G51FAI
Fundamentals of AI

Instructor: Siang Yew Chong

Probabilistic Reasoning

Outline

- Probability Theory
  - overview
  - disjoint, independence
- Bayesian Theorem
  - from the joint distribution
  - using independence/factoring
  - from sources of evidence
- Bayes' Classifier

Basic Concepts

- A probability space (S,F,P) is a model for some real-world situations. (S,F,P) is defined by:
  - ***space S*** – the set of all possible outcomes
  - ***σ-algrebra F*** – the set of all events A (subset of S), e.g., power set 2^S
  - ***probabilities P(A)*** assigned to events A, each event is a subset of F containing zero or more outcomes
  - ***Example***: Experiment involving a single coin-flip with head (H) and tail (T). $S = \{H,T\}$, $F = \{H,T\} \subset 2^S = \{\{\},H,T,\{H,T\}\}$, $P(H) = P(T) = 0.5$ (assuming fair coin).

- The first two basic rules of probability are:
  - ***Rule 1:*** *Any probability P(A) is a number between 0 and 1 ($0 \leq P(A) \leq 1$)*
  - ***Rule 2:*** *The probability of the sample space S is equal to 1 ($P(S) = 1$)*

- Two events are ***disjoint*** if they have no outcome in common:
  - ***Rule 3:*** *If two events A and B are disjoint, then the probability of either event is the sum of the probabilities of the two events: $P(A \vee B) = P(A) + P(B)$.*
  - *The chance of any (one or more) of two or more events occurring is called the **union** of the events.*
  - *A single fair coin-flip's outcome is either H or T, so $P(H \vee T) = P(H) + P(T) = 1$.*

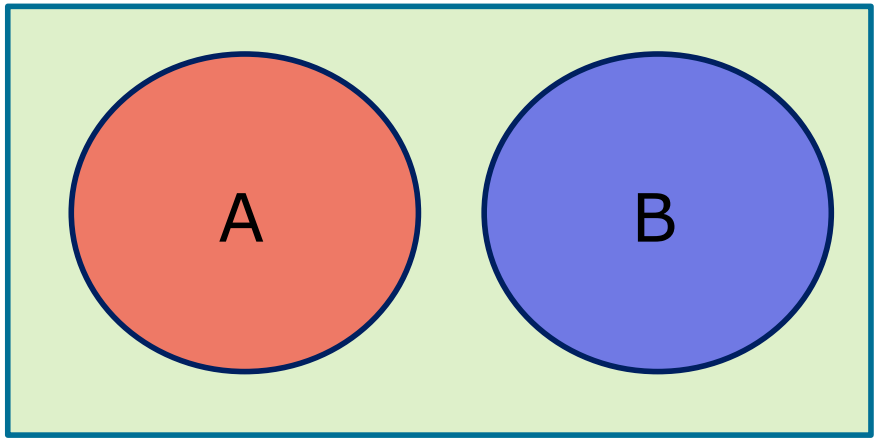
Basic Concepts

- The *complement* A^c of event A refers to all other events.
  - ***Rule 4:*** *The probability that any event A does not occur is* $P(A^c) = 1 - P(A)$.
  - Consider a bag of 3 red and 2 blue marbles. The probability of drawing any red marble is 3/5 = 0.6. The sample space is S = (r, b), which has total probability of 1, so the probability of not drawing red, i.e., its complement is the same as the probability of drawing a blue marble, 2/5 = 0.4.

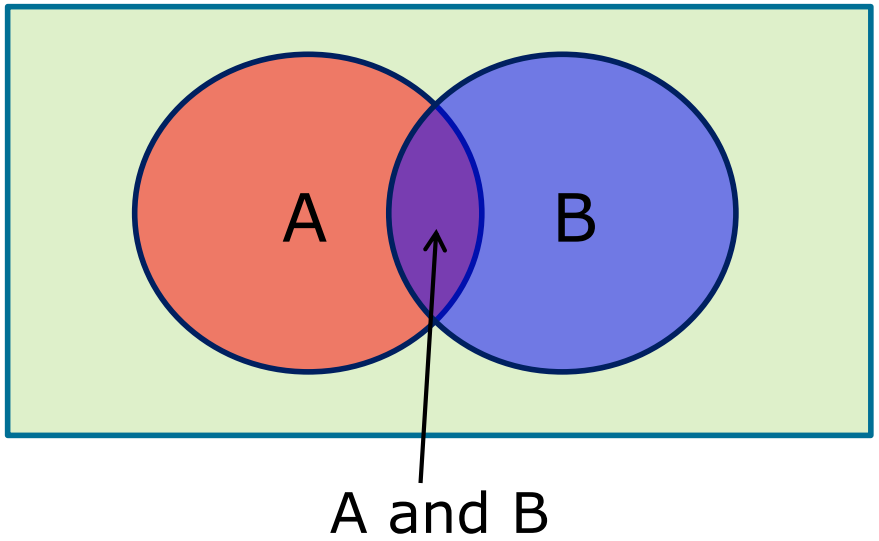
- Consider 2 events occurring in succession, such as two coin-flips. If they are ***independent*** (outcome of the first event has no effect on the second one), then the *multiplication rule states:*
  - ***Rule 5:*** *If two events A and B are independent, then the probability of both events happening is the product of the probabilities for each event:* $P(A \wedge B) = P(A)P(B)$.
  - The chance of *all* of two or more events occurring is called the ***intersection*** of events.
  - For independent events, the probability of the intersection of two or more events is the product of the probabilities, $P(H \wedge H) = P(H)P(H) = 0.5*0.5$.

Venn Diagrams

- Venn diagram is a graphical tool to study the complements, intersections, and unions of events within a sample space *S*
  - top diagram -> events A and B are *disjoint*. Events A and B cannot both occur, so there is no overlapping.



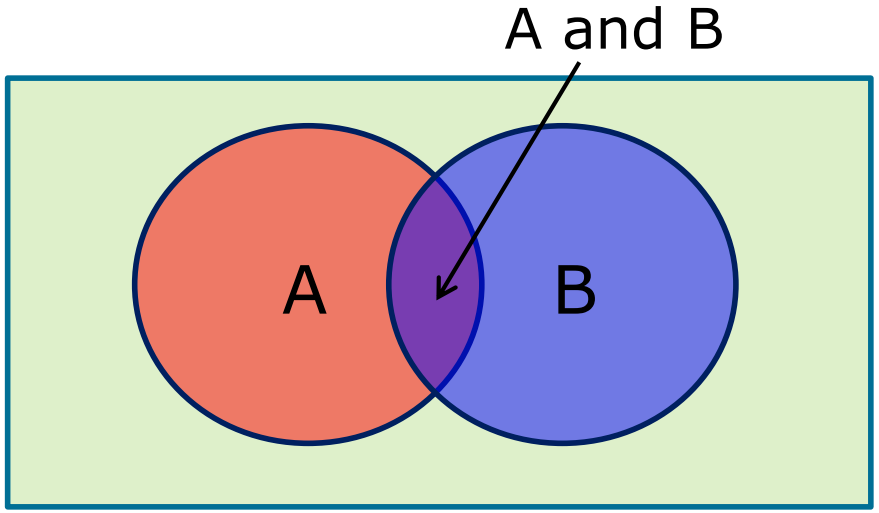
  - bottom diagram -> events A and B are not disjoint. The probability for the union of both events $P(A \wedge B)$ is given by the area of the *intersection*.



Venn Diagrams

- If two events *A* and *B* are *not* disjoint, then the probability of their union (the event that *A* or *B* occurs) is equal to the sum of their probabilities *minus* the sum of their intersection

  Rule 6: $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

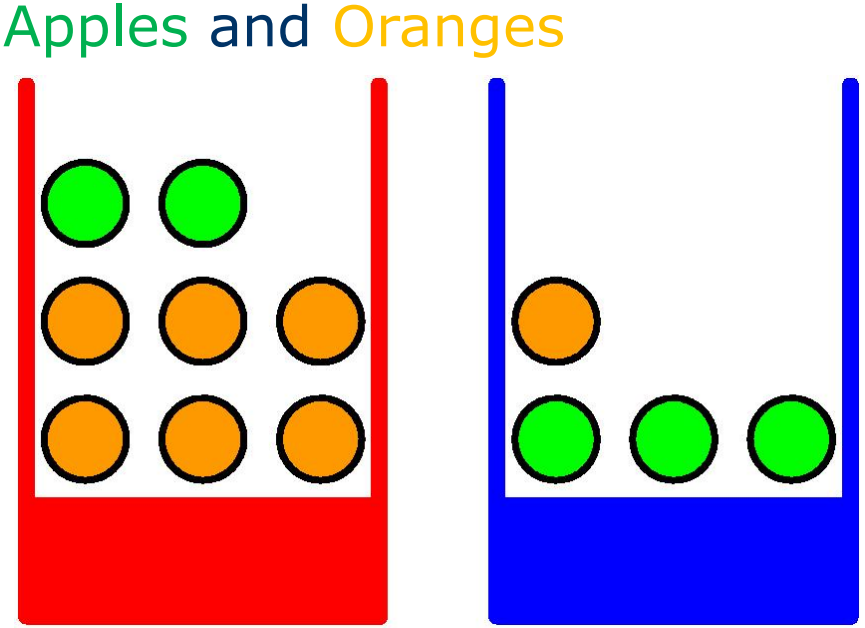


- **Example**: Consider the tossing of 2 coins. The probability of a head on any toss is 1/2. Since the tosses are independent, the probability of a head on both tosses (the intersection) is 1/2*1/2 = 1/4. The probability of a head on either toss (the union) is the sum of the probabilities of a head on each toss minus the probability of the intersection, 1/2 + 1/2 - 1/4 = 3/4

Probability Theory

- One of the boxes is randomly picked and an item of fruit selected from that box. The fruit is replaced back in the same box after observation ***(sampling with replacement)***
- Let B be the random variable denoting the identity of the box and F the random variable indicating the identity of fruit drawn
- If P(B=r) = 4/10 and P(B=b) = 6/10

1. What is the overall probability that this selection procedure will pick an apple?
2. Given that we have chosen an orange, what is the probability that the box chosen is the blue one?



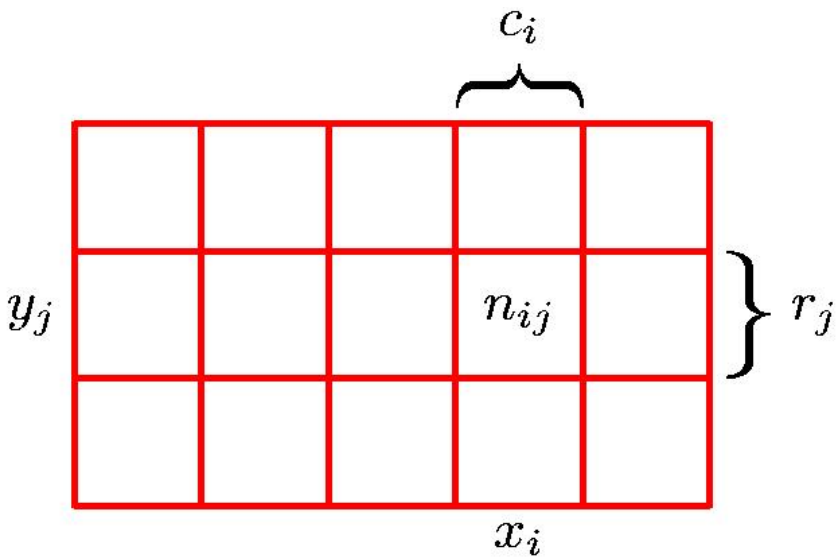
Probability Theory

Suppose X can take any of the values x_i where i = 1,…., M and Y can take the values y_j where j = 1,…., L. After a total of N trials in which both variables X and Y are sampled.

If the number of trials in which X = x_i and Y = y_j is n_{ij} then:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- Joint Probability



Let the number of trials in which X takes the value x_i, irrespective of the value that Y takes, be denote by c_i and similarly the number of trials in which Y takes the value y_j be denoted by r_j.

The probability that X takes the value x_i irrespective of the value of Y is:

$$p(X = x_i) = \frac{c_i}{N}.$$

- Marginal Probability (prior)

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

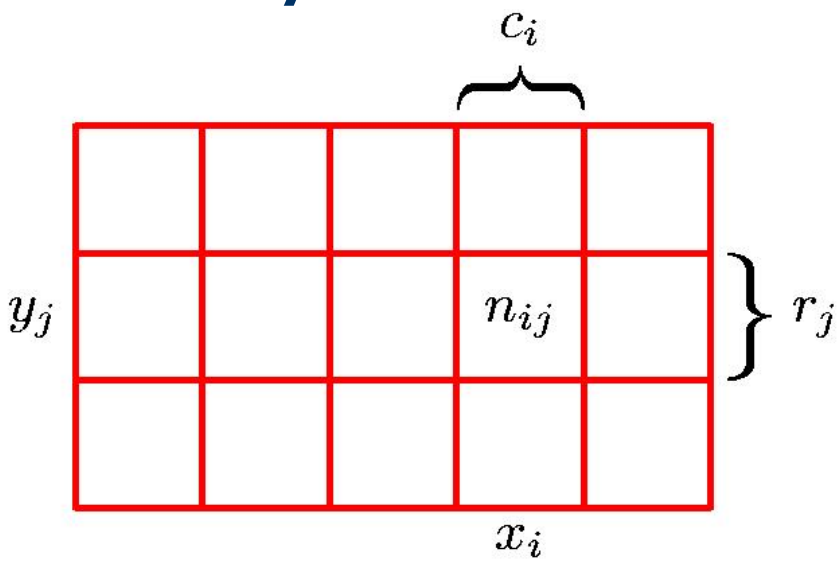
Pattern Recognition and Machine Learning by Christopher Bishop

Probability Theory

If we consider only those instances for which X = x_i then the fraction of such instances for which Y = y_j is:

$$p(Y = y_j|X = x_i) = \frac{n_{ij}}{c_i}$$

- **Conditional Probability (posterior)**



The Rules of Probability:

Sum Rule:

$$p(X) = \sum_Y p(X, Y)$$

Product Rule:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$
$$= p(Y = y_j|X = x_i)p(X = x_i)$$

$$p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem

Product Rule:

$$p(X,Y) = p(Y|X)p(X)$$

Together with the **symmetry property *P*(*X*,*Y*) = *P*(*Y*,*X*)**, the relationship between conditional probabilities can be derived, known as Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

posterior $\propto$ likelihood × prior

Using the sum rule, the denominator can be expressed in terms of the quantities appearing in the numerator

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

Exercise

- Given:
  - P(r) = 4/10; P(b) = 6/10
  - P(a|r) = 2/8; R(o/r) = 6/8
  - P(a|b) = 3/4; P(o/b) = 1/4
- Joint probability
  - P(a, b) = P(a|b) * P(b) = 3/4 * 6/10 = 9/20
  - P(a, r) = P(a|r) * P(r) = 2/8 * 4/10 = 2/20
  - P(o, b) = P(o|b) * P(b) = 1/4 * 6/10 = 3/20
  - P(o, r) = P(o|r) * P(r) = 6/8 * 4/10 = 6/20

Note: P(a, b) + P(a, r) + P(o, b) + P(o, r) = 1

P(a|b) obeys the same rules as probabilities,
i.e., P(a | b) + P(NOT(a) | b) = 1
e.g. P(a|r) + P(o|r) = 1
and P(a|b) + P(o|b) = 1

Joint Probability Table

	Box = b	Box = r
Fruit = a	9/20	2/20
Fruit = o	3/20	6/20

- Prior probability

$$p(X) = \sum_{Y} p(X, Y)$$

  - P(a) = P(a, b) + P(a, r)
    = 11/20
- Conditional probability
  - P(b|o) = P(o|b) * P(b)/P(o)
    = (1/4 * 6/10)/(9/20) = 1/3
  - P(r|o) = 1 – 1/3 = 2/3

Random Variables

- A random variable is the basic element of probability, representing an event with some degree of uncertainty as to the event's outcome
- Let's start with the simplest type of random variables – Boolean ones, which can take on values *true* or *false,* indicating that the event as occurring or not occurring
- Examples (Let *A* be a Boolean random variable):

  A = getting heads on a coin flip

  A = it will rain tomorrow

  A = the US president in 2016 will be female

  A = Mary has influenza

The Joint Probability Distribution

- Joint probabilities can be between any number of variables

  e.g. *P(A = true, B = true, C = true)*
- For each combination of variables, we need to say how probable that combination is
- The probabilities of these combinations need to sum to 1

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Sums to 1

The Joint Probability Distribution

- Once you have the joint probability distribution, you can calculate any probability involving *A*, *B*, and *C*
- Note: May need to use marginalisation and Bayes' rule

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Examples of things you can compute:

- *P(A=true)* = sum of *P(A,B,C)* in rows with *A=true*
- *P(A=true, B=true | C=true) =*

  P(A=true, B=true, C=true) / P(C = true)

Weng-Keen Wong, Oregon State University ©2005

Example

- **Prior probability**: degree of belief without any other evidence

- **Joint probability**: matrix of combined probabilities of a set of variables

- alarm, burglary
  - Boolean, discrete, continuous
- (alarm=true $\wedge$ burglary=false) or equivalently

  (alarm $\wedge$ ¬burglary)

- P(burglary) = 0.1
- P(alarm, burglary) =

	alarm	¬alarm
burglary	0.09	0.01
¬burglary	0.1	0.8

Example

- **Conditional probability**: probability of effect given causes
- **Computing conditional problems**:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: normalizing constant
- **Product rule**:
  - $P(a \wedge b) = P(a \mid b)\, P(b)$
- **Marginalising**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$ (**conditioning**)

- $P(\text{burglary} \mid \text{alarm}) = 0.47$
  $P(\text{alarm} \mid \text{burglary}) = 0.9$
- $P(\text{burglary} \mid \text{alarm}) =$
  $P(\text{burglary} \wedge \text{alarm}) / P(\text{alarm})$
  $= 0.09 / 0.19 = 0.47$
- $P(\text{burglary} \wedge \text{alarm}) =$
  $P(\text{burglary} \mid \text{alarm})\, P(\text{alarm}) =$
  $0.47 * 0.19 = 0.09$
- $P(\text{alarm}) =$
  $P(\text{alarm} \wedge \text{burglary}) +$
  $P(\text{alarm} \wedge \neg\text{burglary}) =$
  $0.09 + 0.1 = 0.19$

Problem with Joint Probability

- Lots of entries in the table to fill up!
- For *k* Boolean random variables, you need a table of size 2^k
- How do we use fewer numbers? Need the concept of independence

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Weng-Keen Wong, Oregon State University ©2005

Independence

Variables *A* and *B* are independent if any of the following hold:

- $P(A,B) = P(A)\,P(B)$
- $P(A \mid B) = P(A)$
- $P(B \mid A) = P(B)$

This says that knowing the outcome of *A* does not tell me anything new about the outcome of *B*.

How is independence useful?

- Suppose you have n coin flips and you want to calculate the joint distribution $P(C_1, \ldots, C_n)$
- If the coin flips are not independent, you need 2^n values in the table
- If the coin flips are independent, then

$$P(C_1,\ldots,C_n) = \prod_{i=1}^{n} P(C_i)$$

Each $P(C_i)$ table has 2 entries and there are *n* of them for a total of 2*n* values

Weng-Keen Wong, Oregon State University ©2005

Conditional Independence

Variables *A* and *B* are conditionally independent given *C* if any of the following hold:

- $P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$
- $P(A \mid B, C) = P(A \mid C)$
- $P(B \mid A, C) = P(B \mid C)$

Knowing *C* tells me everything about *B*. I don't gain anything by knowing *A* (either because *A* doesn't influence *B* or because knowing *C* provides all the information knowing *A* would give)

Weng-Keen Wong, Oregon State University ©2005

Bayes' Rule

- Bayes's rule is derived from the product rule:
  - P(Y | X) = P(X | Y) P(Y) / P(X)
  - commonly expressed as P(H/E) = P(E/H) P(H) / P(E) where

    P(H/E) is the probability that hypothesis H is true given evidence E;

    P(E/H) is the probability that we will observe E given hypothesis H;

    P(H) is the a priori probability that the hypothesis H is true in the absence of any specific evidence.
- Often useful for diagnosis
- In general form:

- A priori probability
- Conditional probability
- Posteriori probability

$$P(H|E) = \frac{P(E|H) \bullet P(H)}{P(E|H) \bullet P(H) + P(E|\sim H) \bullet P(\sim H)}$$

Bayes' Rule - Example

- A child has rash
- A doctor knows
  - 10% of the sick children have flu, and 3% of children has rash
  - Also, 5% of children with flu develop a rash
  - Has the child got a flu?

- Diagnostic hypothesis:
  - P(H) is priori probability that a sick child has flu
  - P(E) is the probability that a child has rash
  - P(E|H) is the probability that a child with flu develops a rash
  - P(H|E) is the probability that the child has flu given evidence of rash

  - P(H) = 0.1; P(E) = 0.03; P(E|H) = 0.05
  - P(H|E) = 0.05*0.1/0.03 = 0.17

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Bayes' Rule – General Form

- Engineers' past experience
  - 5% of items from M1 is faulty; 3.5% of items from M2 is faulty and 2.5% of items from M3 is faulty
  - On a given day, M1 has produced 15% of the total output, M2 has produced 30% and M3 the remainder

Let H_n be the hypothesis that M_n produced the item,
E the evidence that the item is faulty

A priori	Evidence if Hypothesis True
P(H1) = 0.15	P(E | H1) = 0.05
P(H2) = 0.30	P(E | H2) = 0.035
P(H3) = 0.55	P(E | H3) = 0.025

$$P(H_i|E) = \frac{P(E|H_i) \cdot P(H_i)}{\sum_{n=1}^{k} P(E|H_n) \cdot P(H_n)}$$

$$\sum_{n=1}^{k} P(E|H_n) \cdot P(H_n) = 0.0075 + 0.0105 + 0.01375 = 0.0317$$

$$P(H_1|E) = \frac{P(E|H_1) \cdot P(H_1)}{\sum_{n=1}^{k} P(E|H_n) \cdot P(H_n)} = \frac{0.0075}{0.03175} = 0.2362$$

$$P(H_2|E) = 0.0105 / 0.03175 = 0.3307 \qquad P(H_3|E) = 0.01375 / 0.03175 = 0.43307$$

Choosing Hypotheses

- Generally want the most probable hypothesis given the training data
- **Maximum a posteriori** hypothesis h_{MAP}:

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

$$= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D|h)P(h)$$

P(h) = prior probability of hypothesis h
P(D) = prior probability of training data *D*
P(h|D) = probability of *h* given *D*
P(D|h) = probability of *D* given *h*

Example

- *A patient takes a lab test and the result comes back positive. The test returns a correct positive result (+) in only 98% of the cases in which the disease is actually present, and a correct negative result (-) in only 97% of the cases in which the disease is not present*

- *Furthermore, 0.008 of the entire population have this cancer*

Does the patient have cancer or not?

Solution

- If a positive result (+) is returned...

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$
$$P(+|cancer) = 0.98 \qquad P(-|cancer) = 0.02$$
$$P(+|\neg cancer) = 0.03 \qquad P(-|\neg cancer) = 0.97$$
$$P(+|cancer) \cdot P(cancer) = 0.98 \cdot 0.008 = 0.0078$$
$$P(+|\neg cancer) \cdot P(\neg cancer) = 0.03 \cdot 0.992 = 0.0298$$
$$h_{MAP} = \neg cancer$$

- Normalisation

$$\frac{0.0078}{0.0078 + 0.0298} = 0.20745 \qquad \frac{0.0298}{0.0078 + 0.0298} = 0.79255$$

Naïve Bayes' Classifier

- To identify the best classification, the posterior probability of each possible classification is calculated as $P(c_i \mid d_1, \ldots, d_n)$ where c_i is the ith class.
- Bayes' theorem is used to rewrite it as: $P(d_1, \ldots, d_n|c_i) * P(c_i) / P(d_1, \ldots, d_n)$
- The denominator is a common constant independent of c_i, hence can be ignored.

- Using the training data on the right, if new data (x=2, y=3, z=4) is presented, the posterior probability of each class are:
  - P(A) * P(x=2 | A) * P(y=3 | A) * P(z=4 | A)
  - P(B) * P(x=2 | B) * P(y=3 | B) * P(z=4 | B)
  - P(C) * P(x=2 | C) * P(y=3 | C) * P(z=4 | C)

- For class A, B, C we have respectively
  - 8/15 * 5/8 * 2/8 * 4/8 = 0.0417 <--- √
  - 4/15 * 1/4 * 1/4 * 2/4 = 0.0083
  - 3/15 * 1/3 * 2/3 * 1/3 = 0.015

x	y	z	Classification
2	3	2	A
4	1	4	B
1	3	2	A
2	4	3	A
4	2	4	B
2	1	3	C
1	2	4	A
2	3	3	B
2	2	4	A
3	3	3	C
3	2	1	A
1	2	1	B
2	1	4	A
4	3	4	C
2	2	4	A

Summary

- Probability Theory
  - overview
  - disjoint, independence
  - sum and product rules
- Bayesian Theorem
  - from the joint distribution
  - using independence/factoring
  - simple and general form
- Bayes' Classifier
  - Naïve's Bayes' classifier

Acknowledgements

Most of the lecture slides are adapted from the same module taught in Nottingham UK
by
Dr. Rong Qu
and
other slides which are credited individually