**Set of problems to work**

1. Consider the following set of training examples.

a) What is the entropy of this collection of training example with respect to the target

function classification?

b) What is the information gain of A2 relative to these training examples?

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | Classification | A1 | A2 |
| 1 | + | T | T |
| 2 | + | T | T |
| 3 | - | T | F |
| 4 | + | F | F |
| 5 | - | F | T |
| 6 | - | F | T |

1. A property owner is faced with a choice of:

(a) A large-scale investment (A) to improve her flats. This could produce a substantial pay-off in terms of increased revenue net of costs but will require an investment of £1,400,000. After extensive market research it is considered that there is a 40% chance that a pay-off of £2,500,000 will be obtained, but there is a 60% chance that it will be only £800,000.

(b) A smaller scale project (B) to re-decorate her premises. At £500,000 this is less costly but will produce a lower pay-off. Research data suggests a 30% chance of a gain of £1,000,000 but a 70% chance of it being only £500,000.

(c) Continuing the present operation without change (C). It will cost nothing, but neither will it produce any pay-off. Clients will be unhappy and it will become harder and harder to rent the flats out when they become free.

How will a decision tree help the taking of the decision? Analyze the advantages and disadvantages of using decision tree

1. In an airport security checking system, the passengers are checked to find out any intruder. Let I with i ϵ {0,1} be the random variable which indicates whether somebody is an intruder (i=1) or not (i=0) and A with a ϵ {0,1} be the variable indicating alarm. An alarm will be raised if an intruder is identified with probability P(A=1| I = 1) =0.98 and a non-intruder with probability P(A=1| I = 0) = 0.001, which implies the error factor. In the population of passengers, the probability of someone is intruder is P(I = 1) =0.00001. What is the probability that an alarm is raised when a person actually is an intruder?
2. An antibiotic resistance test (random variable T) has 1% false positives(i.e. 1% of those not resistance to an antibiotic show positive result in the test) and 5% false negatives (i.e. 5% of those actually resistant to an antibiotic test negative). Let us assume that 2% of those tested are resistant to antibiotics. Determine the probability that somebody who tests positive is actually resistant (random variable D).
3. For preparation of the exam, a student knows that one question is to be solved in the exam which is either of types A,B or C. The probabilities of A, B or C appearing in the the exame are 30%, 20% and 50% respectively. During the preparation, the student solved 9 of 10 problems of type A, 2 of 10 problems of type B and 6 of 10 problems of type C
   1. What is the probability that the student will solve the problem of the exam?
   2. Given that the student solved the problem, what is the probability that it was of type A?
4. A CCTV is installed in a bank to monitor the incoming customers and take a photograph. Though there are continuous flows of customers, we create bins of timeframe of 5 min each. In each time frame of 5 min, there may be a customer moving into the bank with 5% probability or there is no customer (again, for simplicity, we assume that either there is 1 customer or none, not the case of multiple customers). If there is a customer, it will be detected by the CCTV with a probability of 99%. If there is no customer, the camera will take a false photography by detecting other thing’s movement with a probability of 10%.
   1. How many customers enter the bank on average per day (10 jpirs)?
   2. How many false photographs (there is a photograph taken even though there is no customer) and how many missed photographs(there is no photograph even though there is a customer) are there on average per day?
   3. If there is photography, what is the probability that there is indeed a customer?
5. Why is cosine similarity a suitable measure in context of text categorization? Two rows in a document-term matrix have values – (2,3,2,0,2,3,3,0,1) and (2,1,0,0,3,2,1,3,1). Find the cosine similarity.
6. Compare the Jaccard index and similarity matching coefficient of two features having values (1,1,0,0,1,0,1,1) and (1,0,0,1,1,0,0,1)