Chapter 8

Decision Trees

> If I had eight hours to chop down a tree, I'd spend six sharpening my axe.
>
> Abraham Lincoln

THIS chapter explores the use of decision trees in the context of decision science. Decision trees are often used to analyze complex business decisions and allow to represent strategic alternatives using a graphical tool[1]. The major contribution related to the use of decision trees is twofold:

- On the one hand, decision trees are useful in identifying all the relevant components of a problem. The basic tool of this method is the use of a tree, where each branch represents a different decision and/or alternative. Thus, while building the tree, the decision maker is forced to analyze a number of alternatives and, in the process, the problem itself and all its ramifications will become clearer.
- On the other hand, with the use of a little bit of math, the decision maker will attempt to compute the *expected value*, or *expected utility*, associated to the different alternatives, thus being able to compare, discriminate, and sort the different course of actions that might arise.

A Decision Tree Example. Dante Alighieri Corporation is being sued by Virgilio. Virgilio can settle out of court and win €40,000, or go to court. If Virgilio goes to court, there is a 30% chance that he will win the case. In he wins, a small and large settlement are equally likely (a small settlement nets €50,000, and a large settlement nets €300,000). If Virgilio is risk neutral, what should he do?

Let us define all the alternatives Virgilio is left with. The first decision he needs to make is whether he wants to go to court (Trial) or he prefers to settle out of court(No Trial). These two decisions are at the root node of a tree, where two branches are created, one per decision. With respect to the settlement, the case is closed with a payoff of €40,000. We can notice that there is no uncertainty connected with this decision and, thus, the expected value of the "No Trial" decision is equal to €40,000.

We now want to explore how attractive the alternative decisions are. If Virgilio decides to go to court, then he can either win (Winning) or lose (Losing) the trial. If he loses, then the payoff is zero, while if he wins, he can now win a large sum (Large) or a small sum (Small) with equal probability.

[1] An interesting presentation of the use of Decision Trees in the decision making process, with special emphasis on the application of Decision Trees to legal cases is provided by D. Philbin, "The One Minute Manager Prepares for Mediation: A Multidisciplinary Approach to Negotiation Preparation."

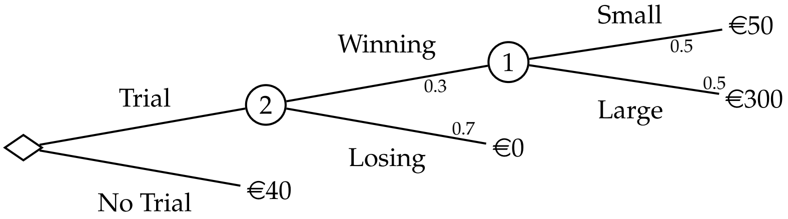


Figure 8.1: A simple decision tree for a lawsuit case.

Figure 8.1 presents a decision tree that summarizes the different alternatives identified in the previous paragraph. Each time a decision is associated to a payoff, we indicate such payoff at the end of the corresponding leave. On the other hand, each time an event of the process is associated with uncertainty, we indicate the associated probability under the branch of the tree.

Considering Figure 8.1, there are a number of elements that need to be described and formalized. However, the two major advantages of using a decision tree can readily be understood after a quick glance. Using decision trees is useful, on the one hand, to identify all the available alternatives and, thus, can be used to grasp a better understanding of the problem itself; on the other hand, the tree allows to quantify the "goodness" of each alternative. For example, given Figure 8.1, we see that the alternative "No Trial" will lead to an expected value of €40,000. During the remainder of this chapter, we will develop techniques aimed at computing the expected value of every branch of the tree. Eventually, our recommendation to the customer will be the action associated with the best (maximum or minimum) expected value.

One important issue connected with the use of Decision Trees is the estimation of probabilities of observing a certain outcome. In the example of Figure 8.1, the decision maker needs to estimate, *e.g.*, the probability of winning and losing a trial. Obviously, such probabilities cannot be more than simple estimates, since there is an intrinsic level of uncertainty associated to decisions depending on other actors, *i.e.*, the judge or the jury. At times, what the decision maker can do is run a preliminary test, or a study, aimed at increasing the chances of "guessing" what the outcome of a certain event will be. For example, in the context of a trial, we could hire jury experts (or we could study similar cases, the profile of the judge, etc.) and thus refine the probabilities of winning and losing the case. Once the result of the test becomes available, we re-compute the probabilities of a certain outcome, *e.g.*, winning, *given* a certain result of the study, *e.g.*, a favorable result of the test. Such revised probabilities are called *conditional* probabilities, since the probability of observing a certain outcome is conditioned to having observed a specific result of the study.

In the next section, we will briefly review the concept of conditional probability. Next, in Section 8.2, we will present an application of Bayes' Theorem to a real-world case. Finally, in Section 8.3, we will show how conditional probability is used in the context of decision trees.

8.1 Conditional Probability

Let us consider the following situations:

- We want to forecast the approval rate of a political party at the next general elections. Since investigating the intentions of every voter is unpractical, we decide to draw a sample out of the population. We then measure the intention of vote of each person in the sample. Once the result on the sample is known, we extrapolate such results over the population. Thus, after running the sample and obtaining a certain result, *e.g.*, positive, the question is: What is the probability that the given party will win the next general elections *given that* the result of the test was positive?

- We want to control the quality of our production process. Let us assume we manufacture electronic chips in batches of a hundred units and we want to determine whether a certain lot is of good quality, *e.g.*, less than ten defective chips, or of bad quality, *e.g.*, with ten or more than ten defective chips. Since inspecting the whole lot is infeasible, we design the following quality control test: We select five chips out of the batch and we thoroughly inspect the chips in the sample. Once the result in know, *i.e.*, number of defective chips in the sample, we want to estimate what is the probability that the lot is of good or bad quality.

- Consider now the example presented in Section 1.2: "Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?" The source of confusion for this problem is related to the difference between conditional as opposed to unconditional probability. While the probability *a priori* of a door being the winning door is 1/3, and thus we should be indifferent between picking any two door, the probability of winning *given that* the host has opened a door should be computed as conditional probability. The answer to this question is related to the probability of winning/losing by switching *given that* the host has open a door, as opposed to the initial, unconditional probability.

All the examples presented above are referring to situations in which a *conditional probability* needs to be computed. Simply stated, given two events A and B, the unconditional probability (or *a priori* probability) of an event is indicated as $p(A)$ or $p(B)$. On the other hand, the conditional probability of A given B is indicated as $p(A|B)$ and is related to the probability of A occurring *given that* B has occurred.

In the remainder of this section, we will illustrate how to compute conditional probabilities using a simple example drawn from the medical field.

The accuracy of a medical test. Consider the following situation (also presented in Section 1.6 with slightly different numbers): Assume that there exists a screening test for a certain type of cancer. Let us also assume that, typically, such test is 99% accurate, *i.e.*, if a person has cancer, the test will give a positive result 99% of the time. Consequently, the test will not give a positive result 1% of the time, despite the fact that a person has cancer (*false negatives*). Let us also assume that, on the other side, if somebody does not have the cancer, the test will give a positive result (a *false positive*) only 5% of the time. Finally, let us assume that 0.5% of the population has this type of cancer.

Now, suppose that a certain person takes the test and the result of such test turns out to be positive. Does this mean that the person is likely to have the cancer?

Before proceeding with the computation of the conditional probability, let us focus on what we want to achieve with such test. A new medical test is designed with the goal of *forecast* the state of an unknown person, *i.e.*, given a new person whose state is not known, we apply the test on such person and, once the result of the test is known, we make an "educated guess" about the state of the person. For example, we might like to conclude that, given that the test gave a positive result, the person has 99% probability of having that cancer.

However, how do we determine the power of accuracy of a new medical test? We might want to apply first the test on a sample of the population, for which the state is known. Once we have a measure of the degree of accuracy of the test on the know sample, we extrapolate this information on the overall population. Let us assume we now select a hundred sick people, *i.e.*, we do know that all these people have cancer. We now apply the test on these people and we find out that the test gave "positive" on 99 people and "negative" on one person, *i.e.*, 1% false negative. We thus conclude that the probability that the test gives positive given that the person is sick is equal to 99%.

Let us now introduce a formal notation. Let us indicate with:

$$S = \{I, NI\}$$

the *states of nature*, *i.e.*, the states we want to forecast, where I stands for infected (with cancer) and NI stands for non-infected. Now, from historical information we can derive the probability of observing each state of nature, *i.e.*, the probability a priory of any given person being sick or healthy. We know that $p(I) = 0.005$ and $p(NI) = 0.995$. This means that, whenever a new person walks into the hospital, we can state that such person has, a priori, a probability of 0.5% of being sick. Thus, we call the probabilities of each state of nature *a priori* probabilities, since they represent our belief that any given person is sick *a priori*, *i.e.*, before any other study or test is run. However, we want to be more precise that that, and that is the reason why we apply a test to the person and, based on the result of the test, we infer about his/her status.

Let us now indicate with:

$$O = \{+, -\}$$

the *observations* of the test, *i.e.*, the test we designed can only give two outcomes, either positive (+) or negative (-). Obviously, we could have designed a test with more than two outcomes and, thus, O would contain more that two possible levels. Now, when we applied the test on a set of a hundred infected people, we obtained 99 positive observations and one negative observation. We can thus express such finding using conditional probability, *i.e.*:

$$p(+|I) = 0.99 \quad \text{and} \quad p(-|I) = 0.01,$$

that is, the probability of having a positive result of the test given that the person is infected is 99%; similarly, the probability of having a negative result of the test given that the person is infected is 1%. These two are called conditional probabilities applied on a known sample or, more precisely, *likelihoods*. Similarly, let us now assume we apply the test on a hundred healthy people, *i.e.*, we select a group of people whose state is already known and we apply the test to all of them. We know that 95 of them will give negative results, while 5 of them will give a positive result, despite the fact that these people are healthy (false negative). We can thus write the two likelihood probabilities as:

$$p(-|NI) = 0.95 \quad \text{and} \quad p(+|NI) = 0.05,$$

Let us now focus on what type of information we have at this point of the analysis. Using historical information, we derive *a priori* probabilities, *i.e.*, we estimate what is the probability that a certain phenomenon is observed. With respect to the current example, we believe that any given person has a 0.5% probability, *a priori*, of developing the disease. Let us thus first collect the *a priori* probabilities $p(I) = 0.005$ and $p(NI) = 0.995$.

Next, we have applied the medical test on known samples, *i.e.*, first a set of a hundred sick people and, then, on a set of a hundred healthy people. We have then measured how good the test is in describing the "past," *i.e.*, how accurate our test is in detecting that a person is sick or healthy *given* that the state of the person is already known. However, such type of 'retroactive' test is of little use.

After all, if we already know that a person is sick, why do we want to apply a test to such person? In reality, what we would like to know is the following: We apply the test on a new person, whose state is unknown, and we get, *e.g.*, a positive result of the test. The question is: What is the probability that the person is sick *given* a positive result of the test? Thus, in the formulas below, we want to derive a method that allows to compute the *a posteriori* probability given the *likelihood* probability, *i.e.*:

$$p(+|I) = 0.99 \quad \Rightarrow \quad p(I|+) = ?$$

Generally speaking, the framework of reference is the following: Using historical information, we will be able to compute the *a priori* probability, *i.e.*, the probability of every state of nature $p(S)$. Next, again using historical information, we derive the likelihoods, *i.e.*, probabilities of O given S. In other words, after selecting a sample of the population whose state is known S, we apply the test and we collect how many times a certain observation O is obtained. Thus, we compute $p(O|S)$. However, what we really want to know is the opposite probability, *i.e.*, the *a posteriori* probability, the probability of S given O. We want to know what is the probability of an individual belonging to a certain state given a result of the test (or *after* the result of the test is known). Thus, we need a method to derive the *a posteriori* probability starting from the *likelihood*:

$$p(O|S) \quad \Rightarrow \quad p(S|O)$$

Bayes' theorem is used to compute the *a posteriori* probability, using the well known formula:

$$p(S_i|O_j) = \frac{p(S_i \text{ and } O_j)}{p(O_j)} \tag{8.1}$$

where:

- S_i is a given state of nature, *e.g.*, $S_i = I$;
- O_j is a given observation, *i.e.*, $O_j = +$;
- $p(S_i)$ is the *a priori* probability;
- $p(O_j)$ is the probability of having observation O_j; and
- $p(S_i \text{ and } O_j)$ is the joint probability of S_i and O_j.

Equation (8.1) can be transformed into:

$$p(S_i|O_j) = \frac{p(S_i) \times p(O_j|S_i)}{\sum_k p(S_k)p(O_j|S_k)} \tag{8.2}$$

where *likelihoods* and *a priori* probabilities (both computed using historical information and, thus, available to the decision maker) are used.

For example, using Equation (8.2), let us indicate how $p(I|+)$ can be computed according to Bayes' theorem. From Equation (8.2), we have:

$$p(I|+) = \frac{p(I) \times p(+|I)}{p(I) \times p(+|I) + p(NI) \times p(+|NI)} = \frac{0.005 \times 0.99}{0.005 \times 0.99 + 0.995 \times 0.05} = 0.09 = 9\% \tag{8.3}$$

Thus, using Bayes' theorem, we discover that our test is quite poor in terms of accuracy: Given that the result of the test is positive, there is only a 9% probability that the person is actually sick. Quite a low predicting power!

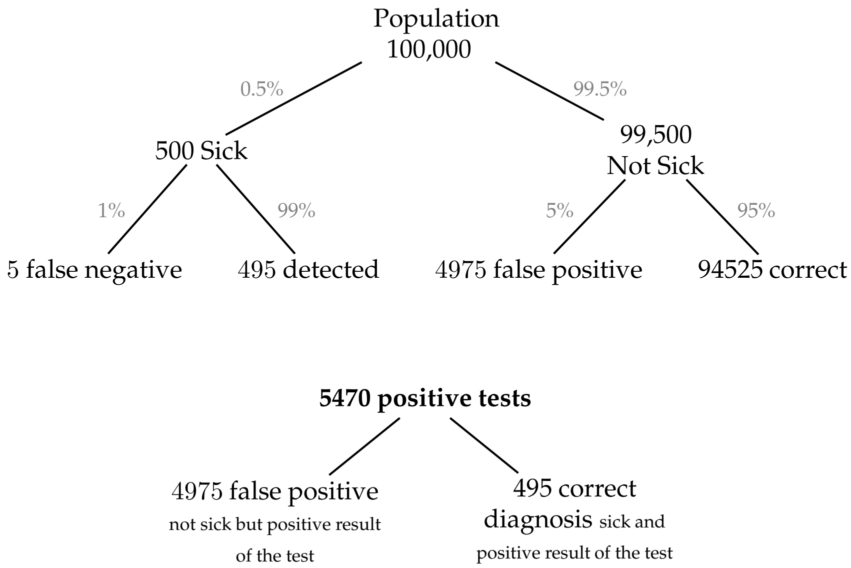


Figure 8.2: How accurate is a medical test?

Why is the medical test so bad, despite the fact that, when used on a known population, it did not seem that bad? In other words, why is it that, when applied on a known population, we only have a 5% probability of obtaining a false positive and a 1% probability of obtaining a false negative, while, when used on an unknown population, the test becomes almost useless? Why is $p(+|I) = 99\%$ so different from $p(I|+) = 9\%$?

To understand why, consider Figure 8.2. Let us assume we have a population of 100,000 people. Given that the level of incidence of the disease is 0.5%, we know that 500 people will be sick while 99,500 people will be healthy. Given the accuracy rate of the test, out of 500 sick people, the test will give a positive result $0.99 \times 500 = 495$ times, while we will get a negative result (despite the fact that these people are sick) 5 times.

Similarly, if we consider the healthy population of $99,500$ individuals, the test will give a negative result $0.95 \times 99,500 = 94,545$ times, while we will get a positive result of the test (despite the fact that these people are healthy) $0.05 \times 99,500 = 4975$ times (false positive).

Consequently, the total number of times that we will observe a positive result of the test is equal to $495 + 4975 = 5470$. Out of these observations, $495/5470 = 9\%$ of the time, the response of the test will be right, while $4975/5470 = 91\%$ of the time the response will be wrong.

Consequently, the probability that a person that tested positive has a cancer is equal to 9%. While the math is quite simple, the findings are a bit counter-intuitive and, therefore, we tend to reject or ignore such results. This is not strange, considering the limitations that many of us display when it comes to dealing with probabilities. In the sequel, you will see that such 'surprising' result is easily obtained using Bayes' theorem. Similarly, you will see how such theorem finds applications in the justice system, *e.g.*, prosecutor's fallacy, DNA match, in quality control and other industrial and business applications.

Again, why is this happening? Why is there such a huge difference between $p(+|I)$ and $p(I|+)$? The answer lies in the *a priori* probabilities. What Bayes' theorem does is weighting the relative importance of each state of nature with respect to the overall population. Let us have a further look at Figure 8.2. We can quickly see that, given the overall population, the overwhelming majority of people will be healthy (99.5% of them). Thus, even a tiny mistake on such population (false positive) will have a high impact on the accuracy of the test. How do we derive the 9% accuracy level? Once more, we had:

$$9\% = \frac{\text{number of correct positive results}}{\text{total number of positive results}} = \frac{495}{495 + 4975}$$

Thus, it is not surprising that the overall accuracy of the test is low: The test has only a 5% false positive level but, since the great majority of people is healthy, this 5% leads to a total of 4975 false positive tests out of a total of 5470 positive tests. That is, we observe 5470 positive results of the test and the overwhelming majority of these tests are mistakes (false positive). Thus, it should not be surprising that the test has such a low level of accuracy!

Let us now consider what happens if we distribute the population in different proportions, *i.e.*, we change the weight of each state of nature. For example, let us assume that, without altering the nature of the test itself, we now apply the test on a different country, for which the *a priori* probability of being sick is equal to $p(I) = 0.9$ and, conversely, the probability of being healthy is equal to $p(NI) = 0.1$. Let us now revise the numbers of Equation (8.2):

$$p(I|+) = \frac{p(I) \times p(+|I)}{p(I) \times p(+|I) + p(NI) \times p(+|NI)} = \frac{0.9 \times 0.99}{0.9 \times 0.99 + 0.1 \times 0.05} = 0.99 = 99\% \tag{8.4}$$

Surprisingly, without changing the test itself, *i.e.*, the test is as good as before, we now have a radically different result: The test has a 99% degree of accuracy when it comes to detecting a certain disease. The reason why we now have a radically different result lies in the distribution of the population. With the new distribution of the population, where the overwhelming majority of people is sick, the number of false positive results is extremely reduced and, therefore, almost all the positive results of the test that we obtain are referring to people that are actually sick. To better grasp this point, you might want to draw a tree similar to the one presented in Figure 8.2 using the new *a priori* probabilities.

Obviously, since it is not possible to increase the level of accuracy of the test without worsening other features of the same test, something else must have changed. Let us now consider other conditional probabilities, *e.g.*, $p(NI|-)$, that is, the probability that a person is healthy given a negative result of the test. Lets us compute this probability in the two cases, (i) when the disease is rare ($p(I) = 0.005$), and (ii) when the disease is widespread ($p(I) = 0.9$):

$$\begin{aligned} \text{case (i): } p(I) = 0.005 \quad \Rightarrow \quad p(NI|-) &= \frac{p(NI) \times p(-|NI)}{p(NI) \times p(-|NI) + p(I) \times p(-|I)} \qquad (8.5) \\ &= \frac{0.995 \times 0.95}{0.995 \times 0.95 + 0.005 \times 0.01} = 0.9999 = 99.99\% \\ \text{case (ii): } p(I) = 0.9 \quad \Rightarrow \quad p(NI|-) &= \frac{p(NI) \times p(-|NI)}{p(NI) \times p(-|NI) + p(I) \times p(-|I)} \qquad (8.6) \\ &= \frac{0.1 \times 0.95}{0.1 \times 0.95 + 0.9 \times 0.01} = 0.51 = 51\% \end{aligned}$$

Therefore, as we can see from Equation (8.5), when the disease is rare, the test is really bad in detecting the disease (9% accuracy) while, on the other hand, it is almost perfect in discarding the disease (99.99% accuracy), *i.e.*, when the test gives a negative results, we can be almost sure that a person does not have the disease. Conversely, when we change the weight of the two populations, healthy *vs* sick, we see that the test becomes good at detecting the disease (99% from Equation 8.4) but, on the other hand, it is a poor test when it comes to discarding the disease (51% accuracy, *i.e.*, if the test is negative, the person has only a 51% probability of being healthy).

What is the conclusion that we can draw? The first finding is that we should always be aware of the difference between the probability *a priori* and the probability *a posteriori*. That means that, while we have a certain believe about the general probability of occurrence of a state of nature ($p(S)$), the probability of occurrence of the same state after observing a certain results of a test should be revised. Considering the example presented in this section, we know that, a priory, each person has a 0.5% probability of being sick. However, the probability of being sick must be revised after we obtain the result of the test.

A second important finding is related to how the formula of Bayes' theorem works. As we have seen from the examples above, Bayes' formula weights in the relative importance of each class. That is, given the same likelihood, the fact that a certain observation is drawn out of a large population, as opposed to a small one, is taken into account by Bayes' theorem. Therefore, even a small error (false positive or false negative) on a very large population can have quite a large impact on the level of accuracy of the test itself. With respect to the example presented in this section, it is wrong to conclude that since there is a 99% probability of getting a positive result of the test out of a sick person, the test is 99% accurate. As we have seen, the level of accuracy depends not only on the percentage of false positives and false negatives but also on the relative weight, or importance, of those errors and is related to the way in which the population is divided.

Let us now conclude this section by illustrating how Bayes' formula can easily be computed using an auxiliary structure, a small tree that allows to define all the relevant terms of Equation (8.1). In order to compute a conditional probability using Bayes' theorem, we can recur to a simple two-level tree. The first level identifies all the possible states of nature, *i.e.*, out of the root node, we create a branch for each state of nature. Considering the medical test example, we only have two possible states of natures, *i.e.*, infected (I) and non-infected (NI). The second level of the tree defines the possible observations. Therefore, with respect to each first-level branch (nodes (i) and (ii) in Figure 8.3), we create one branch per observation, thus identifying all the possible configurations. We thus obtain one leave for each possible pair state of nature-observation. For example, the first leave of Figure 8.2 corresponds to people that are infected (I) and returned a positive result of the test (+). We can thus compute all the joint (AND) probabilities, *i.e.*, the numerator of Equation (8.1). As presented in Figure 8.3, along each path we can obtain the associated joint probabilities by multiplying the probabilities along the path itself. For example, if we want to compute $p(I \text{ AND } +)$, we just need to multiply the probability over the first path, *i.e.*, the probability of the first level (0.005) with the probability of the second level (0.99). We thus obtain:

$$p(I \text{ AND } +) = 0.005 \times 0.99 = 0.00495$$

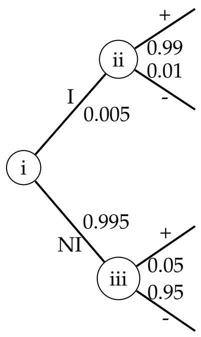
In a similar fashion, we can quickly obtain all the possible joint probabilities, as the product of the probabilities along each path, as indicated in Figure 8.3. Once we have the four joint probabilities, we need to compute the denominator of Equation (8.1). The probability of each observation can be obtain as the sum of the joint probabilities containing that observation. For example, to compute $p(+)$, we simply need to add up all the joint probabilities in which the observation $+$ appears, *i.e.*, (1) and (3). Thus, we have:

$$p(+) = (1) + (3) = 0.00495 + 0.0498 = 0.0547$$

Similarly, we have:

$$p(-) = (2) + (4) = 0.00005 + 0.94525 = 0.9453$$

It is worth noting that $p(+) + p(-) = 1$. In general, the sum of the probabilities of all the possible observations is equal to one.



	p(S AND O)		p(O)
p(I AND +)	$0.005 \times 0.99 = 0.00495$	(1)	p(+) = (1) + (3)
p(I AND -)	$0.005 \times 0.01 = 0.00005$	(2)	
p(NI AND +)	$0.995 \times 0.05 = 0.0498$	(3)	p(-) = (2) + (4)
p(NI AND -)	$0.995 \times 0.95 = 0.9453$	(4)	
	$\sum = 1$		

Figure 8.3: Bayes' formulas computed using an auxiliary tree.

We now have all the ingredients we need to compute Bayes' probabilities using Equation (8.1). We can easily see that:

$$\begin{aligned}
p(I|+) &= \frac{(1)}{(1)+(3)} = \frac{0.00495}{0.0547} = 0.09 \\
p(NI|+) &= \frac{(3)}{(1)+(3)} = \frac{0.04975}{0.0547} = 0.91 \\
p(I|-) &= \frac{(2)}{(2)+(4)} = \frac{0.00005}{0.9453} = 0.00005 \\
p(NI|-) &= \frac{(4)}{(2)+(4)} = \frac{0.94525}{0.9453} = 0.9999
\end{aligned}$$

We can finally observe that the sum of the conditional probabilities for any given observation should be equal to one, *i.e.*, $p(I|+) + p(NI|+) = 1$ and $p(I|-) + p(NI|-) = 1$.

8.2 Bayes' Theorem and the Prosecutor's Fallacy

This section is based on "Beyond Reasonable Doubt," available at the following web page: `http://plus.maths.org/issue21/features/clark/index.html`. Let us analyze the following information about a real case happened in England in 1997 (Statistics are taken from the Confidential Enquiry for Stillbirth and Deaths in Infancy - CESDI). A British family, the Clarks, experienced the death of their 11-week-old baby. The baby died in his sleep and, since there was no forensic evidence supporting any other explanation, the death was assumed to be a case of Sudden Infant Death Syndrome (SIDS), also known as cot dead.

A year later, the Clarks had a second baby. The second child died after 8 weeks. Once more, there was little or no forensic evidence supporting any alternative hypothesis and, eventually, the investigation lead to only two possible alternatives:

- Murder: The mother, Sally Clark, murdered both children, or

- Sudden Infant Death Syndrome (SIDS): The Clark family experienced a double case of cot dead.

At the trial, a famous pediatrician, Sir Roy Meadow, speaking as expert witness of the prosecutor, pointed out that the probability that a child dies of SIDS is 1 in 8500. Thus, he concluded, the probability *a priori*, of observing a cot dead is $p(SIDS) = 1/8500$. Thus, the argument of the pediatrician exploited by the prosecutor is that the probability of having two cot dead in the same family is equal to:

$$p(2SIDS) = \frac{1}{8500} \times \frac{1}{8500} = \frac{1}{73M}$$

The prosecutor concluded that there is 1 chance in 73 millions that a normal family experiences two cot dead and, therefore, that Sally Clark had one chance in 73 million of being innocent. As strange as it may sound, this argument was enough to convict Sally Clark, and Sally Clark was sentenced to jail.

However, there are at least two major flaws in this reasoning, one of them connected with the use of conditional probability. You might be interesting in learning that the proper use of conditional probability by the Royal Society of Statistics was at the basis of the reopening of the case and, eventually, to the acquittance of Sally Clark.

How can you assess the likelihood, in absence of any reliable forensic evidence, that Sally Clark murdered her two little boys?

If you are interested in the case, you are encouraged to research around the topic, starting from the following source:

- `http://plus.maths.org/issue21/features/clark/index.html`: A web site dedicated to the topic. A proposal about how to assess the likelihood of Sally Clark being guilty is offered.

- A You Tube video of Peter Donnelly, whose title is "How juries are fooled by statistics." The talk was presented at the TED conference in 2005 and, around minute 13, he briefly presents the case of Sally Clark.

Let us just report here the conclusion of the case:

- "We do not want to answer the question "Is Sally Clark guilty?". In British trials, there is a presumption of innocence - it is for prosecutions to prove "beyond reasonable doubt" that defendants are guilty, not for defendants to prove their innocence. So all we need to do is to make reasonable estimates and show that these lead to reasonable doubt." (Plus Magazine)

- It is with the very greatest sadness that Sally Clark's family announces that Sally was found dead at her home this morning, having passed away during the night. Sally, aged 42, was released in 2003 having been wrongfully imprisoned for more than 3 years, falsely accused of the murder of her two sons. Sadly, she never fully recovered from the effects of this appalling miscarriage of justice. Sally, a qualified solicitor, was a loving and talented wife, mother, daughter and friend. She will be greatly missed by all who knew her. (Sally Clark - Friday 16th March 2007, `http://www.sallyclark.org.uk`)

8.3 Decision Trees

Decision Trees: An Example. A company produces electronic chips in lots of 10 units. A lot can either be of good quality (when it has 20% of defective chips) or of bad quality (when it has 50% of defective chips.)[2] Historical data highlight that 80% of produced lots is of good quality, while 20% is of bad quality. The lot is provided as input to a second phase, the *processing* phase. However, the cost of the processing phase depends on the quality of the lot provided in input. It is known that when processing a good quality lot, the company incurs a cost of €1000, while processing a bad quality lot generates a cost of €4000.

The company knows that there exists another option, *i.e.*, re-working the entire lot, *before* processing it, at an extra cost of €1000. After the re-working phase, every lot is transformed into a good quality lot, regardless of the initial state of the lot itself. Once the lot is re-worked, such lot is ready for the processing phase.

The company is studying the option of designing a "quality test" with the following format: Once a new lot is received, and *before* the processing phase, two chips are randomly selected from each lot to infer about the status of the whole lot. The cost of such test is €50 per lot.

What is the minimum cost policy for the company?

Figure 8.4 presents a flow chart of the problem. As we can easily see from the picture, the company needs to make a number of subsequent decisions over time. The typical setting in which we might want to use decision trees is when a problem is composed of multiple decisions over time. The decision maker will first make a decision; next, the result of this decision is observed; based on the observation, a second decision is made; a new observation is collected, *etc.* Thus, when we have a problem in which a set of pairs decision-observation is identified, we might want to explore the possibility of using decision trees.

From Figure 8.4, we see that whenever a new lot comes in, the state of the lot (good or bad) is unknown. Once a new lot is received, the decision maker has the option of running a test to discover what is the status of the lot itself. If the decision maker decides not to use the test, then he needs to determine whether it is preferable to rework the lot (no matter what the status of the lot is), or to process the lot (again, regardless of the status of the lot). If the lot is reworked, we pay €1000 for the reworking phase, plus €1000 for the subsequent processing phase (since we know that *every* lot, after the reworking phase, becomes a good quality lot). On the other hand, if we decide to process the lot, we can incur a cost of €1000, if the lot is good, or €4000, if the lot is bad. Such difference in costs could be imputed to the fact that, *e.g.*, a bad quality lot slows down the production, maybe due to the generation of waste, and, thus, increases the production cost per lot.

With respect to the test, as mentioned in the example, rather than examining the whole lot, the decision maker decided to randomly select two chips out of a lot and thoroughly inspect such two chips. The test returns as output the number of defective chips out of the sample (either zero, one, or two). Once this information is known, the decision maker needs to extrapolate the information from the sample to the population, *i.e.*, from the two chips to the entire lot. Based on the result of the test, the decision maker will decide whether to process or to rework the lot.

The first step in working with a decision tree is connected with the identification of *states of nature* and *observations*. We must remember that the states of nature are the states we want to forecast, while the observations are the possible outcomes of a test. With respect to the current example, it is quite simple to identify the following:

[2] These assumptions are obviously an oversimplification of the reality. We are here assuming that every lot will either have 20% of defective chips (and, thus, it is labeled "good"), or 50% of defective chips (and, consequently, it is labeled "bad"). However, a more realistic setting would be to have good lots whenever we have 20% *or less* defective chips, while a bad lot is a lot with a number of defective chips between 20% (excluded) and 50%. However, to keep the computations simple, let us work with the initial, simplified, assumption. This will allow to simplify our computation without altering the logical steps of the exercise. During the development of the exercise, we will try to relax such assumptions.

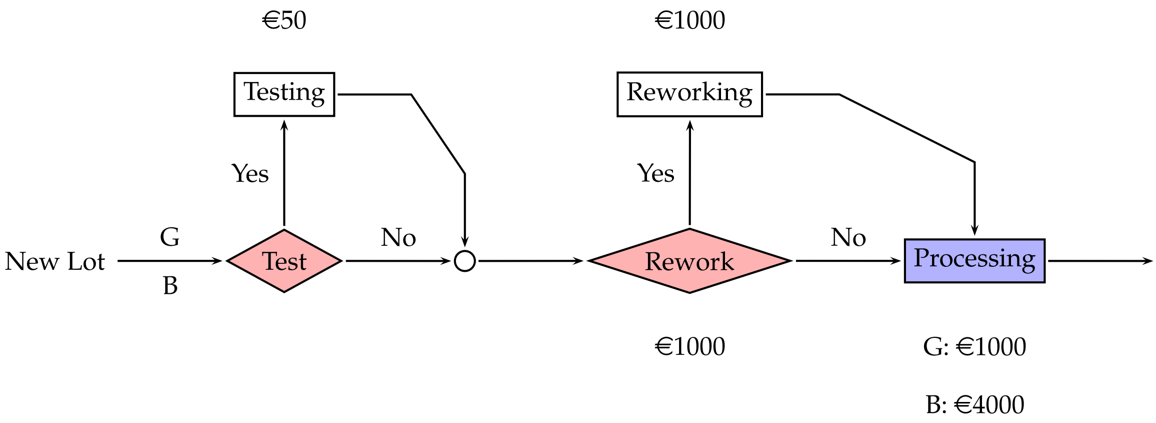


Figure 8.4: A flow chart of the process.

- States of nature: The state of each incoming lot. The company is primarily concerned with "guessing," *i.e.*, identifying, the correct status of each lot. If we knew whether a lot is good or bad, we could easily define the optimal policy: *i.e.*, every good lot is directly processed, while every bad lot is reworked and then processed. Thus, we can indicate with:

$$S = \{G, B\} \tag{8.7}$$

  the states of nature, where G indicates that *a lot* is of good quality, while B indicates that *a lot* is of bad quality.

- Observations: The observations are the results of the test. With respect to the test proposed in this example, we can only obtain three possible outcomes: Once two chips are selected and inspected, either both of them are working (zero defective), one of them is working and one is broken (one defective), or both of them are broken (two defectives). Let us indicate the observations with:

$$O = \{D_o, D_1, D_2\} \tag{8.8}$$

  where D_0 indicates zero defective chips, D_1 indicates one defective chip, and D_2 indicates two defective chips.

The second step, once the states of nature and observations have been identified, is to collect historical information to define the probabilities a priori and the likelihoods, *i.e.*, $p(S_i)$ and $p(O_j|S_i)$. With respect to the current example, we easily derive:

- Probabilities a priori: $p(G) = 0.8$, and $p(B) = 0.2$.

- Likelihoods: Probabilities of the type $p(O|S)$ are not explicitly given in the exercise. Therefore, we need to compute such probabilities based on the available information. For example, let us begin by computing the following three probabilities: $p(D_o|G), p(D_1|G)$, and $P(D_2|G)$. That is, let us assume we take a good lot: We now want to compute the probability of selecting two chips and having both of them working, one working and one broken, or both of them broken, respectively.

  Using Figure 8.5, we can easily see that computing $p(D_0|G)$, *i.e.*, computing the probability of extracting two working chips out of a good lot (lot of size ten), is equivalent to computing the

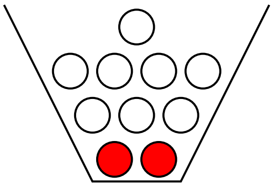


Figure 8.5: A good quality lot and a bag of white and red balls.

probability of extraction of two white balls out of a set of ten, without reinsertion (a bag in which 80% of the balls is white, *i.e.*, working, and 20% is red, *i.e.*, broken – according to the definition of a "good" lot).

Thus, we can easily see that:

$$p(D_0|G) = 8/10 \times 7/9 = 0.622 \tag{8.9}$$

where the first term $(8/10)$ corresponds to the probability of selecting the first white ball, while the second term $(7/9)$ is the probability of extracting the second white ball, *without* reintroducing the first white ball into the bag.

With a similar reasoning, we can assimilate the probability $p(D_2|G)$, *i.e.*, the probability of obtaining two defective chips out of a good lot, to the probability of selecting two red balls out of the bag. Thus, we have:

$$p(D_2|G) = 2/10 \times 1/9 = 0.022 \tag{8.10}$$

where the first term $(2/10)$ corresponds to the probability of selecting the first red ball, while the second term $(1/9)$ is the probability of extracting the second red ball, *without* reintroducing the first red ball into the bag.

Finally, we can now compute $p(D_1|G)$, *i.e.*, the probability of obtaining one defective chip (and one working chip) out of a good lot, to the probability of selecting one red ball and one white ball out of the bag. We now need to consider two possible scenarios: (i) the first ball is red and the second ball is white, or (ii) the first ball is white and the second ball is red. Thus, we have:

$$p(D_1|G) = 2/10 \times 8/9 + 8/10 \times 2/9 = 0.355 \tag{8.11}$$

where the first and second terms are referred to scenario (i) and correspond to the probability of selecting the first red ball $(2/10)$, and the probability of extracting the second ball as white, *without* reintroducing the first red ball into the bag $(8/9)$. Similarly, the third and fourth terms are referred to scenario (ii) and correspond to the probability of selecting the first white ball $(8/10)$, and the probability of extracting the second ball as red, *without* reintroducing the first white ball into the bag $(2/9)$.

In a similar fashion, we can now compute the likelihoods referred to a "bad" lot, *i.e.*, a lot that has exactly 50% of defective chips and 50% of working chips. Figure 8.6 builds the analogy between a bad quality lot and an extraction of balls from a bag.

Thus, we have:

$$p(D_0|B) = 5/10 \times 4/9 = 0.222 \tag{8.12}$$

$$p(D_1|B) = 5/10 \times 5/9 + 5/10 \times 5/9 = 0.555 \tag{8.13}$$

$$p(D_2|B) = 5/10 \times 4/9 = 0.222 \tag{8.14}$$

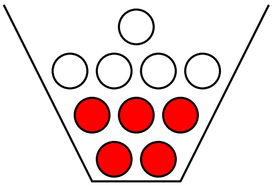


Figure 8.6: A bad quality lot and a bag of white and red balls.

Thus, to summarize, from Equations (8.9)– (8.14) we obtained the following likelihoods:

$$\begin{aligned}
p(D_0|G) &= 0.622\\
p(D_1|G) &= 0.355\\
p(D_2|G) &= 0.022\\
p(D_0|B) &= 0.222\\
p(D_1|B) &= 0.555\\
p(D_2|B) &= 0.222
\end{aligned}$$

It is worth noting that the sum of the probabilities for any given state is equal to one, *i.e.*, $p(D_0|G) + p(D_1|G) + p(D_2|G) = 1$, as well as $p(D_0|B) + p(D_1|B) + p(D_2|B) = 1$. This must always be the case, since the observations are exhaustive, *i.e.*, they are covering the whole spectrum of possibilities.

The third step is focused on drawing the decision tree. The idea is to start from a root node and to identify every possible alternative, or course of action. In the following, we are going to use the following notation:

○ : event node

□ : decision node

As indicated, we are going to use a circle to identify an *event node*, *i.e.*, a point in which an uncontrollable event takes place. We cannot decide about the outcome of such event but, rather, we can estimate the likelihood of a certain outcome with a probability value. Thus, out of every event node, we will have a set of branches, each identifying a possible scenario and each associated to a probability value, *i.e.*, the probability of that event occurring.

On the other hand, we identify decision nodes with a square. A decision node is a point of the problem in which the decision maker will make a decision. For every decision node, we will have a set of branches listing all the possible course of actions. At every decision node, a rational decision maker will select the alternative that maximizes his utility function.

Typically, a decision tree will start with a decision node, *i.e.*, the decision maker will have to identify what are the possible alternatives at hand. In the case of the exercise, the decision maker is called to choose among the three main alternatives, namely: (i) rework every lot, (ii) process every lot, or (iii) perform a test (and decide what to do next, depending on the result of the test). Thus, we could create a root node for the three (a decision node), with the outgoing branches, one per alternative.

However, a typical way of starting the tree is by focusing on the test, thus identifying two main alternatives, one in the case in which we execute the test (and, therefore, postpone the decision), and one in which we decide not to perform the test (and, therefore, decide immediately what to do next). Figure 8.7 presents a possible tree for the example studied in this section.

As we can see from Figure 8.7, and in line with what presented in Figure 8.4, we first analyze whether we should, or should not, execute the test. Let us first have a look at the top part of the tree, following the branch "no test." If we decide not to do the test, we need to decide whether we want to process every lot or to rework every lot. If we decide to rework every lot, we incur a cost of €2000 per lot, without uncertainty. Thus, we know that one available strategy for the company, *i.e.*, rework every lot twice, will generate an average cost of €2000 per lot. On the other hand, if we decide to process every lot, we then have an event node, *i.e.*, we just need to wait and see whether the lot is of good or bad quality. If the lot turns out to be of good quality, the processing phase generates a cost of €1000; however, if the lot turns out to be of bad quality, the cost for the processing phase is €4000.

At this point, we should ask: Assuming that we have decided not to take the test (top branch of the tree), which alternative shall we recommend to the company? The answer is quite simple: On the one hand, we know that the expected cost per lot in the case of "reworking" is €2000, while on the other hand, we know that if we "process" the lots, 80% of the time we will pay €1000 while 20% of the time we will pay €4000. Thus, we can compute the *expected value* (cost) at node 1 as:

$$EV(1) = 0.8 \times 1000 + 0.2 \times 4000 = 1600 \tag{8.15}$$

That is, on the long run, if we process every lot, we incur an average cost of €1600 per lot. Thus, up to this point, we already know that, with respect to the branch "no test," the best option is "processing," with an expected cost of €1600.

Next, in a similar fashion, we are going to work on the bottom part of the tree, the "test" branch. If the expected cost coming out of the bottom part of the tree is less than 1600, we will recommend to execute the test. Otherwise, we will recommend to process every lot.

Let us now analyze the bottom part of the tree. If we decide to execute the test, we then have an event node, accounting for the result of the test. That is, the test is executed, and the only thing that we can do is to observe the result of the test. Thus, we must now use an event node (node 5 in Figure 8.7). The test can produce one of the three observations defined in $O = \{D_0, D_1, D_2\}$. Thus, we create three branches, one per observation. Let us now select, *e.g.*, branch D_0, *i.e.*, we ran the test, and both chips were working. Obviously, since the test is based on a sample of a larger population, we cannot automatically conclude that the batch is of good quality, even if the two chips were working. We might guess that, since the result of the test was satisfactory, the probability of having a good lot *after* such a result of the test should be higher that the probability a priory of having a good lot ($p(G)$). However, we cannot exclude the possibility that such a sample were drawn out of a "bad" population. Thus, after receiving the response of the test, all the options should be reconsidered, *i.e.*, processing *and* reworking. That means, that the part of the tree to the right of the branch D_0 is identical to the top part of the tree corresponding to the branch "no test." As we will see, the only things changing are the probabilities we use at the leaves of these branches.

Consequently, the tree of Figure 8.7 is now completed: For each possible result of the test, we leave all the options open, and we replicate the subtree with two branches (processing and reworking). Let us now focus on event node 2: The result of the test was D_0 and we decided to process the entire lot. The question is: What is the probability of this lot being "good" as opposed to "bad"? Can we just

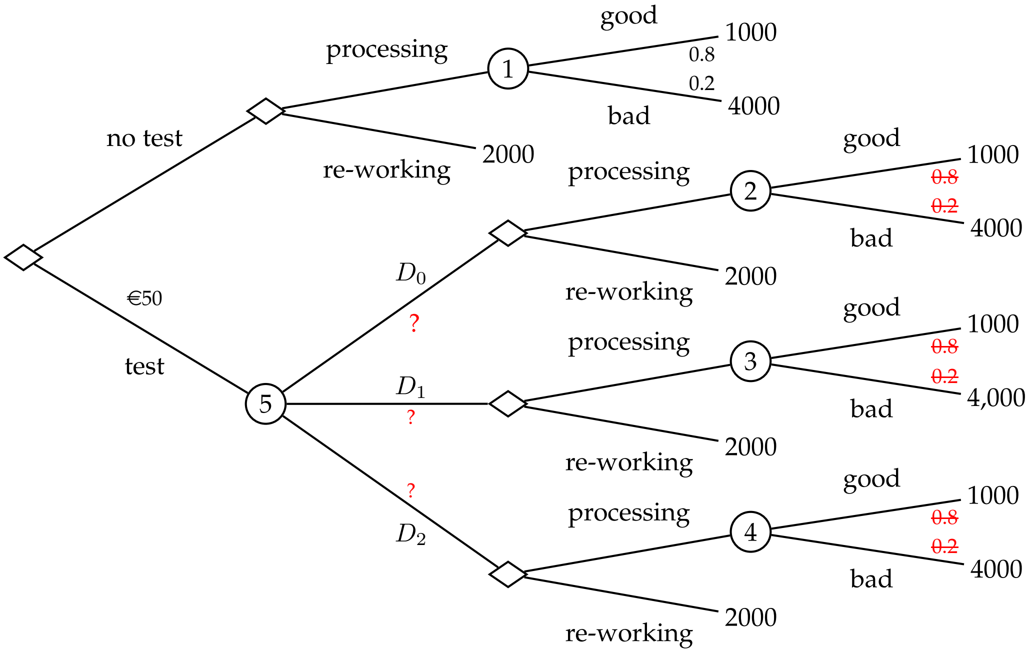


Figure 8.7: A decision tree for the Electronic Chips problem.

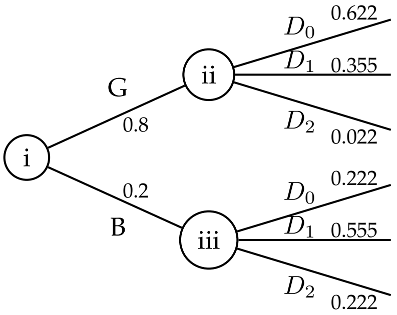
say that, *after* knowing the result of the test, the probabilities of having a good/bad lot are still the same?

The answer to the aforementioned question is no. If we used the *a priori* probabilities for the branches at nodes 2, 3, and 4, we would be making the same mistake of the prosecutor, presented in Section 8.2. The probabilities that should be used for the leaves of nodes 2, 3, and 4 are the *conditional probabilities* obtained using Bayes' theorem. That is, the probability of having a good lot *after* having D_0 as result of the test is $p(G|D_0)$. Similarly, the probability of having a bad lot *after* having D_0 as result of the test is $p(B|D_0)$. Thus, at this point in time, we need to compute the conditional probabilities using Bayes' theorem.

We will now present two alternative approaches to compute Bayes' probabilities. The first approach is based upon the use of the standard formula of Equation (8.2). Next, we will use the method presented in Figure 8.3 to compute the same probabilities.

Let us first use the standard formula to compute $p(G|D_0)$:

$$\begin{aligned} p(G|D_0) &= \frac{p(G) \times p(D_0|G)}{p(G) \times p(D_0|G) + p(B) \times p(D_0|B)} \\ &= \frac{0.8 \times 0.622}{0.8 \times 0.622 + 0.2 \times 0.222} \\ &= \frac{0.4976}{0.542} = 0.918 \end{aligned}$$



$p(S \text{ AND } O)$			$p(O)$
$p(G \text{ AND } D_0)$	$0.8 \times 0.622 = 0.498$	(1)	$p(D_0) = (1) + (4)$
$p(G \text{ AND } D_1)$	$0.8 \times 0.355 = 0.284$	(2)	
$p(G \text{ AND } D_2)$	$0.8 \times 0.022 = 0.018$	(3)	$p(D_1) = (2) + (5)$
$p(B \text{ AND } D_0)$	$0.2 \times 0.222 = 0.044$	(4)	
$p(B \text{ AND } D_1)$	$0.2 \times 0.556 = 0.111$	(5)	$p(D_2) = (3) + (6)$
$p(B \text{ AND } D_2)$	$0.2 \times 0.222 = 0.044$	(6)	
	$\sum = 1$		

Figure 8.8: Bayes' formulas for the Electronic Chips exercise computed using an auxiliary tree.

In a similar fashion, we can compute the remaining probabilities:

$$\begin{aligned}
p(G|D_1) &= \frac{p(G)p(D_1|G)}{p(G)p(D_1|G) + p(B)p(D_1|B)} \\
&= \frac{0.8 \times 0.355}{0.8 \times 0.355 + 0.2 \times 0.555} = \frac{0.284}{0.395} = 0.718 \\
p(G|D_2) &= \frac{p(G)p(D_2|G)}{p(G)p(D_2|G) + p(B)p(D_2|B)} \\
&= \frac{0.8 \times 0.022}{0.8 \times 0.022 + 0.2 \times 0.222} = \frac{0.018}{0.062} = 0.286
\end{aligned}$$

Probabilities $p(B|D_0), p(B|D_1)$, and $p(B|D_2)$ can be obtained either applying the formula of Equation (8.2), or simply as the complementary probability, *e.g.*, $p(B|D_0) = 1 - p(G|D_0)$, *etc.*

Alternatively, we can use the method presented in Figure 8.3, *i.e.*, through the use of an auxiliary tree.

Once we have the auxiliary tree, Bayes' probabilities can easily be computed as:

$$\begin{aligned}
p(G|D_0) &= \frac{(1)}{(1) + (4)} = 0.918 \\
p(G|D_1) &= \frac{(2)}{(2) + (5)} = 0.719 \\
p(G|D_2) &= \frac{(3)}{(3) + (6)} = 0.286 \\
p(B|D_0) &= \frac{(4)}{(1) + (4)} = 0.082 \\
p(B|D_1) &= \frac{(5)}{(2) + (5)} = 0.281 \\
p(B|D_2) &= \frac{(6)}{(3) + (6)} = 0.714
\end{aligned}$$

We can easily see that the sum of the complementary probabilities is always equal to one, *e.g.*, $p(G|D_0) + p(B|D_0) = 1$. In addition, we might want to verify that the sum of the probabilities of the observations is also equal to one, *i.e.*, $p(D_0) + p(D_1) + p(D_2) = 1$.

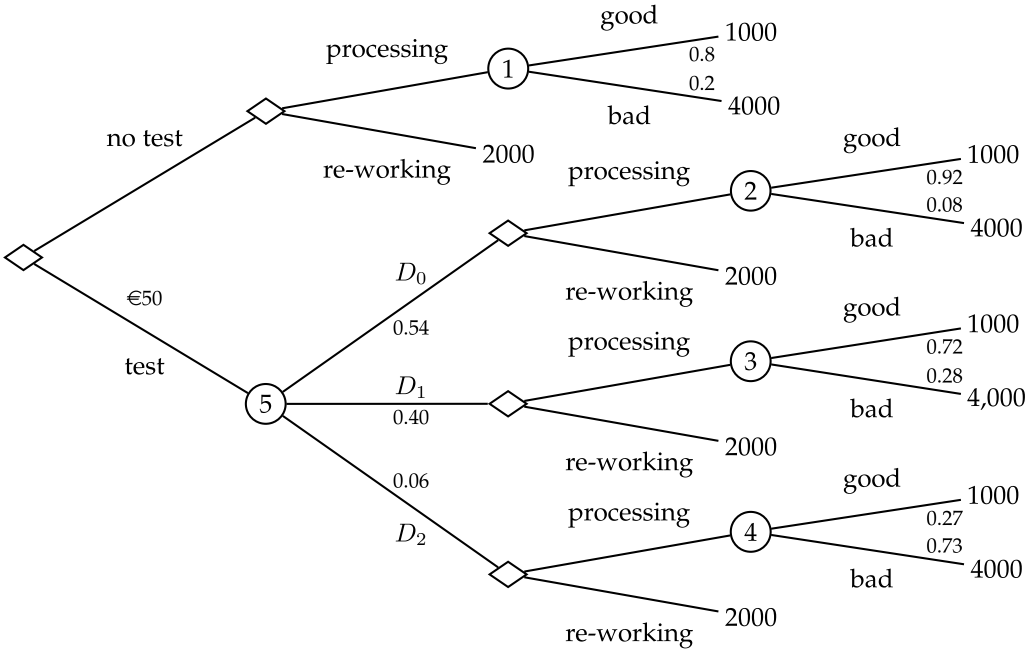


Figure 8.9: A decision tree for the Electronic Chips problem. Bayes' probabilities are used in the "test" branch.

We can now insert Bayes probabilities into the decision tree of Figure 8.7. Figure 8.9 presents the final version of the decision tree for the Electronic Chips exercise. Using the tree of Figure 8.9, we can now work backward, *i.e.*, from the leaves back to the root of the tree, to identify the best alternative.

Let us review one more time how to work the tree backward. Let us focus on the top part of the tree, the "no test" side. As already mentioned, we know that $EV(1) = 1600$. Thus, next to the event node 1 we can write the cost of €1600, corresponding to the expected value at that event node. Next, we go one step backward, to the decision node. Since the decision node is characterizing a point in which the decision maker needs to choose among different alternatives, at every decision node we will select the best option, depending on the goal of the problem. Provided that in this problem we are looking for the minimum cost policy, at every decision node we will pick the alternative the minimizes the expected cost. In the case of the decision node at the end of the "no test" branch, two alternatives must be considered: Either reworking, with an expected cost of €2000, or processing, with an expected cost of €1600. Thus, a rational decision maker will select the alternative processing. Consequently, below the decision node, we write down the cost associated to the best alternative, selecting the minimum cost option, *i.e.*, processing with a cost of €1600. Figure 8.10 presents the backward analysis of a portion of the tree. The general rule employed is the following:

- At every event node, we compute the expected value, as sum-product of the utility values (costs, revenues, *etc.*) with the conditional probabilities.

- At every decision node, we select the best alternative, minimizing or maximizing, depending on the goal of the problem.

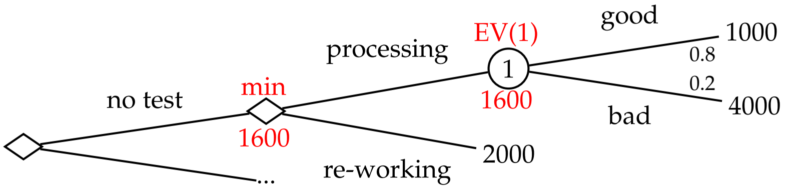


Figure 8.10: A decision tree for the Electronic Chips problem. Bayes' probabilities are used in the "test" branch.

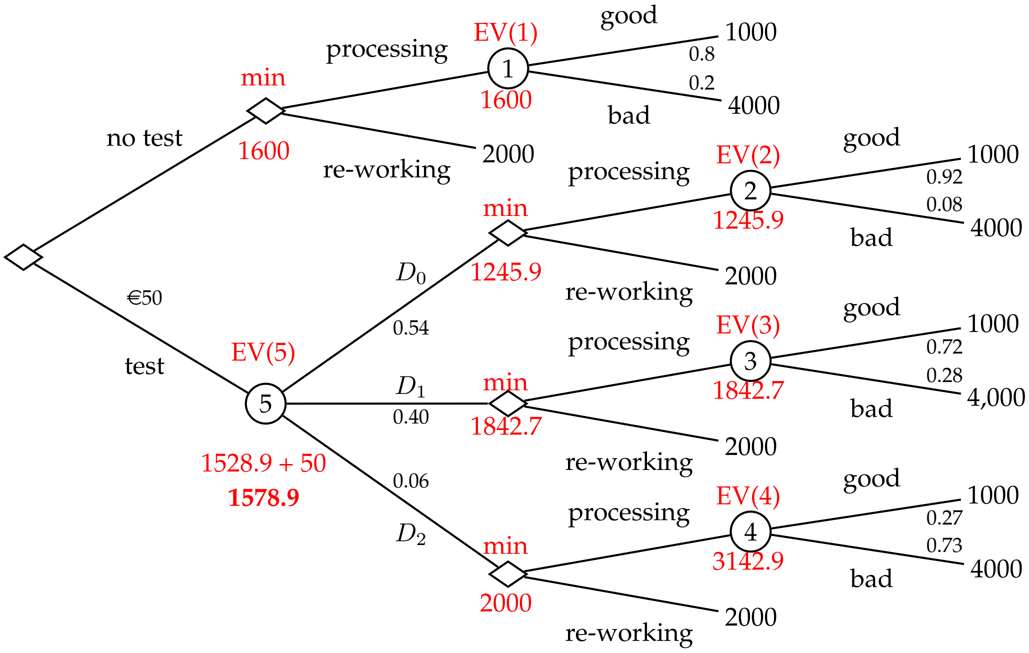


Figure 8.11: A decision tree for the Electronic Chips problem. Bayes' probabilities are used in the "test" branch.

In a similar fashion, we now need to work the tree backward for the remaining leaves, thus computing the expected values at nodes (2) to (5). Figure 8.11 presents the backward procedure, while the equations below illustrate how the expected values have been obtained. Once we obtain the expected value at node 5, we just need to include the cost of the test (€50) to obtain the expected cost of the "test" branch, €1578.9.

$$\begin{aligned}
VE(2) &= \quad 0.918 \times 1000 + 0.082 \times 4000 = 1245.9 \\
VE(3) &= \quad 0.719 \times 1000 + 0.281 \times 4000 = 1842.7 \\
VE(4) &= \quad 0.286 \times 1000 + 0.714 \times 4000 = 3142.9 \\
VE(5) &= \quad 1245.9 \times 0.54 + 1842.7 \times 0.40 + 2000 \times 0.06 = 1528.9
\end{aligned}$$

Therefore, we now have the final recommendation for the company. The minimum cost policy is the following:

- Do the test: If the test gives zero or one defective chips (D_0 or D_1), then we recommend to process the lot; if the test gives two defective chips (D_2), then we recommend to rework the lot.

- The expected cost per lot associated to the recommended strategy is €1578.9.

- What is the maximum price that you should be willing to pay for such a test? The answer is related to the value of the information conveyed by the test itself. Let us assume that the test were free. In this case, we would have an expected cost per lot of €1528.9. On the other hand, without test, we would incur a cost of €1600 per lot. Thus, we can conclude that test allows to save costs and, more precisely, the information conveyed by the test is worth 1600 - 1528.9 = €71.1, *i.e.*, the difference between the expected cost of the best option without test and the cost with the test.