is –

  Mean

  Standard Deviation

- For each feature, we first estimate the mean and standard deviation of the feature for each class.

Parametric decision making (Continued)

- If a group of features – multivariate normally distributed, estimate mean and standard deviation and covariance.
- Covariance is a measure of the relationship between two random variables, in statistics.
- The covariance indicates the relation between the two variables and helps to know if the two variables vary together.
- In the covariance formula, the covariance between two random variables X and Y can be denoted as Cov(X, Y).
- x_i is the values of the X-variable
- y_j is the values of the Y-variable
- x^- is the mean of the X-variable
- y^- is the mean of the Y-variable
- N is the number of data points

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n}$$

Positive and negative covariance

- Positive Co variance: If temperature goes high sale of ice cream also goes high. This is positive covariance. Relation is very close.

- On the other hand cold related disease is less as the temperature increases. This is negative covariance.

- No co variance : Temperature and stock market links. Value of the covariance will be zero

Example1: Two set of data X and Y

Day	x	y
1	30	5
2	35	8
3	40	8
4	25	4
5	35	5
Mean	33	6

Compute x-x(mean) and y-y(mean)

Day	x	y	$x-\bar{x}$	$y-\bar{y}$
1	30	5	−3	−1
2	35	8	+2	+2
3	40	8	+7	+2
4	25	4	−8	−2
5	35	5	+2	−1
Mean	33	6		

Apply Covariance formula

Day	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	30	5	–3	–1	3
2	35	8	+2	+2	4
3	40	8	+7	+2	14
4	25	4	–8	–2	16
5	35	5	+2	–1	-2
Mean	33	6			Sum = 35

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Final result will be 35/5 = 7 = is a positive covariance

- Example-2 John is an investor. His portfolio primarily tracks the performance of the S&P 500 and John wants to add the stock of ABC Corp. Before adding the stock to his portfolio, he wants to assess the direct relationship between the stock and the S&P 500.
- John does not want to increase the unsystematic risk of his portfolio. Thus, he is not interested in owning securities in the portfolio that tend to move in the same direction.
- John can calculate the covariance between the stock of ABC Corp. and S&P 500

	S&P 500	ABC Corp.
2013	1,692	68
2014	1,978	102
2015	1,884	110
2016	2,151	112
2017	2,519	154

$$\text{Mean (S\&P 500)} = \frac{1{,}692 + 1{,}978 + 1{,}884 + 2{,}151 + 2{,}519}{5} = 2{,}044.80$$

$$\text{Mean (ABC Corp.)} = \frac{68 + 102 + 110 + 112 + 154}{5} = 109.20$$

			Step 3		Step 4
	S&P 500	ABC Corp.	a	b	a x b
2013	1,692	68	-352.80	-41.20	14,535.36
2014	1,978	102	-66.80	-7.20	480.96
2015	1,884	110	-160.80	0.80	-128.64
2016	2,151	112	106.20	2.80	297.36
2017	2,519	154	474.20	44.80	21,244.16
Mean	**2,044.80**	**109.20**	**Sum**		**36,429.20**

- $\text{Cov(S\&P-500,ABC Corp)} = \frac{36{,}429.20}{5} = 7285.84$

The positive covariance indicates that the price of the stock and the S&P 500 tend to move in the same direction.

Example-3

Let's say you are the new owner of a small ice-cream shop in a little village near the beach. You noticed that there was more business in the warmer months than the cooler months. Before you alter your purchasing pattern to match this trend, you want to be sure that the relationship is real.

How can you be sure that the trend you noticed is real?

Temperature	Number of Customers
98	15
87	12
90	10
85	10
95	16
75	7

Temperature $(x - \bar{x})$	Customers $(y - \bar{y})$	Product $(x - \bar{x})(y - \bar{y})$
98 - 88.33= 9.67	15 - 11.67= 3.33	32.20
87 - 88.33= -1.33	12 - 11.67= 0.33	-0.44
90 - 88.33= 1.67	10 - 11.67= -1.67	-2.79
85 - 88.33= -3.33	10 - 11.67= -1.67	5.56
95 - 88.33= 6.67	16 - 11.67= 4.33	28.88
75 - 88.33= -13.33	7 - 11.67= -4.67	62.25

In the final step is to divide by n = 6 .
Covariance =125.66/6 = 20.94

The table below describes the rate of economic growth (xi) and the rate of return on the S&P 500 (yi). Using the covariance formula, determine whether economic growth and S&P 500 returns have a positive or inverse relationship. Before you compute the covariance, calculate the mean of x and y.

Economic Growth % (xi)	S&P 500 Returns % (yi)
2.1	8
2.5	12
4.0	14
3.6	10

E(X) = 3.1
E(Y) = 11

Xi	Yi	Xi-E(X)	Yi-E(Y)
2.1	8	-1	-3
2.5	12	-0.6	1
4.0	14	0.9	3
3.6	10	0.5	-1

$\mathrm{Cov}(X,Y) = 1.533$

Parametric Decision making (Statistical) continued

- Parametric Methods can perform well in many situations but its performance is at peak (top) when the spread of each group is different.
- Goal of most classification procedures is to estimate the probabilities that a pattern to be classified belongs to various possible classes, based on the values of some feature or set of features.

  Ex1. To classify the fish on conveyor belt as salmon or sea bass

  Ex2. To estimate the probabilities that a patient has various diseases given some symptoms or lab tests

- In most cases, we decide which is the most likely class.
- We need a mathematical decision making algorithm, to obtain classification or decision.

Bayes Theorem

This refers to choosing the most likely class, given the value of feature/s. The probabilities of class membership is calculated from Bayes Theorem.

Revisiting conditional probability

Suppose that we are interested in computing the probability of event *A* and we have been told event *B* has occurred.

Then the conditional probability of *A* given *B* is defined to be:

$$P\left[A|B\right] = \frac{P\left[A \cap B\right]}{P\left[B\right]} \qquad \text{if} \quad P\left[B\right] \neq 0$$

Similarly, $P[B|A] = \frac{P[A \cap B]}{P[A]}$ if P[A] is not equal to 0

From the previous two expressions

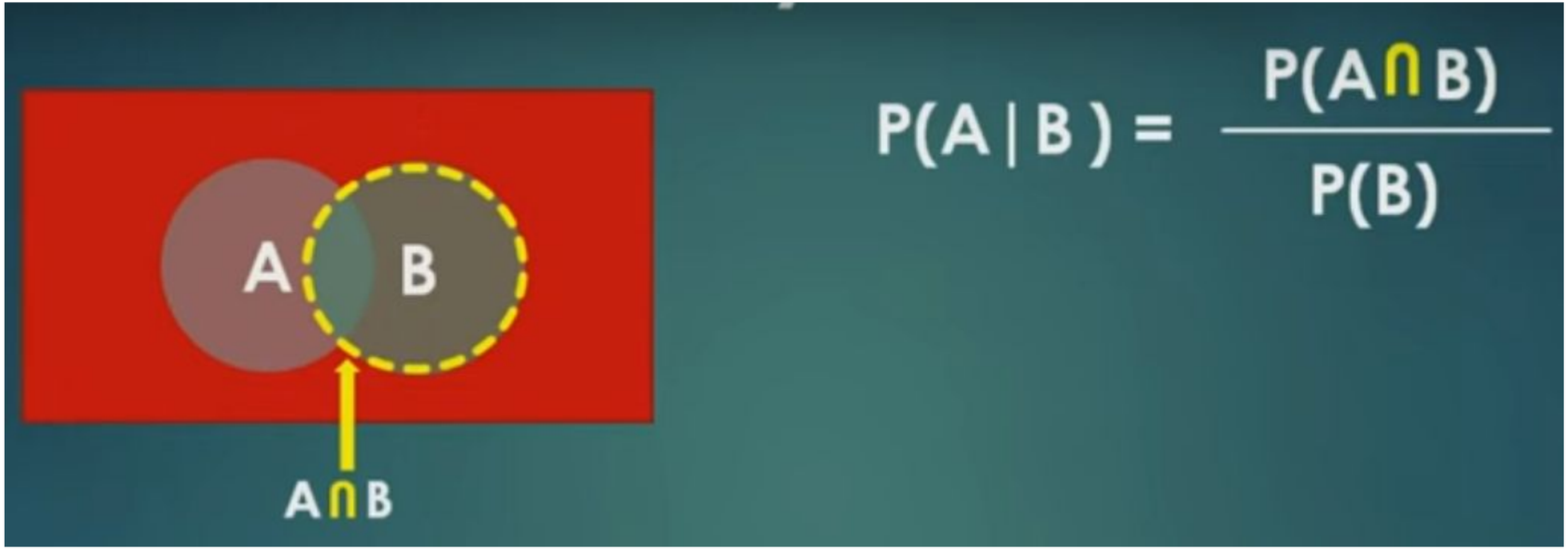
$P[A \cap B] = P[B].P[A|B]$

and $P[A \cap B] = P[A].P[B|A]$

This can also be used to calculate $P[A \cap B]$

The Multiplication Rule

- In many of the cases, P(A) may not depend on whether B has occurred. We say that the event A is independent of B if P(A) = P(A|B).
- An important consequence of the definition of independence is multiplication rule, which is obtained by substituting P(A) for P(A|B) in the above expressions
- $P[A \cap B] = P[A].P[B]$ whenever A is independent of B



- Original Sample space is the red coloured rectangular box.
- What is the probability of A occurring given sample space as B.
- Hence P(B) is in the denominator.
- And area in question is the intersection of A and B

- **Bayes Theorem:** $$P(w_i \mid X) = \frac{P(X \mid w_i)\, P(w_i)}{P(X)}$$
- P(X) is the probability distribution for feature X in the entire population. Also called unconditional density function or evidence.
- P(w_i) is the prior probability that a random sample is a member of the class w_i.
- P(X | w_i) is the class conditional probability (or likelihood) of obtaining feature value X given that the sample is from class w_i. It is equal to the number of times (occurrences) of X, if it belongs to class w_i.

$$P(\omega_j / x) = \frac{p(x/\omega_j)P(\omega_j)}{p(x)} = \frac{likelihood \times prior}{evidence}$$

Bayes rule

Likelihood **Prior**

Posterior

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)} \propto P(D|\theta) \times P(\theta)$$

Evidence

- How good are our parameters given the data
- **Prior** knowledge is incorporated and used to **update** our beliefs about the parameters

Define the terms:

- P(wi) - Prior Prob. for class wi ;
- P(X) - Prob. (Uncondl.) for feature vector X.
- P(wi |X) - Measured-conditioned or posteriori probability
- P(X | wi) - Prob. (Class-Condnl.) of feature vector X in class w

There are four parts to Bayes' Theorem: **Prior,**

Evidence,

Likelihood, and

Posterior.

Prior or State of Nature

- The a priori or prior probability reflects our knowledge of how likely we expect a certain state of nature before we can actually observe said state of nature.

Priors are known before the training process.

The state of nature is a random variable $\mathbf{P(w_i)}$.

If there are only two classes, then the sum of the priors is $\mathbf{P(w_1) + P(w_2)=1}$, if the classes are exhaustive.

In the previous fish example , it is the probability that we will see either a salmon or a sea bass next on the conveyor belt. **The prior may vary depending on the situation.**

If we get equal numbers of salmon and sea bass in a catch, then the priors are equal, or uniform.

Depending on the season, we may get more salmon than sea bass, for example.

We write P(ω = ω1) or just P(ω1) for the prior the next is a sea bass.

For c states of nature, or classes:

$$1=\sum_{i=1}^{c} P(w_i)$$

Class Conditional Probabilities

It represents the probability of how likely a feature x occurs given that it belongs to the particular class. It is denoted by, **P(X|A)** where x is a particular feature.

Sometimes, it is also known as the **Likelihood**.

It is the quantity that we have to evaluate while training the data.

During the training process, we have input(features) X labeled to corresponding class w and we figure out the likelihood of occurrence of that set of features given the class label.

Evidence

It is the probability of occurrence of a particular feature i.e. **P(X)**.

It can be calculated using the chain rule as, $\mathbf{P(X) = \Sigma_i\, p(X \mid w_i)\, P(w_i)}$

As we need the likelihood of class conditional probability is also figure out evidence values during training.

The best way to describe the evidence, P(x), is through the law of total probability.

This law states that if you have mutually exclusive events (e.g. ω1 and ω2) whose probability of occurrence sum up to 1, then the probability of some feature is the likelihood times the prior summed across all mutually exclusive events.

$$\sum_i p(\mathbf{x}|\omega_i)P(\omega_i)$$

Posterior Probabilities

It is the probability of occurrence of Class A when certain Features are given.

The result of using Bayes' Theorem is called the posterior, P(ω1|x) and P(ω2|x). The posterior represents the probability that an observation falls into class ω1 or ω2 given the measurement x.

Each observation receives a posterior probability for every class, and all the posteriors must add up to 1

$$1=\sum_{i=1}^{c} P(w_i)$$

- # Example1: Two class problem

What is the probability of a person having cold given that he or she has fever?

Cold (C) and not-cold (C'). Feature is fever (f).

Prior probability of a person having a cold, P(C) = 0.01.

Prob. of having a fever, given that a person has a cold is, P(f|C) = 0.4.

Overall prob. of fever P(f) = 0.02.

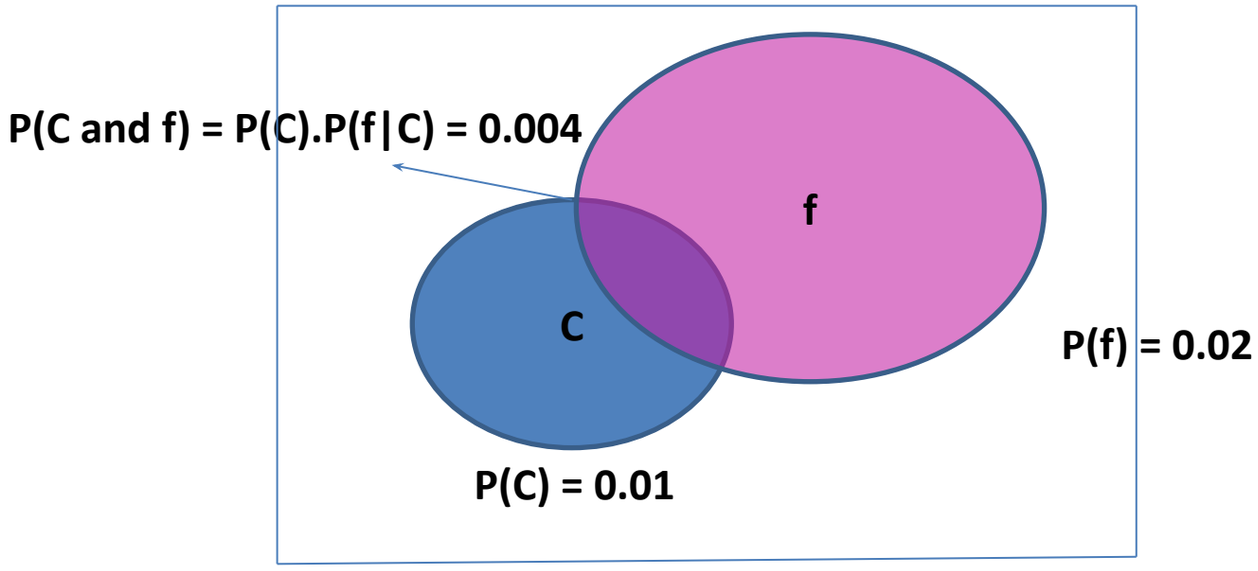
Then using Bayes Th., the Prob. that a person has a cold, given that she (or he) has a fever is:

$$P(C|f) = \frac{P(f|C)\,P(\mathbf{C})}{P(f)} \quad == \frac{0.4*0.01}{0.02} = 0.2$$

- Total Population =1000. Thus, people having cold = 10.
- People having both fever and cold = 4.
- Thus, people having only cold = 10 – 4 = 6.
- People having fever (with and without cold) = 0.02 * 1000 = 20.
- People having fever without cold = 20 – 4 = 16 (may use this later). So, probability (percentage) of people having cold along with fever, out of all those having fever, is:

  4/20 = 0.2 (20%).

Venn Diagram illustrating one feature and two class problem



Probability of a joint event - a sample comes from class C and has the feature value X:

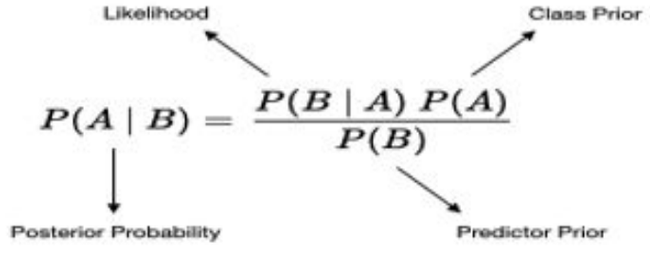
$$P(C \text{ and } X) = P(C).P(X|C) = P(X).P(C|X) = 0.01*0.4 = 0.02*0.2$$

Example2: Two class problem

Find the probability that a king is drawn, given that face card was drawn using Bayes theorem. —----------4/12

- Compute : Probability in the deck of cards (52 excluding jokers)

- Probability of (King|Face)



Bayes Theorem

- It is given by P(King|Face) = P(Face|King) * P(King)/ P(Face)

  = 1 * (4/52) / (12/52)

  = 1/3

Example-2

10% of patients in a clinic have liver disease. Five percent of the clinic's patients are alcoholics. Amongst those patients diagnosed with liver disease, 7% are alcoholics. You are interested in knowing the probability of a patient having liver disease, given that he is an alcoholic.

P(A) = probability of liver disease = 0.10
P(B) = probability of alcoholism = 0.05
P(B|A) = 0.07
P(A|B) = ?

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} = \frac{0.07 \times 0.10}{0.05} = 0.14$$

In other words, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%)

Example-4.

A disease occurs in 0.5% of the population

A diagnostic test gives a positive result in:

- 99% of people with the disease
- 5% of people without the disease (false positive)

If a person receives a positive result, What is the probability of him having the disease, given a positive result?

$$P(disease|positive\ test) = \frac{P(positive\ test|disease) \times P(disease)}{P(positive\ test)}$$

We know:

$P(positive\ test|disease)$ = 0.99
$P(disease)$ = 0.005
$P(positive\ test)$ = ???

$$P(positive\ test) = P(PT|D) \times P(D) + P(PT|{\sim}D) \times P({\sim}D)$$
$$= (0.99 \times 0.005) + (0.05 \times 0.995) = 0.0547$$

Where:

$P(D) =$ chance of having the disease

$P({\sim}D) =$ chance of not having the disease

Remember: $P({\sim}D) = 1 - P(D)$

$P(PT|D) =$ chance of positive test given that disease is present

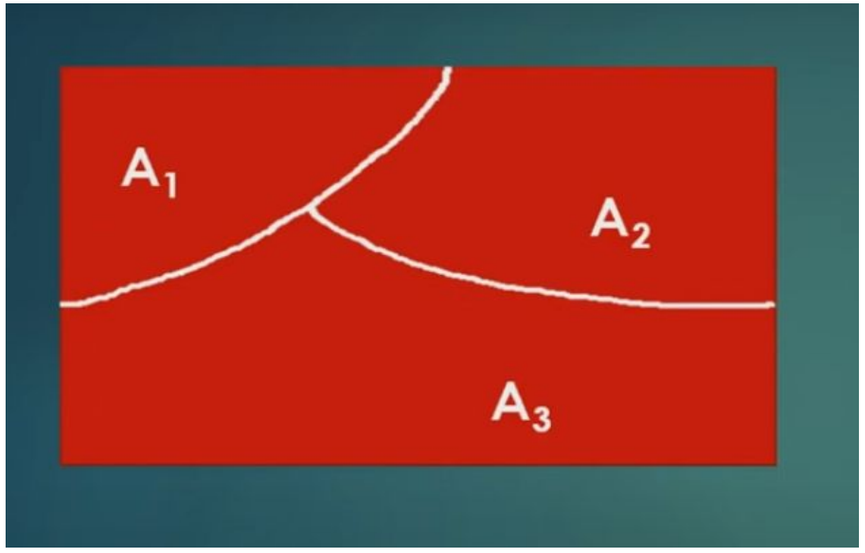
$P(PT|{\sim}D) =$ chance of positive test given that the disease isn't present

P(positive test|~disease) = 0.05

Therefore:

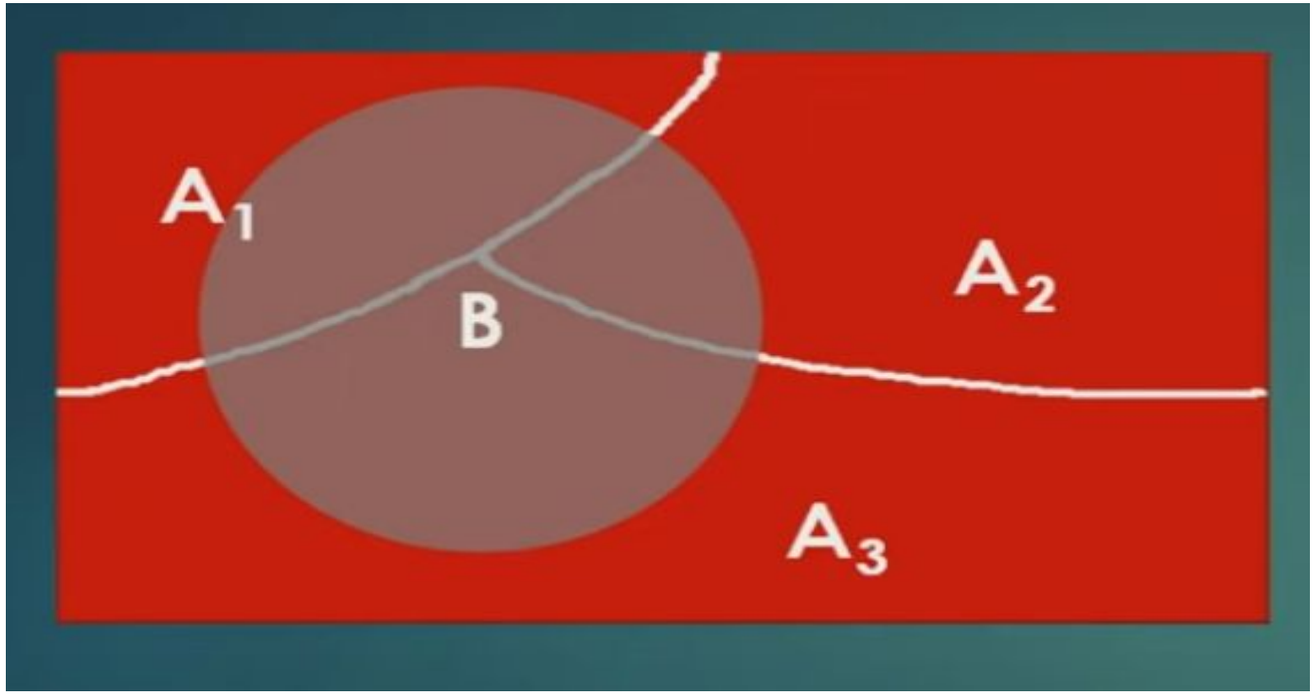
$$P(disease|positive\ test) = \frac{0.99\ \times 0.005}{0.0547} = 0.09$$
$$i.e.\ 9\%$$

Generalized Bayes Theorem



- Consider we have 3 classes A1, A2 and A3.
- Area under Red box is the sample space
- Consider they are mutually exclusive and collectively exhaustive.
- Mutually exclusive means, if one event occurs then another event cannot happen.
- Collectively exhaustive means, if we combine all the probabilities, i.e P(A1), P(A2) and P(A3), it gives the sample space, i.e the total rectangular red coloured space.

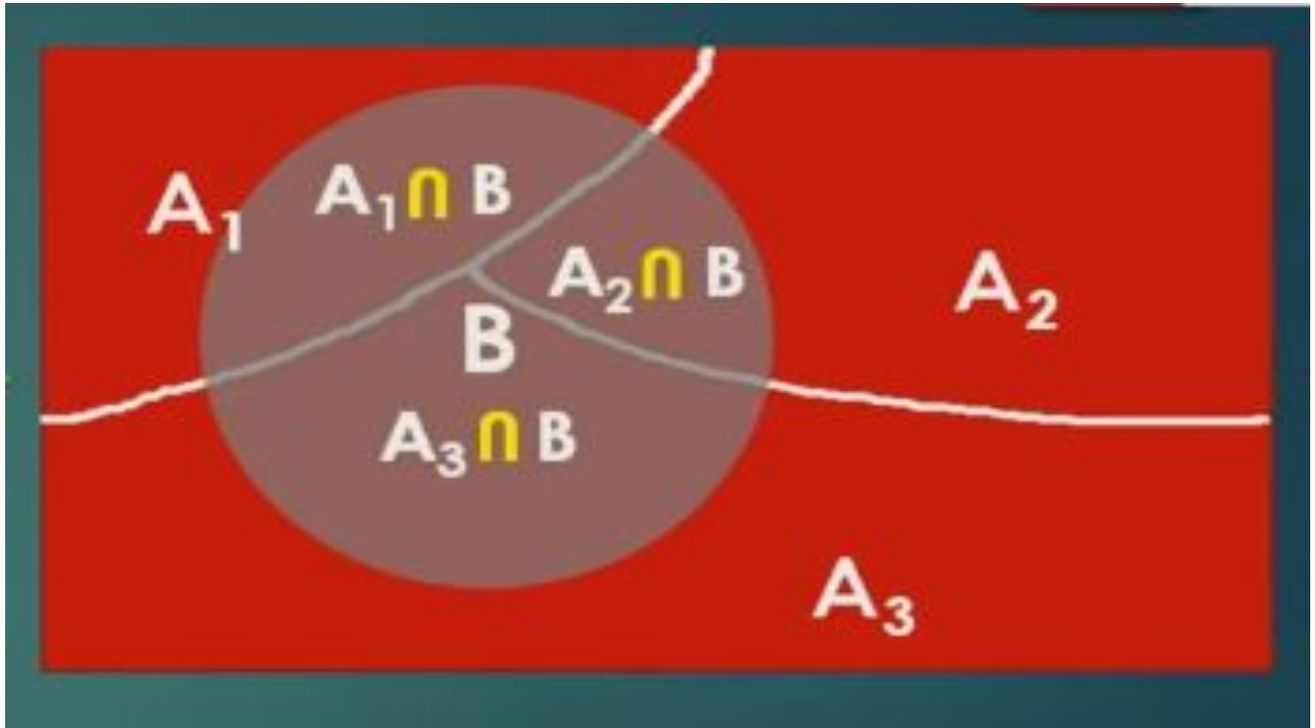
- Consider now another event B occurs over A1,A2 and A3.
- Some area of B is common with A1, and A2 and A3.
- It is as shown in the figure below:



- Portion common with A1 and B is shown by: $P(A_1 \cap B)$
- Portion common with A2 and B is given by : $P(A_2 \cap B)$
- Portion common with A3 and B is given by: $P(A_3 \cap B)$

- Probability of B in total can be given by

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$



- Remember :

$$P(A \cap B) = P(A \mid B) * P(B) = P(B \mid A) * P(A)$$

- Equation from the previous slide:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

- Replacing first in the second equation in this slide, we will get:

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

$$P(B) = P(B \mid A_1) * P(A_1) + P(B \mid A_2) * P(A_2) + P(B \mid A_3) * P(A_3)$$

Further simplified P(B)

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

$$P(B) = P(B \mid A_1)*P(A_1) + P(B \mid A_2)*P(A_2) + P(B \mid A_3)*P(A_3)$$

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)*P(A_i)$$

Arriving at Generalized version of Bayes theorem

$$P(A_i \mid B) = \frac{P(B \mid A_i)*P(A_i)}{P(B)}$$

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)*P(A_i)$$

$$P(A_i \mid B) = \frac{P(B \mid A_i)*P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i)*P(A_i)}$$

Example-5.

Given 1% of people have a certain genetic defect

90% tests positive for the gene defect (true positives).

9.6% of the tests are false positives

If a person gets a positive test result, **what are the odds they actually have the genetic defect?**

A = chance of having the faulty gene. That was given in the question as 1%.

That also means the probability of *not* having the gene (~A) is 99%.

X = A positive test result.

P(A|X) = Probability of having the gene given a positive test result.

P(X|A) = Chance of a positive test result given that the person actually has the gene = 90%.

p(X|~A) = Chance of a positive test if the person *doesn't* have the gene. That was given in the question as 9.6%

Now we have all of the information we need to put into the equation:

P(A|X) = (.9 * .01) / (.9 * .01 + .096 * .99) = 0.0865 (8.65%).

The probability of having the faulty gene on the test is 8.65%.

Example-6

Given the following statistics, what is the probability that a woman has cancer if she has a positive mammogram result?

One percent of women over 50 have breast cancer.

Ninety percent of women who have breast cancer test positive on mammograms.

Eight percent of women will have false positives.

Let women having cancer is W and

Positive test result is PT.

- P(W)=0.01
- P(~W)=0.99
- P(PT|W)=0.9
- P(~PT|w)=0.08 Compute P(testing positive)

(0.9 * 0.01) / ((0.9 * 0.01) + (0.08 * 0.99) = 0.10.

Example-7 Box P has 2 red balls and 3 blue balls and box Q has 3 red balls and 1 blue ball. A ball is selected as follows:

(i) Select a box (ii) Choose a ball from the selected box such that each ball in the box is equally likely to be chosen. The probabilities of selecting boxes P and Q are (1/3) and (2/3), respectively.

Given that a ball selected in the above process is a red ball, the probability that it came from the box P

Solution:

R --> Event that red ball is selected

B --> Event that blue ball is selected

P --> Event that box P is selected

Q --> Event that box Q is selected

We need to calculate P(P|R)?

$$P(P|R) = \frac{P(R|P)P(P)}{P(R)}$$

$P(R|P)$ = A red ball selected from box P

$= 2/5$

$P(P) = 1/3$

$P(R) = P(P)*P(R|P) + P(Q)*P(R|Q)$

$= (1/3)*(2/5) + (2/3)*(3/4)$

$= 2/15 + 1/2$

$= 19/30$

Putting above values in the Bayes's Formula

$P(P|R) = (2/5)*(1/3) / (19/30)$

$= 4/19$

Example-8 An aircraft emergency locator transmitter (ELT) is a device designed to transmit a signal in the case of a crash. The Altigauge Manufacturing Company makes 80% of the ELTs, the Bryant Company makes 15% of them, and the Chartair Company makes the other 5%. The ELTs made by Altigauge have a 4% rate of defects, the Bryant ELTs have a 6% rate of defects, and the Chartair ELTs have a 9% rate of defects (which helps to explain why Chartair has the lowest market share).

If a randomly selected ELT is then tested and is found to be defective, find the probability that it was made by the Altigauge Manufacturing Company.

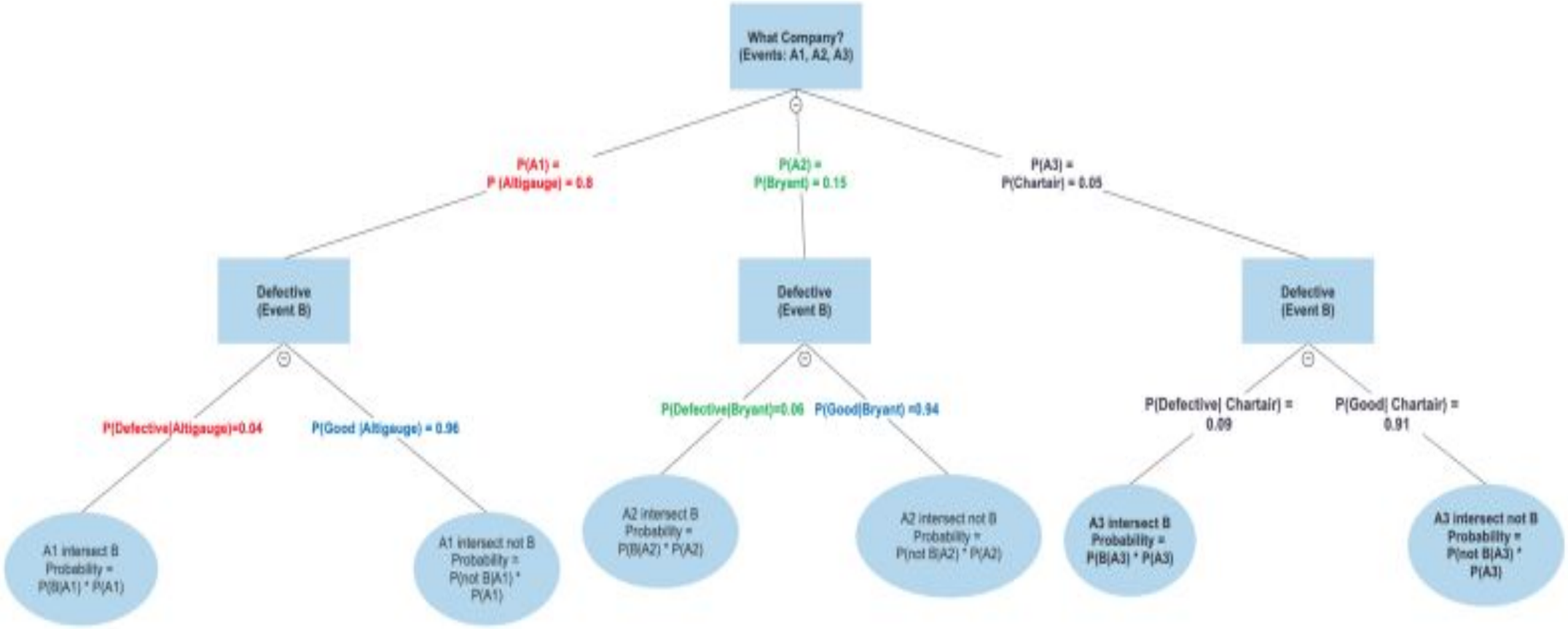
We need to find P(Altigauge|Defective).

P(Altigauge|Defective) = P(Defective|Altigauge) * P(Altigauge) / P(Defective)

P(Defective|Altigauge) * P(Altigauge) = 0.04*0.8 = 0.032

P(Defective) = 0.04*0.8 + 0.06*0.15 + 0.09*0.05 = 0.0455

P(Altigauge|Defective) = 0.032 / 0.0455 = 0.7032


What Company?
(Events: A1, A2, A3)
P(A1) =
P (Altigauge) = 0.8
P(A2) =
P(Bryant) = 0.15
P(A3) =
P(Chartair) = 0.05
Defective
(Event B)
Defective
(Event B)
Defective
(Event B)
P(Defective|Altigauge)=0.04
P(Good |Altigauge) = 0.96
P(Defective|Bryant)=0.06
P(Good|Bryant) =0.94
P(Defective| Chartair) =
0.09
P(Good| Chartair) =
0.91
A1 intersect B
Probability =
P(B|A1) * P(A1)
A1 intersect not B
Probability =
P(not B|A1) *
P(A1)
A2 intersect B
Probability =
P(B|A2) * P(A2)
A2 intersect not B
Probability =
P(not B|A2) * P(A2)
A3 intersect B
Probability =
P(B|A3) * P(A3)
A3 intersect not B
Probability =
P(not B|A3) *
P(A3)

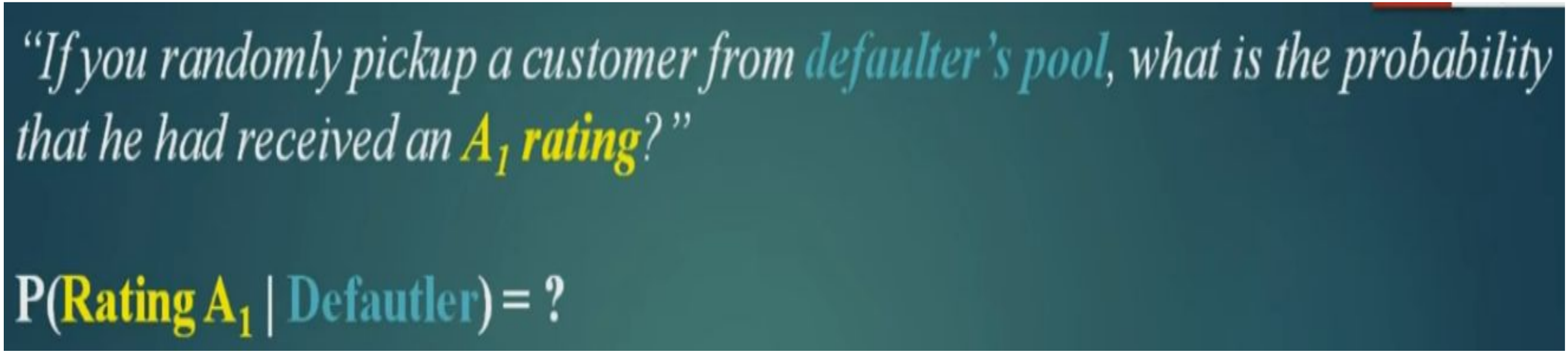
Problem on Bayes theorem with 3 class case

n order to manage the Credit Risk, a bank regularly rates each of its borrowers as A_1 or A_2 or A_3, based on their Credit history. A_1 implies lowest risk and A_3 mplies highest risk. Risk means the chance that a borrower might fail to payback he loan amount.

Based on historical data, on an average, 30% customers are rated A_1, 60% are ated A_2, and 10% are rated A_3. It was found that 1% of the customers who were ated A_1, 10% of the customers who were rated A_2, and 18% of the customers who were rated A_3, eventually became defaulters(failed to payback).

f you randomly pickup a customer from defaulter's pool, what is the probability hat he had received an A_1 rating?

- While solving problem based on Bayes theorem, we need to split the given information carefully:
- Asked is:



- Note, the flip of what is asked will be always given:

Note: Flip of what is being asked i.e. P(Defautler | Rating A_1) will always be given in such problems.

*"It was found that **1% of the customers who were rated** A_1, 10% of the customers who were rated* A_2*, and 18% of the customers who were rated* A_3*, eventually became defaulters (failed to payback)."*

P(Defautler | Rating A_1) = 1% or 0.01
P(Defautler | Rating A_2) = 10% or 0.10
P(Defautler | Rating A_3) = 18% or 0.18

- What else is given:

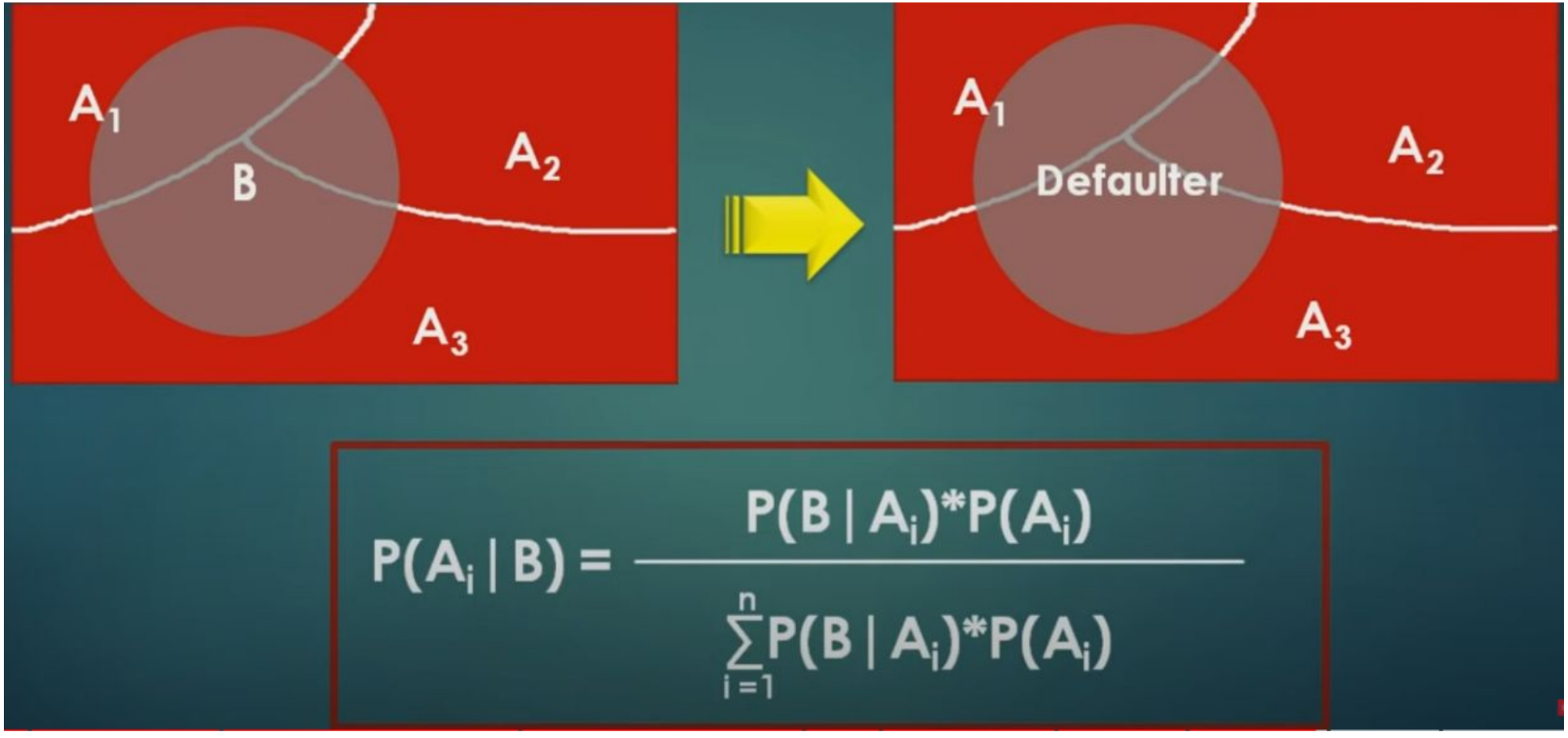
"Based on historical data, on an average, 30% customers are rated A_1, 60% are rated A_2, and 10% are rated A_3."

- Represented by:

$$P(\text{Rating } A_1) = 30\% \text{ or } 0.30$$
$$P(\text{Rating } A_2) = 10\% \text{ or } 0.60$$
$$P(\text{Rating } A_3) = 18\% \text{ or } 0.10$$

Note: $P(\text{Rating } A_1) + P(\text{Rating } A_2) + P(\text{Rating } A_3) = 0.30 + 0.60 + 0.10 = 1$

So.. Given Problem can be represented as:



$$P(A_i \mid B) = \frac{P(B \mid A_i)*P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i)*P(A_i)}$$

Numerator:

$$P(\text{Defautler} \mid \text{Rating } A_1)*P(\text{Rating } A_1) = 0.01*0.30 = 0.003$$

Denominator:

$$\sum_{i=1}^{n} P(B \mid A_i)*P(A_i) = P(B \mid A_1)*P(A_1) + P(B \mid A_2)*P(A_2) + P(B \mid A_3)*P(A_3)$$
$$= 0.01*0.30 + 0.10*0.60 + 0.18*.10$$
$$= 0.081$$

$$P(\text{Rating } A_1 \mid \text{Defautler}) = \frac{0.003}{0.081} = 0.0370 \text{ or } 3.7\%$$

Decision Regions

- An alternative for classifying samples by comparing posterior probabilities $P(C_i|x)$ of the classes using

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

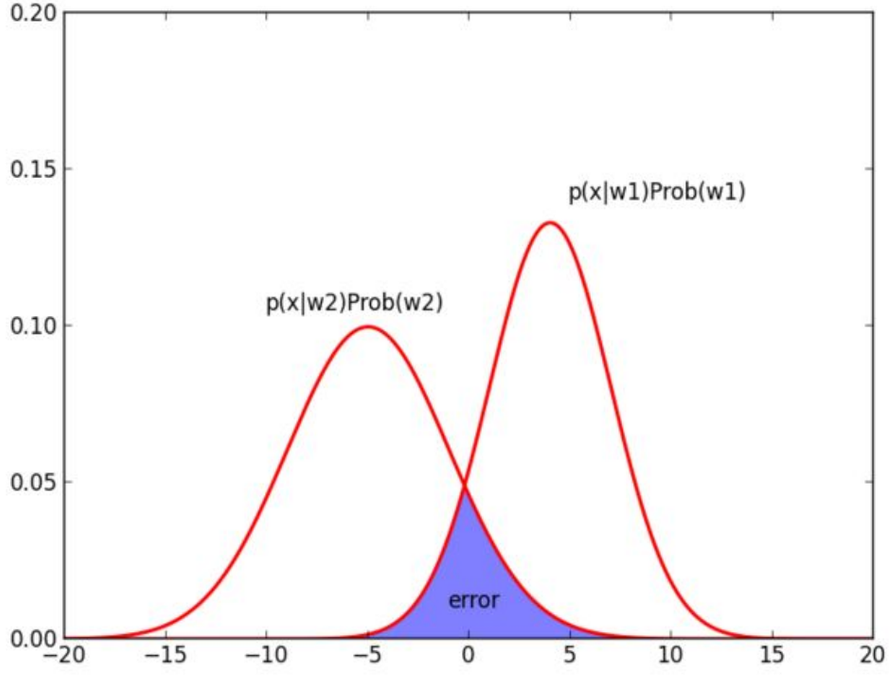
to calculate the decision regions or ranges of x in advance in which a particular class is most likely.

- Each decision is associated with a class.
- To classify the sample with feature value x, we determine which decision region contains x and assign x to the class identified with that region.
- A boundary between the decision regions is called decision boundary
- Optimal decision boundaries separate the feature space into decision regions R1, R2…..Rn such that class Ci is the most probable for values of x in Ri than any other region

- A decision rule prescribes what action to take based on observed input.
- Idea Check: What is a reasonable Decision Rule if the only available information is the prior, and the cost of any incorrect classification is equal?
- Decide ω1 if P(A) > P(B); otherwise decide ω2.

- Decision or Classification algorithm according to Baye's Theorem:

Choose:
$$\begin{cases} A & \text{if } P(X|A)P(A) > P(X|B)P(B) \\ B & \text{if } P(X|B)P(B) > P(X|A)P(A) \end{cases}$$



- To compute the optimal decision boundary between two classes A and B, we can equate the posterior probabilities if the densities are continuous and overlapping **P(A|x) = P(B|x).**
- Substituting Bayes Theorem and cancelling p(x) term **P(A)p(x|A)** = **P(B)p(x|B)**
- If the feature x in both the classes are normally distributed
- $P(A)\frac{1}{\sigma A\sqrt{2\pi}}e^{-(x-\mu A)^2/2\sigma A^2} = P(B)\frac{1}{\sigma B\sqrt{2\pi}}e^{-(x-\mu B)^2/2\sigma B^2}$
-
- Cancelling $\sqrt{2\pi}$ and taking natural logarithm

- $-2\ln(P(A)/\sigma A) + (\frac{x-\mu A}{\sigma A})^2 = -2\ln(P(B)/\sigma B) + (\frac{x-\mu B}{\sigma B})^2$

-
- $D = -2\ln(P(A)/\sigma A) + (\frac{x-\mu A}{\sigma A})^2 + 2\ln(P(B)/\sigma B) + (\frac{x-\mu B}{\sigma B})^2$
- D equals 0 on the decision boundary;
- D is positive in the decision region in which B is most likely the class;
- and D is negative in the decision region in which A is most likely.

What is our probability of error?

- For the two class situation, we have
- P(error|x) = { P(A|x) if we decide B
  { P(B|x) if we decide A
- We can minimize the probability of error by following the posterior:

  Decide A if P(A|x) > P(B|x)

Probability of error becomes P(error|x) = min [P(A|x), P(B|x)]

Equivalently, Decide A if p(x|A)P(A) > p(x|B)P(B);

otherwise decide B I.e., the evidence term is not used in decision making.

Conversely, if we have uniform priors, then the decision will rely exclusively on the likelihoods.

T**ake Home Message: Decision making relies on both the priors and the likelihoods** and Bayes Decision Rule combines them to achieve the minimum probability of error.

Example 3.4 *Computing optimal one-dimensional decision boundaries.*

To compute the decision boundaries for Example 3.3, we substitute the following da into (3.15): $\mu_G = 26$, $\sigma_G = 2$, $\mu_{\overline{G}} = 22$, $\sigma_{\overline{G}} = 3$, $P(G) = 0.8$, and $P(\overline{G}) = 0.2$. Th produces

$$-2\ln(0.8/2) + \left(\frac{x-26}{2}\right)^2 = -2\ln(0.2/3) + \left(\frac{x-22}{3}\right)^2 \quad (3.15)$$

$$2 \cdot 36 \ln\left(\frac{0.2 \cdot 2}{0.8 \cdot 3}\right) = 4(x-22)^2 - 9(x-26)^2$$

$$5x^2 - 292x + 4018.99 = 0,$$

so

$$x = \frac{292 \pm \sqrt{292^2 - 4 \cdot 5 \cdot 4018.99}}{2 \cdot 5} = 22.2 \text{ and } 36.2.$$

These two decision boundaries partition the feature space into three decision regi and produce the following decision rule:

1. Classify the sample as class G if $22.2 < x < 36.2$.
2. Classify the sample as class $\overline{G}$ if $x < 22.2$ or $36.2 < x$.

Multiple Features

- A single feature may not discriminate well between classes.
- Recall the example of just considering the length of fish Vs considering width and lightness to reduce the misclassification (better classification).
- If the joint conditional density of multiple features is known for each class, Bayesian classification is very similar to classification with one feature.
- Replace the value of single feature x by feature vector X which has single feature as the component.
- $P(w_i \mid X) = \frac{P(w_i)P(X \mid w_i)}{\sum_{j=1}^{k} P(w_j)P(x \mid w_j)}$ for single feature
- $P(w_i \mid X) = \frac{P(w_i)p(X \mid w_i)}{\sum_{j=1}^{k} P(w_j)p(x \mid w_j)}$
- For multiple features with Vector X replaces the conditional probabilities $P(X|W_i)$ by the conditional densities $p(x|w_i)$

Conditional Independence

- Event A and B are ***conditionally independent given C*** in case

  Pr(AB|C)=Pr(A|C)Pr(B|C)
- A set of events $\{A_i\}$ is conditionally independent given C in case

$$\Pr(\bigcap_i A_i \mid C) = \prod_i \Pr(A_i \mid C)$$

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Solution

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Example. ‘Play Tennis’ data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
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?

Based on the examples in the table, classify the following datum ***x***:

x=(Outl=Sunny, Temp=Cool, Hum=High, Wind=strong)

- That means: Play tennis or not?

$$h_{NB} = \arg\max_{h \in [yes, no]} P(h)P(\mathbf{x} \mid h) = \arg\max_{h \in [yes, no]} P(h)\prod_t P(a_t \mid h)$$

$$= \arg\max_{h \in [yes, no]} P(h)P(Outlook = sunny \mid h)P(Temp = cool \mid h)P(Humidity = high \mid h)P(Wind = strong \mid h)$$

- Working:

$$P(PlayTennis = yes) = 9/14 = 0.64$$

$$P(PlayTennis = no) = 5/14 = 0.36$$

$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$

$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$

$$etc.$$

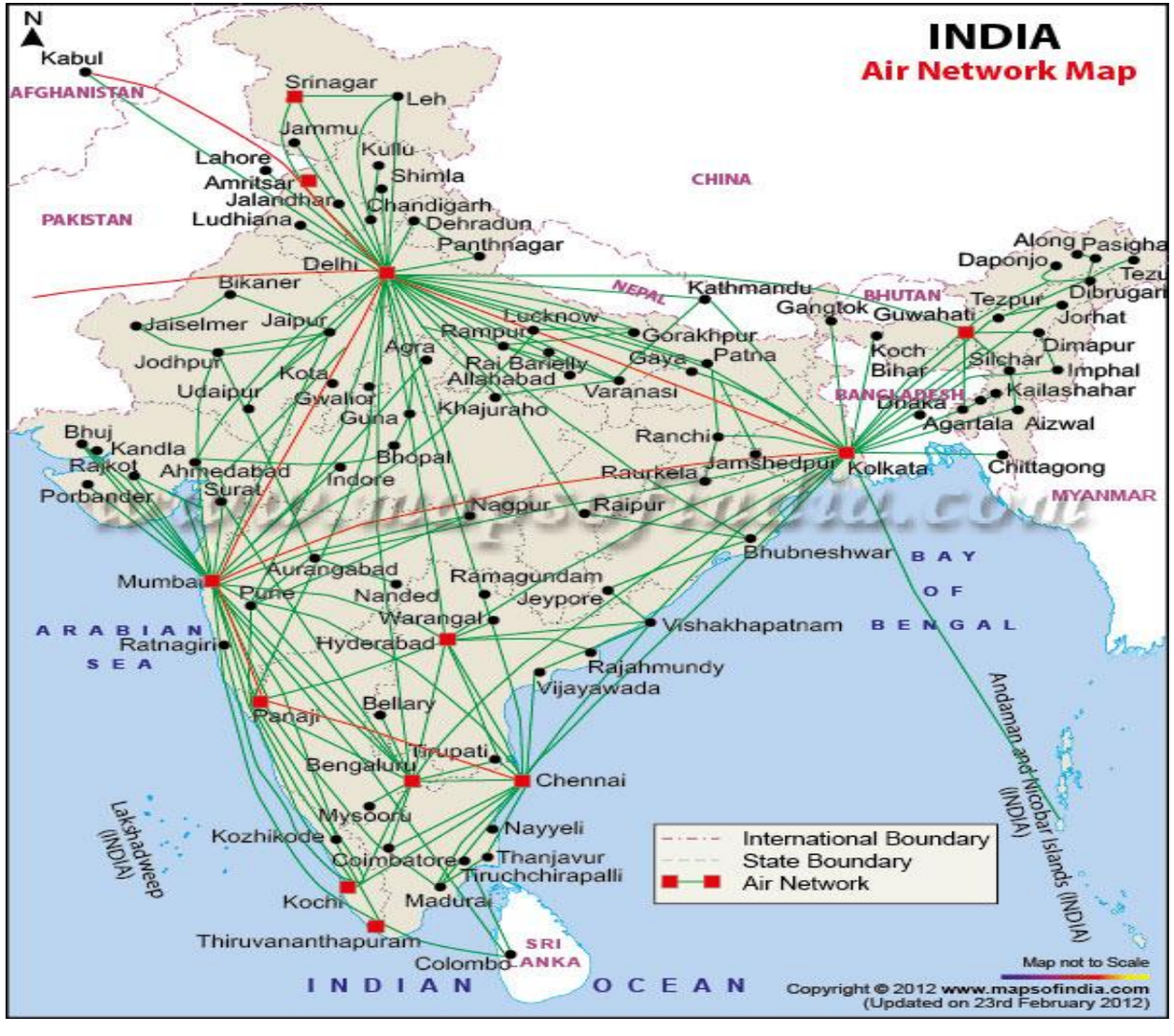
$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$

$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = \mathbf{0.0206}$$

$$\Rightarrow answer : PlayTennis(x) = no$$

Example: Bayesian Classification

- **Example 8.2:** Air Traffic Data
  - Let us consider a set observation recorded in a database
    - Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time

Cond. to next slide…

Air-Traffic Data

Cond. from previous slide…

Days	Season	Fog	Rain	Class
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Heavy	On Time

Air-Traffic Data

- In this database, there are four attributes

  A = [Day, Season, Fog, Rain]

  with 20tuples.
- The categories of classes are:

  C= [On Time, Late, Very Late, Cancelled]

- Given this is the knowledge of data and classes, we are to find most likely classification for any other unseen instance, for example:

- Classification technique eventually to map this tuple into an accurate class.

  Week Day Winter High None ???

Naïve Bayesian Classifier

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	½ = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	½ = 0.5	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	Holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
Season	Spring	4/14 = 0.29	0/2 = 0	0/3 = 0	0/1 = 0
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3= 0.33	0/1 = 0
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

Naïve Bayesian Classifier

Attribute		Class			
		On Time	Late	Very Late	Cancelled
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	None	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	Heavy	1/14 = 0.07	1/2 = 0.5	2/3 = 0.67	1/1 = 1
Prior Probability		14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

Naïve Bayesian Classifier

Instance:

Week Day	Winter	High	Heavy	???

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.50 = 0.0125

Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

Case3 is the strongest; Hence correct classification is **Very Late**

suppose you are
will play a gam
weather-related

end

Temperature	Wind	Sunshine	Play
Cold	Strong	Cloudy	No
Warm	Strong	Cloudy	No
Warm	None	Sunny	Yes
Hot	None	Sunny	No
Hot	Breeze	Cloudy	Yes
Warm	Breeze	Sunny	Yes
Cold	Breeze	Cloudy	No
Cold	None	Sunny	Yes
Hot	Strong	Cloudy	Yes
Warm	None	Cloudy	Yes
Warm	Strong	Sunny	?

There is a 80% chance that Ashish takes bus to the school and there is a 20% chance that his father drops him to school. The probability that he is late to school is 0.5 if he takes the bus and 0.2 if his father drops him. On a given day, Ashish is late to school. **Find the probability that his father dropped him to school on that day.** *Express your answer as a number, rounded to two decimal places.*

P(E1-Bus) = 0.8

P(E2-Father) = 0.2

Let A denote the event Ashish late to school, for 2 possibilities E1-Bus and E2-Father, with probabilities 0.5 and 0.2 respectively.

Then P(A|E1)*P(E1) = 0.8*0.5

P(A|E2)*P(E2) = 0.2*0.2

P(A) = P(A|E1)*P(E1) + P(A|E2)*P(E2) —> **Total probability of occurrence of A**

P(E2|A) = (P(A|E2)*P(E2))/P(A) = 0.04/0.44 = 0.09 = 9%

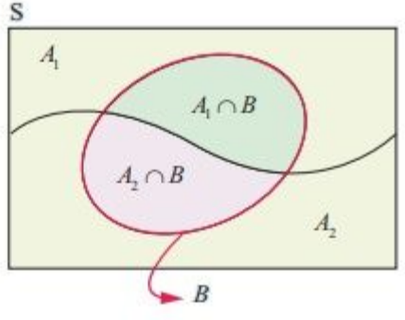
A factory has two machines I and II. Machine I produces 40% of items of the output and Machine II produces 60% of the items. Further 4% of items produced by Machine I are defective and 5% produced by Machine II are defective. An item is drawn at random. If the drawn item is defective, find the probability that it was produced by Machine II.

Solution

Let A_1 be the event that the items are produced by Machine-I, A_2 be the event that items are produced by Machine-II. Let B be the event of drawing a defective item. Now we are asked to find the conditional probability P (A_2 / B). Since A_1 , A_2 are mutually exclusive and exhaustive events, by Bayes' theorem,

$$P(A_2 / B) = \frac{P(A_2)\,P(B / A_2)}{P(A_1)\,P(B / A_1) + P(A_2)\,P(B / A_2)}$$

$$P(A_2 / B) = \frac{(0.60)(0.05)}{(0.40)(0.04) + (0.60)(0.05)} = \frac{15}{23}.$$



$$P(A_2 / B) = \frac{P(A_2)\,P(B / A_2)}{P(A_1)\,P(B / A_1) + P(A_2)\,P(B / A_2)}$$

A construction company employs 2 executive engineers. Engineer-1 does the work for 60% of jobs of the company. Engineer-2 does the work for 40% of jobs of the company. It is known from the past experience that the probability of an error when engineer-1 does the work is 0.03, whereas the probability of an error in the work of engineer-2 is 0.04. Suppose a serious error occurs in the work, which engineer would you guess did the work?

Solution: Let A_1 and A_2 be the events of job done by engineer-1 and engineer-2 of the company respectively. Let B be the event that the error occurs in the work.

We have to find the conditional probability $P(A_1 / B)$ and $P(A_2 / B)$ to compare their errors in their work.

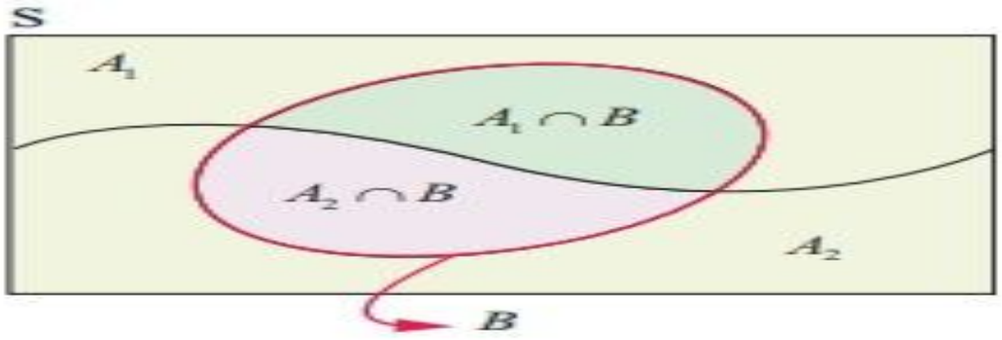
From the given information, we have

$P(A_1) = 0.60, P(B / A_1) = 0.03$

$P(A_2) = 0.40, P(B / A_2) = 0.04$

A_1 and A_2 are mutually exclusive and exhaustive events.

Applying Bayes' theorem,



$$P(A_1 / B) = \frac{P(A_1)\,P(B / A_1)}{P(A_1)\,P(B / A_1) + P(A_2)\,P(B / A_2)}$$

$$= \frac{(0.60)(0.03)}{(0.60)(0.03) + (0.40)(0.04)}$$

$$P(A_1 / B) = \frac{9}{17}.$$

$$P(A_2 / B) = \frac{P(A_2)\,P(B / A_2)}{P(A_1)\,P(B / A_1) + P(A_2)\,P(B / A_2)}$$

$$P(A_2 / B) = \frac{(0.40)(0.04)}{(0.60)(0.03) + (0.40)(0.04)}$$

$$P(A_2 / B) = \frac{8}{17}.$$

Since $P(A1 / B) > P(A2 / B)$, the chance of error done by engineer-1 is greater than the chance of error done by engineer-2. Therefore one may guess that the serious error would have been done by engineer-1

Decision Tree

- Decision tree algorithms transform raw data to rule based decision making trees.
- Herein, ID3 is one of the most common decision tree algorithm.
- Firstly, It was introduced in 1986 and it is acronym of **Iterative Dichotomiser**.
- First of all, dichotomisation means dividing into two parts containing completely different observations.
- Here, the algorithm iteratively divides attributes into two groups using the most dominant attribute to construct a tree.
- This is done by calculating the entropy and information gains of each atrribute.

Then, the most dominant attribute is put on the tree as decision node.

- Thereafter, entropy and gain scores would be calculated again for other attributes in the branch.
- Thus, the next most dominant attribute is found out.
- Finally, this procedure continues until reaching a decision for that branch.
- That’s why, it is called Iterative Dichotomiser.

***Definition**: **Entropy** is the measures of **impurity**, **disorder** or **uncertainty** in a bunch of examples.*

What an Entropy basically does?

*Entropy controls how a Decision Tree decides to **split** the data. It actually effects how a **Decision Tree** draws its boundaries.*

$$Entropy = -\sum p(X) \log p(X)$$

here p(x) is a fraction of examples in a given class

Entropy is measured between 0 and 1

The Equation of Entropy:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

For simplicity's sake let's say we only have two classes , a positive class and a negative class. Therefore 'i' here could be either + or (-).

if we had a total of 100 data points in our dataset with 30 belonging to the positive class and 70 belonging to the negative class then 'P+' would be 3/10 and 'P-' would be 7/10.

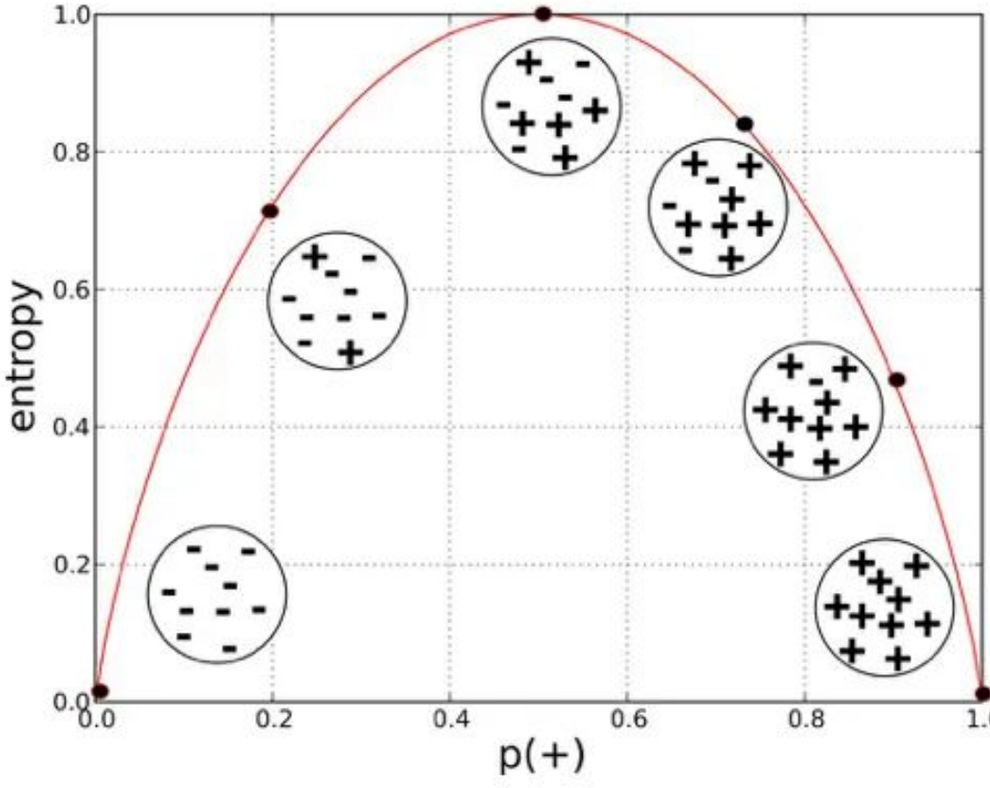
$$-\frac{3}{10} \times \log_2\left(\frac{3}{10}\right) - \frac{7}{10} \times \log_2\left(\frac{7}{10}\right) \approx 0.88$$

This indicates a high entropy, meaning a low level of purity.

The x-axis measures the proportion of data points belonging to the positive class in each bubble and the y-axis axis measures their respective entropies.

Right away, you can see the inverted ‘U’ shape of the graph.

Entropy is lowest at the extremes, when the bubble either contains no positive instances or only positive instances. That is, when the bubble is pure the disorder is 0.



Entropy is highest in the middle when the bubble is evenly split between positive and negative instances. Extreme disorder , because there is no majority.

What is Information gain and why it is matter in Decision Tree?

Next we need a metric to measure the reduction of this disorder in our target variable/class given additional information(features/independent variables) about it. This is where Information Gain comes in.

Information gain (IG) *measures how much "information" a feature gives us about the class.*

Why it matters ?

- **Information gain** is the main key that is used by **Decision Tree Algorithms** to construct a Decision Tree.
- **Decision Trees** algorithm will always tries to maximize **Information gain**.
- Information Gain is applied to quantify which feature provides maximal information about the classification based on the notion of entropy.
- An **attribute** with highest **Information gain** is split first.

The Equation of Information gain:

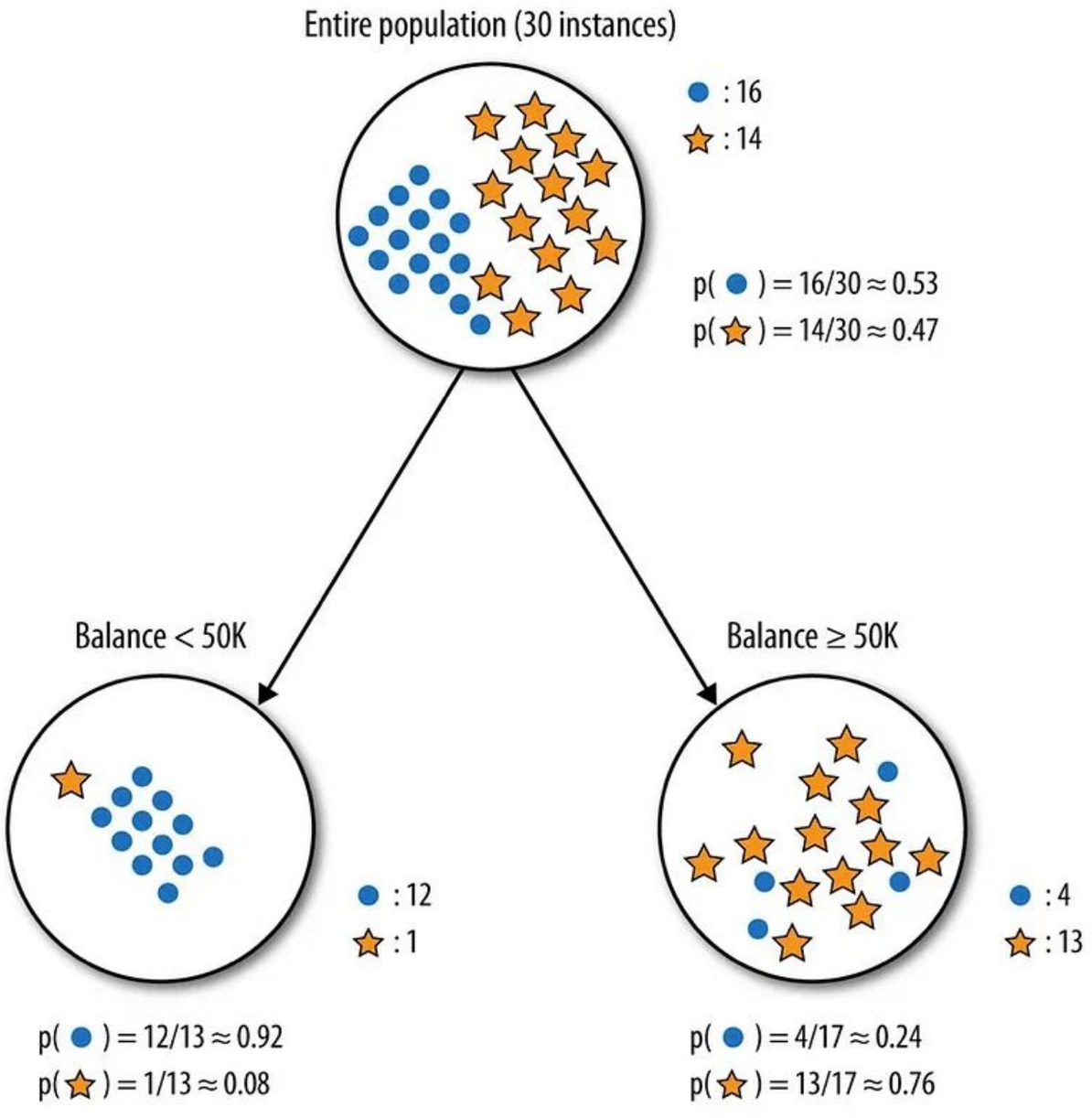
Information from X on Y

$$IG(Y, X) = E(Y) - E(Y|X)$$

Information gain = entropy (parent) – [weightes average] * entropy (children)

Example: Decision Tree

Consider an example where we are building a decision tree to predict whether a loan given to a person would result in a write-off or not. Our entire population consists of 30 instances. 16 belong to the write-off class and the other 14 belong to the non-write-off class. We have two features, namely “Balance” that can take on two values -> “< 50K” or “>50K” and “Residence” that can take on three values -> “OWN”, “RENT” or “OTHER”. The decision tree algorithm would decide what attribute to split on first and what feature provides more information, or reduces more uncertainty about our target variable out of the two using the concepts of Entropy and Information Gain.


Entire population (30 instances)
● : 16
★ : 14
p(●) = 16/30 ≈ 0.53
p(★) = 14/30 ≈ 0.47
Balance < 50K
● : 12
★ : 1
p(●) = 12/13 ≈ 0.92
p(★) = 1/13 ≈ 0.08
Balance ≥ 50K
● : 4
★ : 13
p(●) = 4/17 ≈ 0.24
p(★) = 13/17 ≈ 0.76

$$E(Parent) = -\frac{16}{30}\log_2\left(\frac{16}{30}\right) - \frac{14}{30}\log_2\left(\frac{14}{30}\right) \approx 0.99$$

$$E(Balance < 50K) = -\frac{12}{13}\log_2\left(\frac{12}{13}\right) - \frac{1}{13}\log_2\left(\frac{1}{13}\right) \approx 0.39$$

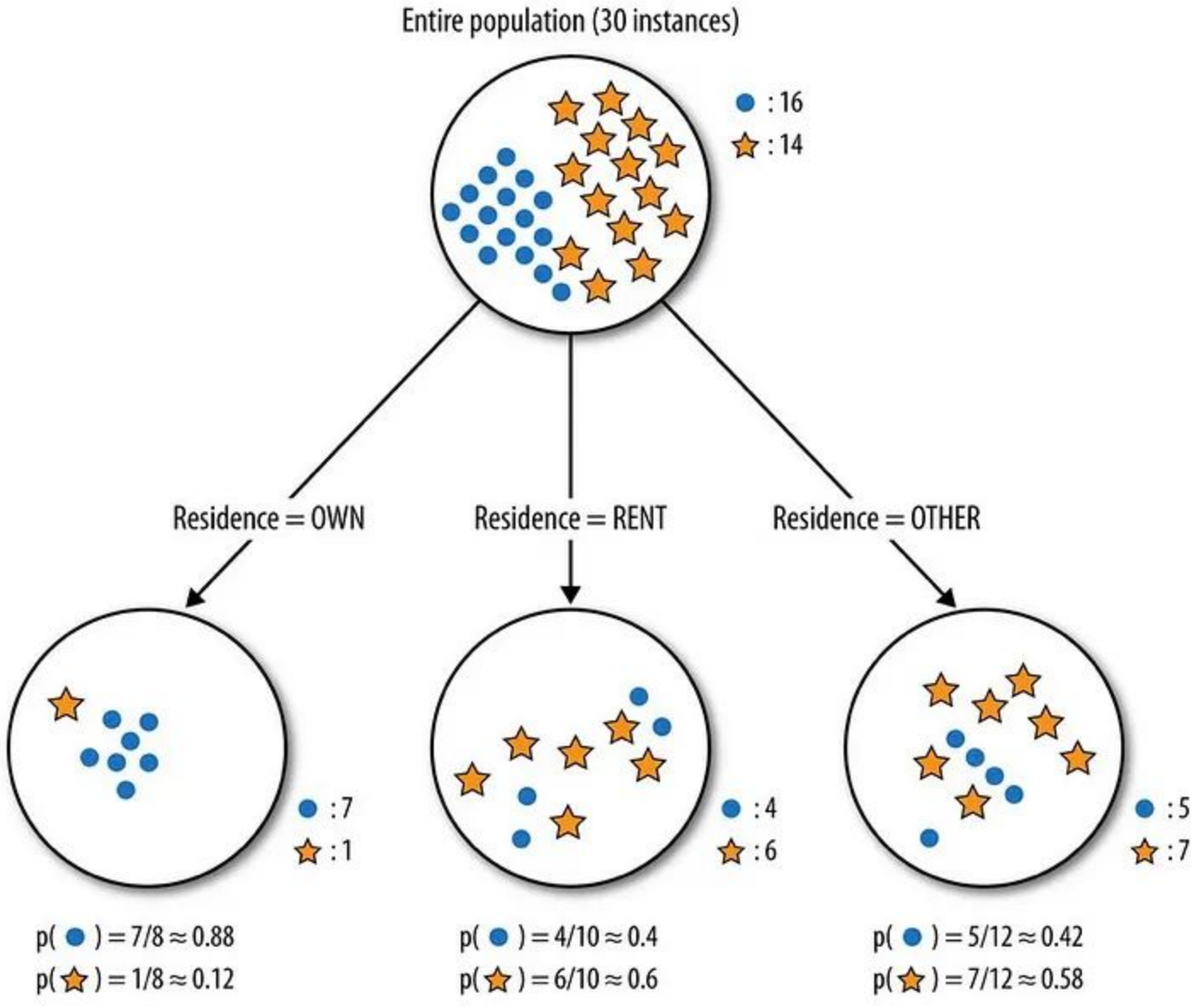
$$E(Balance > 50K) = -\frac{4}{17}\log_2\left(\frac{4}{17}\right) - \frac{13}{17}\log_2\left(\frac{13}{17}\right) \approx 0.79$$

Weighted Average of entropy for each node:

$$\begin{aligned} E(Balance) &= \frac{13}{30} \times 0.39 + \frac{17}{30} \times 0.79 \\ &= 0.62 \end{aligned}$$

Information Gain:

$$\begin{aligned} IG(Parent, Balance) &= E(Parent) - E(Balance) \\ &= 0.99 - 0.62 \\ &= \mathbf{0.37} \end{aligned}$$


Entire population (30 instances)
: 16
: 14
Residence = OWN
Residence = RENT
Residence = OTHER
: 7
: 1
: 4
: 6
: 5
: 7
p( ) = 7/8 ≈ 0.88
p( ) = 1/8 ≈ 0.12
p( ) = 4/10 ≈ 0.4
p( ) = 6/10 ≈ 0.6
p( ) = 5/12 ≈ 0.42
p( ) = 7/12 ≈ 0.58

$$E(\mathit{Residence} = \mathit{OWN}) = -\frac{7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) \approx 0.54$$

$$E(\mathit{Residence} = \mathit{RENT}) = -\frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right) \approx 0.97$$

$$E(\mathit{Residence} = \mathit{OTHER}) = -\frac{5}{12}\log_2\left(\frac{5}{12}\right) - \frac{7}{12}\log_2\left(\frac{7}{12}\right) \approx 0.98$$

Weighted Average of entropies for each node:

$$E(\mathit{Residence}) = \frac{8}{30} \times 0.54 + \frac{10}{30} \times 0.97 + \frac{12}{30} \times 0.98 = 0.86$$

Information Gain:

$$\begin{aligned} IG(\mathit{Parent}, \mathit{Residence}) &= E(\mathit{Parent}) - E(\mathit{Residence}) \\ &= 0.99 - 0.86 \\ &= 0.13 \end{aligned}$$

Consider the Following table having decision making factors to play tennis at outside for previous 14 days.

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	NO
2	Sunny	Hot	High	Strong	NO
3	Overcast	Hot	High	Weak	YES
4	Rain	Mild	High	Weak	YES
5	Rain	Cool	Normal	Weak	YES
6	Rain	Cool	Normal	Strong	NO
7	Overcast	Cool	Normal	Strong	YES
8	Sunny	Mild	High	Weak	NO
9	Sunny	Cool	Normal	Weak	YES
10	Rain	Mild	Normal	Weak	YES
11	Sunny	Mild	Normal	Strong	YES
12	Overcast	Mild	High	Strong	YES
13	Overcast	Hot	Normal	Weak	YES
14	Rain	Mild	High	Strong	NO

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Outlook

$$Values\,(Outlook) = Sunny, Overcast, Rain$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$S_{Sunny} \leftarrow [2+, 3-] \qquad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$S_{Overcast} \leftarrow [4+, 0-] \qquad Entropy(S_{Overcast}) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$$

$$S_{Rain} \leftarrow [3+, 2-] \qquad Entropy(S_{Rain}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$$

$$Gain\,(S, Outlook) = Entropy(S) - \sum_{v \in \{Sunny, Overcast, Rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Outlook)$$

$$= Entropy(S) - \frac{5}{14}Entropy(S_{Sunny}) - \frac{4}{14}Entropy(S_{Overcast})$$

$$- \frac{5}{14}Entropy(S_{Rain})$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14}0.971 - \frac{4}{14}\,0 - \frac{5}{14}0.971 = 0.2464$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Temp

$$Values\ (Temp) = Hot, Mild, Cool$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$$

$$S_{Hot} \leftarrow [2+, 2-] \qquad Entropy(S_{Hot}) = -\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4} = 1.0$$

$$S_{Mild} \leftarrow [4+, 2-] \qquad Entropy(S_{Mild}) = -\frac{4}{6} log_2 \frac{4}{6} - \frac{2}{6} log_2 \frac{2}{6} = 0.9183$$

$$S_{Cool} \leftarrow [3+, 1-] \qquad Entropy(S_{Cool}) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} = 0.8113$$

$$Gain\ (S, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Temp)$$

$$= Entropy(S) - \frac{4}{14} Entropy(S_{Hot}) - \frac{6}{14} Entropy(S_{Mild})$$

$$- \frac{4}{14} Entropy(S_{Cool})$$

$$Gain(S, Temp) = 0.94 - \frac{4}{14} 1.0 - \frac{6}{14} 0.9183 - \frac{4}{14} 0.8113 = 0.0289$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Humidity

$$Values\,(Humidity) = High, Normal$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$S_{High} \leftarrow [3+, 4-] \qquad Entropy(S_{High}) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.9852$$

$$S_{Normal} \leftarrow [6+, 1-] \qquad Entropy(S_{Normal}) = -\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7} = 0.5916$$

$$Gain\,(S, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Humidity)$$

$$= Entropy(S) - \frac{7}{14}Entropy(S_{High}) - \frac{7}{14}Entropy(S_{Normal})$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14}0.9852 - \frac{7}{14}\,0.5916 = 0.1516$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Wind

$$Values\,(Wind) = Strong, Weak$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$$

$$S_{Strong} \leftarrow [3+, 3-] \qquad Entropy(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [6+, 2-] \qquad Entropy(S_{Weak}) = -\frac{6}{8} log_2 \frac{6}{8} - \frac{2}{8} log_2 \frac{2}{8} = 0.8113$$

$$Gain\,(S, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Wind) = Entropy(S) - \frac{6}{14} Entropy(S_{Strong}) - \frac{8}{14} Entropy(S_{Weak})$$

$$Gain(S, Wind) = 0.94 - \frac{6}{14}\ 1.0 - \frac{8}{14}\ 0.8113 = 0.0478$$

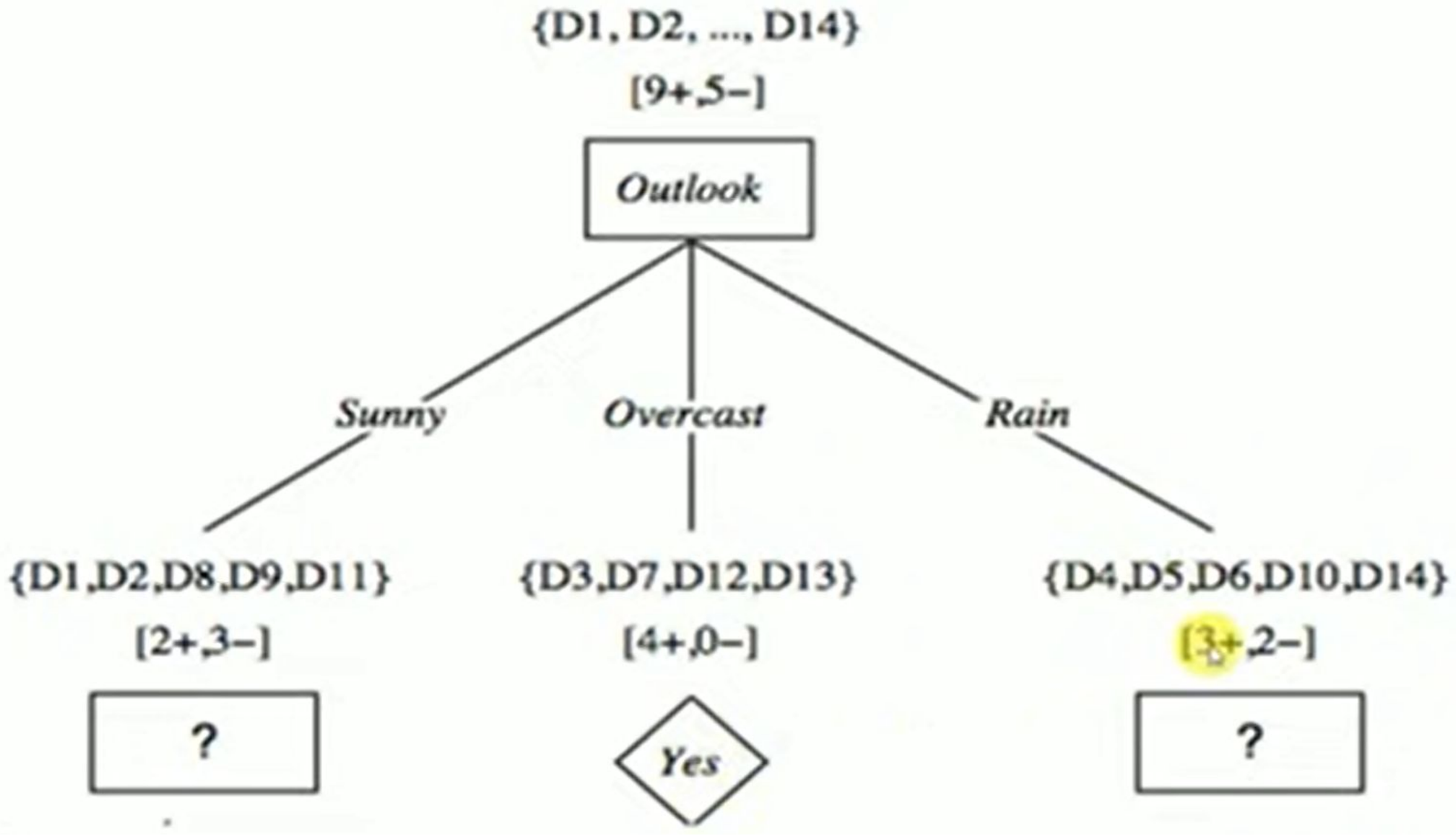
Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	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

$$Gain(S, Outlook) = 0.2464$$

$$Gain(S, Temp) = 0.0289$$

$$Gain(S, Humidity) = 0.1516$$

$$Gain(S, Wind) = 0.0478$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
{D1,D2,D8,D9,D11}
[2+,3−]
?
{D3,D7,D12,D13}
[4+,0−]
Yes
{D4,D5,D6,D10,D14}
[3+,2−]
?

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

Attribute: Temp

$$Values\ (Temp) = Hot, Mild, Cool$$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{Hot} \leftarrow [0+, 2-]$ $\quad Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Mild}) = 1.0$

$S_{Cool} \leftarrow [1+, 0-]$ $\quad Entropy(S_{Cool}) = 0.0$

$$Gain\ (S_{Sunny}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Temp)$$

$$= Entropy(S) - \frac{2}{5}Entropy(S_{Hot}) - \frac{2}{5}Entropy(S_{Mild})$$

$$- \frac{1}{5}Entropy(S_{Cool})$$

$$Gain(S_{sunny}, Temp) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1 - \frac{1}{5}0.0 = 0.570$$

Day	Temp	Humidity	Wind	Play Tennis
Dl	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
Dl1	Mild	Normal	Strong	Yes

Attribute: Humidity

$$Values\,(Humidity) = High, Normal$$

$$S_{Sunny} = [2+, 3-] \qquad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$$

$$S_{high} \leftarrow [0+, 3-] \qquad Entropy(S_{High}) = 0.0$$

$$S_{Normal} \leftarrow [2+, 0-] \qquad Entropy(S_{Normal}) = 0.0$$

$$Gain\,(S_{Sunny}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Humidity) = Entropy(S) - \frac{3}{5}Entropy(S_{High}) - \frac{2}{5}Entropy(S_{Normal})$$

$$Gain(S_{sunny}, Humidity) = 0.97 - \frac{3}{5}\,0.0 - \frac{2}{5}0.0 = 0.97$$

Day	Temp	Humidity	Wind	Play Tennis
Dl	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
Dl1	Mild	Normal	Strong	Yes

Attribute: Wind

$$Values\,(Wind) = Strong, Weak$$

$$S_{Sunny} = [2+, 3-] \qquad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$$

$$S_{Strong} \leftarrow [1+, 1-] \qquad Entropy(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [1+, 2-] \qquad Entropy(S_{Weak}) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9183$$

$$Gain\,(S_{Sunny}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

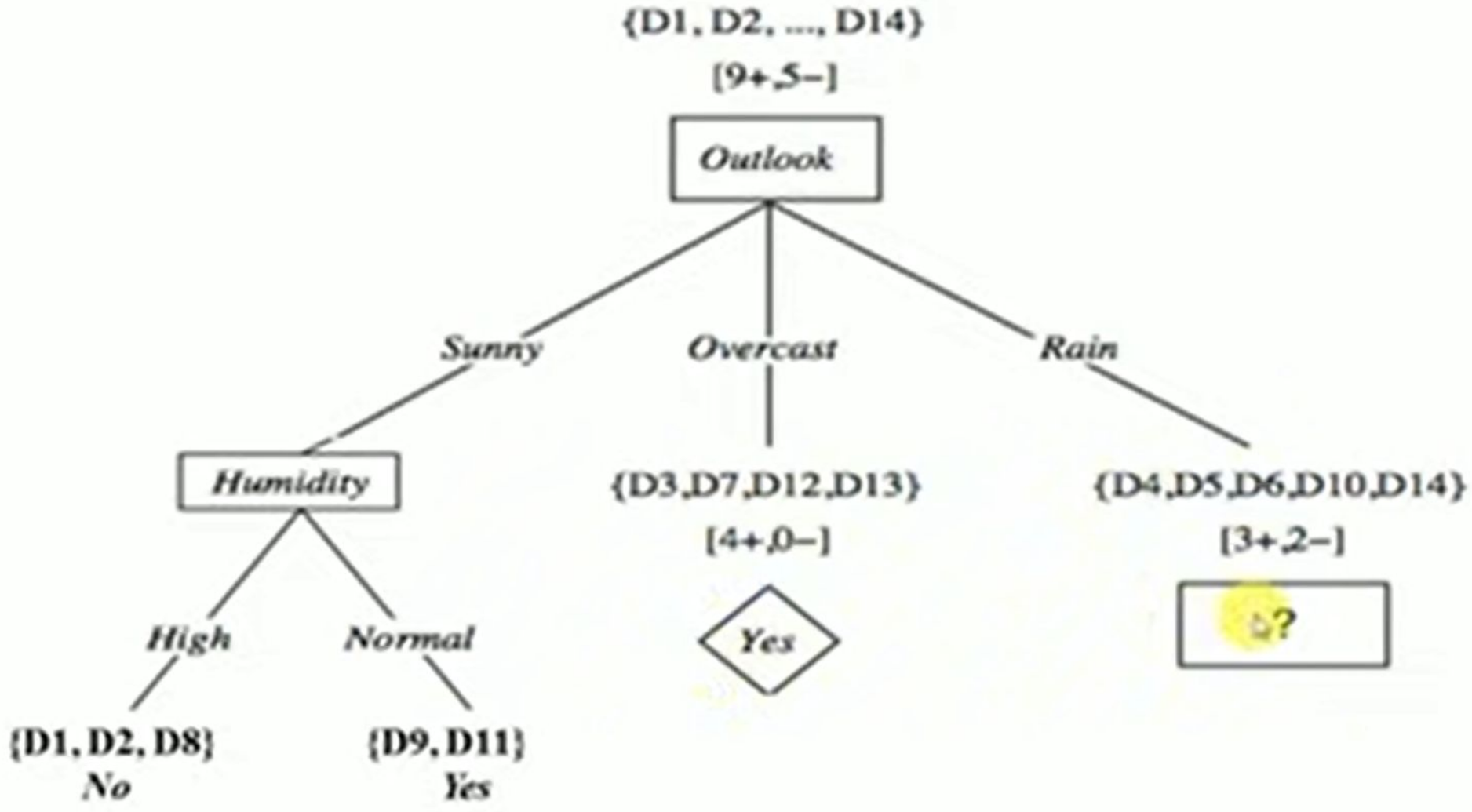
$$Gain(S_{sunny}, Wind) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}\,0.918 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D1	Hot	High	Weak	No
D2	Hot	High	Strong	No
D8	Mild	High	Weak	No
D9	Cool	Normal	Weak	Yes
D11	Mild	Normal	Strong	Yes

$$Gain(S_{sunny}, Temp) = 0.570$$

$$Gain(S_{sunny}, Humidity) = 0.97$$

$$Gain(S_{sunny}, Wind) = 0.0192$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
Humidity
{D3,D7,D12,D13}
[4+,0−]
{D4,D5,D6,D10,D14}
[3+,2−]
High
Normal
Yes
?
{D1, D2, D8}
No
{D9, D11}
Yes

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
D10	Mild	Normal	Weak	Yes
D14	Mild	High	Strong	No

Attribute: Temp

$$Values\,(Temp) = Hot, Mild, Cool$$

$S_{Rain} = [3+, 2-]$

$S_{Hot} \leftarrow [0+, 0-]$

$S_{Mild} \leftarrow [2+, 1-]$

$S_{Cool} \leftarrow [1+, 1-]$

$$Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$$

$$Entropy(S_{Hot}) = 0.0$$

$$Entropy(S_{Mild}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$$

$$Entropy(S_{Cool}) = 1.0$$

$$Gain\,(S_{Rain}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Temp)$$

$$= Entropy(S) - \frac{0}{5}Entropy(S_{Hot}) - \frac{3}{5}Entropy(S_{Mild})$$

$$- \frac{2}{5}Entropy(S_{Cool})$$

$$Gain(S_{Rain}, Temp) = 0.97 - \frac{0}{5}0.0 - \frac{3}{5}\,0.918 - \frac{2}{5}\,1.0 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

Attribute: Humidity

$$Values\,(Humidity) = High, Normal$$

$S_{Rain} = [3+, 2-]$ $\quad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$

$S_{High} \leftarrow [1+, 1-]$ $\quad Entropy(S_{High}) = 1.0$

$S_{Normal} \leftarrow [2+, 1-]$ $\quad Entropy(S_{Normal}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$

$$Gain\,(S_{Rain}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Humidity) = Entropy(S) - \frac{2}{5}Entropy(S_{High}) - \frac{3}{5}Entropy(S_{Normal})$$

$$Gain(S_{Rain}, Humidity) = 0.97 - \frac{2}{5}\,1.0 - \frac{3}{5}0.918 = 0.0192$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

Attribute: Wind

$$Values\ (wind) = Strong, Weak$$

$S_{Rain} = [3+, 2-]$

$S_{Strong} \leftarrow [0+, 2-]$

$S_{Weak} \leftarrow [3+, 0-]$

$$Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$$

$$Entropy(S_{Strong}) = 0.0$$

$$Entropy(S_{weak}) = 0.0$$

$$Gain\ (S_{Rain}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

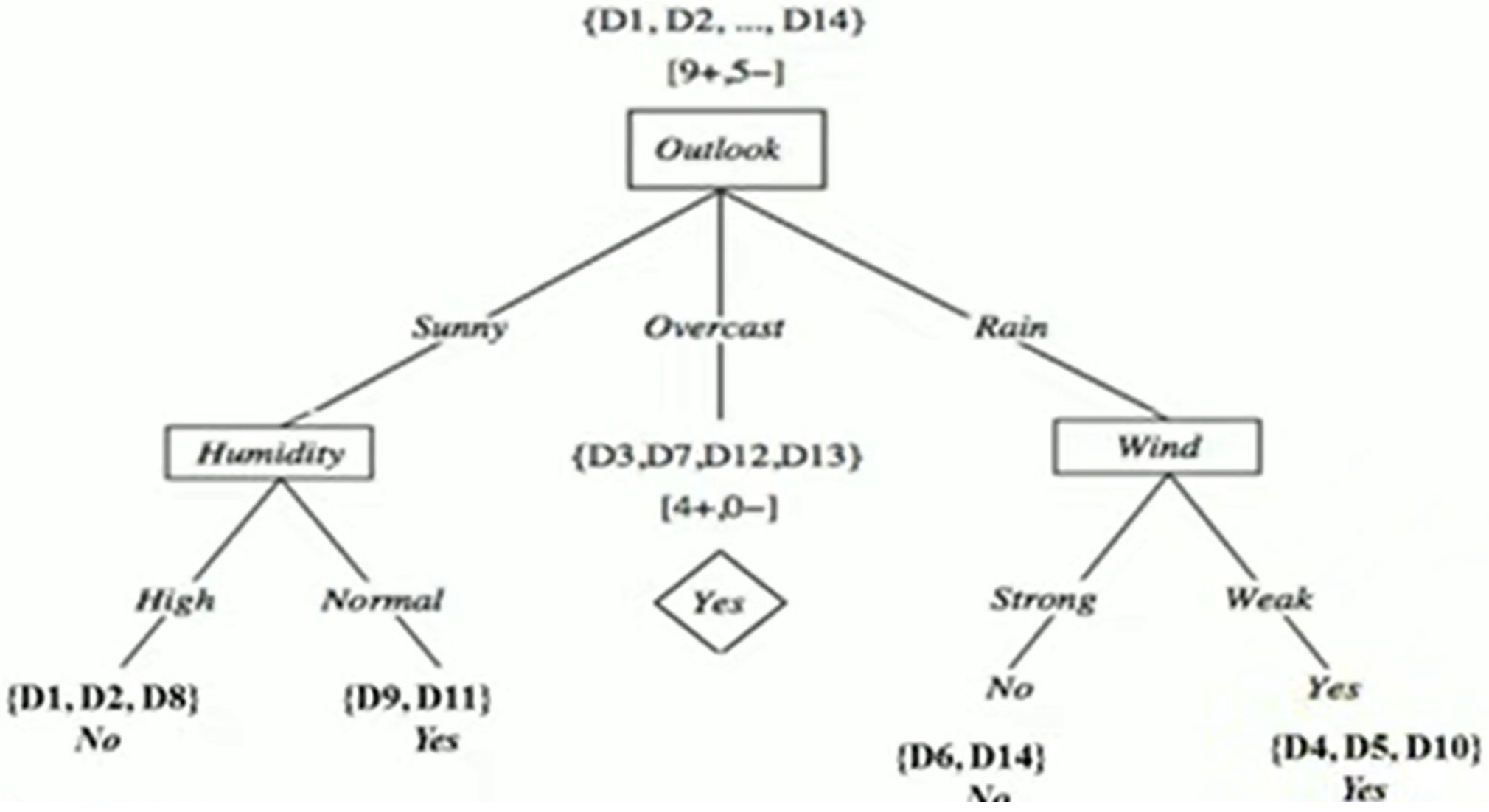
$$Gain(S_{Rain}, Wind) = 0.97 - \frac{2}{5}\ 0.0 - \frac{3}{5}0.0 = 0.97$$

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	Weak	Yes
D6	Cool	Normal	Strong	No
Dl0	Mild	Normal	Weak	Yes
Dl4	Mild	High	Strong	No

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

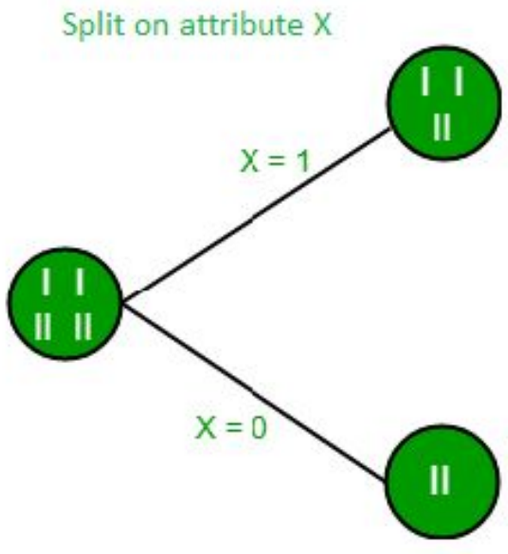
$$Gain(S_{Rain}, Wind) = 0.97$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
Humidity
{D3,D7,D12,D13}
[4+,0−]
Wind
High
Normal
Yes
Strong
Weak
No
Yes
{D1, D2, D8}
No
{D9, D11}
Yes
{D6, D14}
No
{D4, D5, D10}
Yes

- So, decision tree algorithms transform the raw data into rule based mechanism.
- In this post, we have mentioned one of the most common decision tree algorithm named as ID3.
- They can use nominal attributes whereas most of common machine learning algorithms cannot.
- However, it is required to transform numeric attributes to nominal in ID3.
- Besides, its evolved version exists which can handle nominal data.
- Even though decision tree algorithms are powerful, they have long training time.
- On the other hand, they tend to fall over-fitting.
- Besides, they have evolved versions named random forests which tend not to fall over-fitting issue and have shorter training times.

Example: Now, let us draw a Decision Tree for the following data using Information gain. **Training set: 3 features and 2 classes**

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II



$E_{parent} = 1$

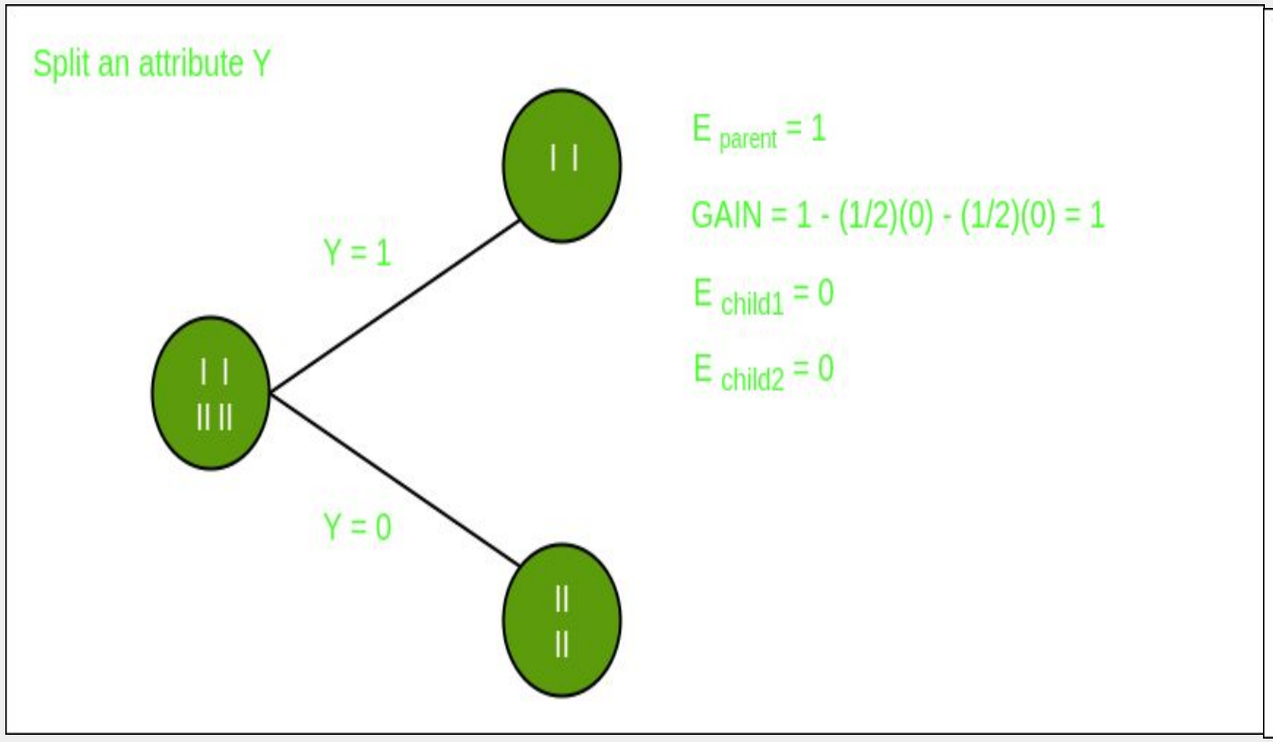
$GAIN = 1 - (3/4)(0.9184) - (1/4)(0) = 0.3112$

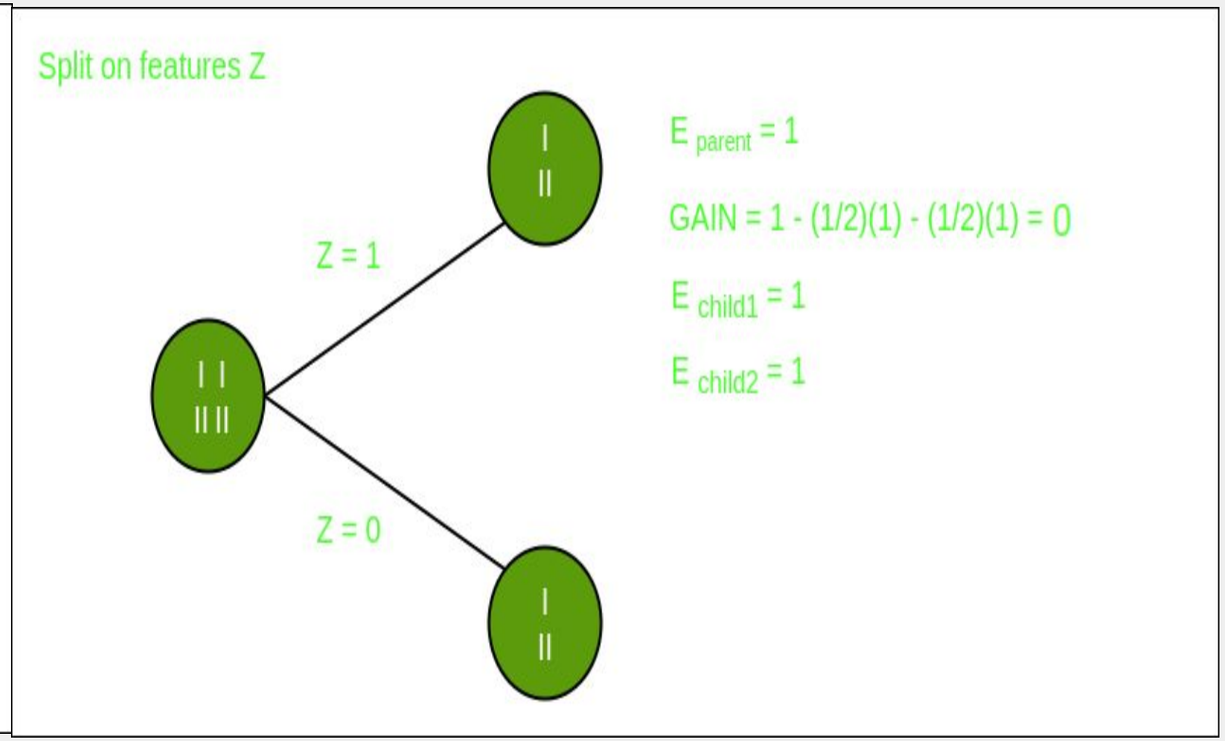
$E_{child1} = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3)$

$= 0.5284 + 0.39$

$= 0.9184$

$E_{child2} = 0$





From the above images, we can see that the information gain is maximum when we make a split on feature Y. So, for the root node best-suited feature is feature Y. Now we can see that while splitting the dataset by feature Y, the child contains a pure subset of the target variable. So we don't need to further split the dataset. The final tree for the above dataset would look like this:

Gini Index in Action

- Gini Index, also known as Gini impurity, **calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.**
- ***Gini index varies between values 0 and 1****, where 0 expresses the purity of classification, i.e. All the elements belong to a specified class or only one class exists there. And 1 indicates the random distribution of elements across various classes. The value of 0.5 of the Gini Index shows an equal distribution of elements over some classes.*
- While designing the decision tree, the features possessing the least value of the Gini Index would get preferred.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

Gini Index: is a metric to measure how often a randomly chosen element would be incorrectly identified.

- It means an attribute with a lower Gini index should be preferred.
- Sklearn supports “Gini” criteria for Gini Index and by default, it takes “gini” value.
- The Formula for the calculation of the Gini Index is given below.

The Gini Index is a measure of the inequality or impurity of a distribution, commonly used in decision trees and other machine learning algorithms. It ranges from 0 to 1, where 0 represents perfect equality (all values are the same) and 1 represents perfect inequality (all values are different).

Some additional features and characteristics of the Gini Index are:

- It is calculated by summing the squared probabilities of each outcome in a distribution and subtracting the result from 1.
- A lower Gini Index indicates a more homogeneous or pure distribution, while a higher Gini Index indicates a more heterogeneous or impure distribution.
- In decision trees, the Gini Index is used to evaluate the quality of a split by measuring the difference between the impurity of the parent node and the weighted impurity of the child nodes.

- ***Classification and Regression Tree (CART) algorithm deploys the method of the Gini Index to originate binary splits.***
- In addition, ***decision tree algorithms exploit Information Gain to divide a node and Gini Index or Entropy is the passageway to weigh the Information Gain.***

Gini Index vs Information Gain

1. The Gini Index **facilitates the bigger distributions** so easy to implement whereas the Information Gain **favors lesser distributions** having small count with multiple specific values.
2. The method of the Gini Index is **used by CART algorithms**, in contrast to it, Information Gain is **used in ID3, C4.5 algorithms**.
3. Gini index **operates on the categorical target variables** in terms of "success" or "failure" and **performs only binary split**, in opposite to that Information Gain **computes the difference between entropy before and after the split** and indicates the impurity in classes of elements.