Updating Predictions

References:

Chapter 18

https://www.inferentialthinking.com/chapters/18/Updating_Predictions.html

Updating Predictions

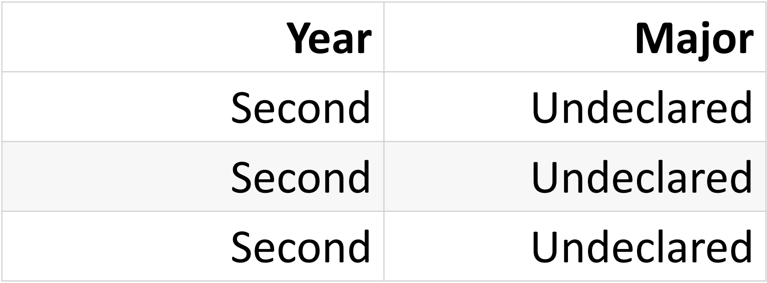
- We know how to use training data to classify a point into one of two categories.
- Our classification is just a prediction of the class, based on the most common class among the training points that are nearest our new point.
- Suppose that we eventually find out the true class of our new point.
- Then we will know whether we got the classification right. Also, we will have a new point that we can add to our training set, because we know its class.
- This *updates* our training set. So, naturally, we will want to *update our classifier* based on the new training set.
- This chapter looks at some simple scenarios where new data leads us to update our predictions. While the examples in the chapter are simple in terms of calculation, the method of updating can be generalized to work in complex settings and is one of the most powerful tools used for machine learning.

A "More Likely Than Not" Binary Classifier

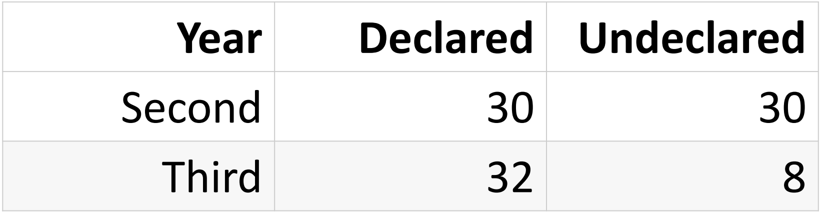
- Let's try to use data to classify a point into one of two categories, choosing the category that we think is more likely than not.
- To do this, we not only need the data but also a clear description of how chances are involved.
- We will start out in a simple artifical setting just to develop the main technique, and then move to a more intriguing example.
- Suppose there is a university class with the following composition:
  - 60% of the students are Second Years and the remaining 40% are Third Years
  - 50% of the Second Years have declared their major
  - 80% of the Third Years have declared their major
- Now suppose **I pick a student at random from the class**. Can you classify the student as Second Year or Third Year, using our "more likely than not" criterion?
- You can, because the student is picked at random and so you know that the chance that the student is a Second Year is 60%. That's greater than the 40% chance of being a Third Year, so you would classify the student as Second Year.
- The information about the majors is irrelevant, as we already know the proportions of Second and Third Years in the class.
- We have a pretty simple classifier! But now suppose I give you some additional information about the student who was picked:
- **The student has declared a major.**
- Would this knowledge change your classification?

Updating the Prediction Based on New Information

- Now that we know the student has declared a major, it becomes important to look at the relation between year and major declaration.
- It's still true that more students are Second Years than Third Years. But it's also true that among the Third Years, a much higher percent have declared their major than among the Second Years.
- Our classifier has to take both of these observations into account.
- To visualize this, we will use a table students that consists of one row for each of 100 students whose years and majors have the same proportions as given in the data.
- The total count is 100 students, of whom 60 are Second Years and 40 are Third Years.
- Among the Second Years, 50% are in each of the Major categories. Among the 40 Third Years, 20% are Undeclared and 80% Declared.
- So this population of 100 students has the same proportions as the class in our problem, and we can assume that our student has been picked at random from among all 100 students.



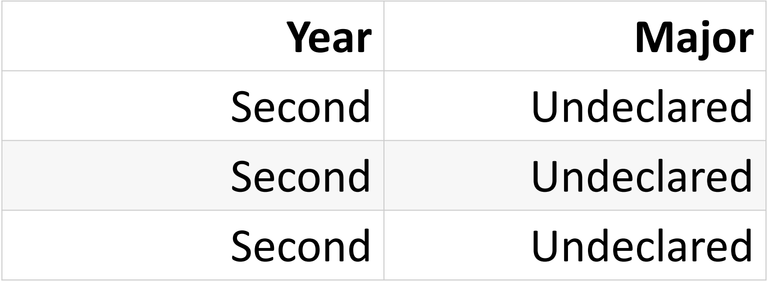
Year	Major
Second	Undeclared
Second	Undeclared
Second	Undeclared



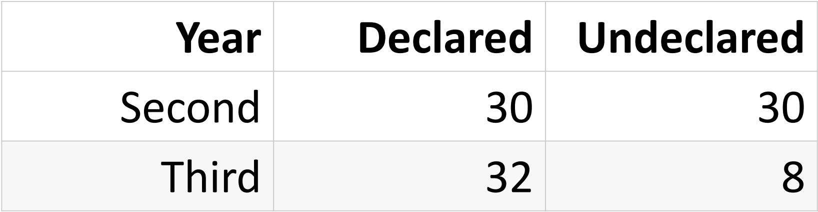
Year	Declared	Undeclared
Second	30	30
Third	32	8

Updating the Prediction Based on New Information

- We have to pick which row the student is most likely to be in. When we knew nothing more about the student, he or she could be in any of the four cells, and therefore were more likely to be in the top row (Second Year) because that contains more students.
- But now we know that the student has declared a major, so the space of possible outcomes has decreased: now the student can only be in one of the two Declared cells.
- There are 62 students in those cells, and 32 out of the 62 are Third Years. That's more than half, even though not by much.
- So, in the light of the new information about the student's major, we have to update our prediction and now classify the student as a Third Year.
- What is the chance that our classification is correct? We will be right for all the 32 Third Years who are Declared, and wrong for the 30 Second Years who are Declared. The chance that we are correct is therefore about 0.516.
- In other words, the chance that we are correct is **the proportion of Third Years among the students who have Declared**.
- 32/(30+32) = 0.516



Year	Major
Second	Undeclared
Second	Undeclared
Second	Undeclared

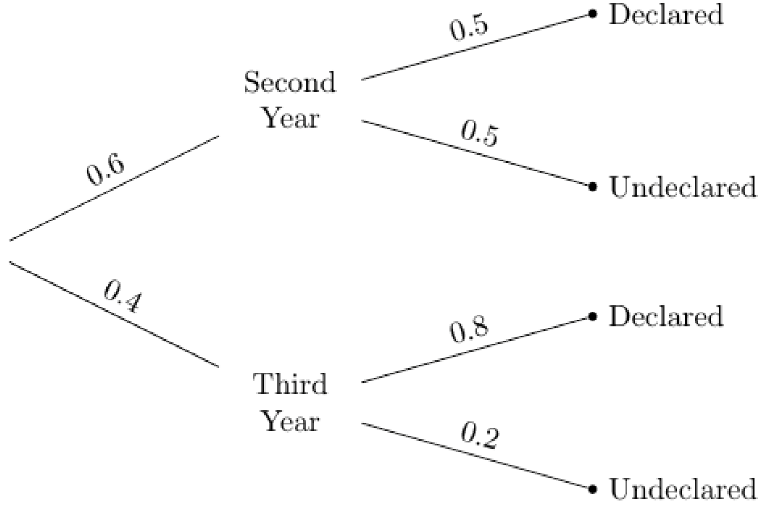


Year	Declared	Undeclared
Second	30	30
Third	32	8

Tree Diagram

- The proportion that we have just calculated was based on a class of 100 students. But there's no reason the class couldn't have had 200 students, for example, as long as all the proportions in the cells were correct.
- Then our calculation would just have been 64/(60 + 64) which is 0.516 as before.
- So the calculation depends only on the proportions in the different categories, not on the counts.
- The proportions can be visualized in a *tree diagram*, shown directly below the pivot table for ease of comparison.
- Like the pivot table, this diagram *partitions* the students into four distinct groups known as "branches". Notice that the "Third Year, Declared" branch contains the proportion 0.4 x 0.8 = 0.32 of the students, corresponding to the 32 students in the "Third Year, Declared" cell of the pivot table.
- The "Second Year, Declared" branch contains 0.6 x 0.5 = 0.3 of the students, corresponding to the 30 in the "Second Year, Declared" cell of the pivot table.
- So, given that the student is Declared, the chance of them being a Third Year can be calculated directly from the tree. The answer is the proportion in the "Third Year, Declared" branch relative to the total proportion in the two "Declared" branches.
- That is, the answer is **the proportion of Third Years among students who are Declared**, as before.
- (0.4 * 0.8)/(0.6 * 0.5 + 0.4 * 0.8) = 0.516

Year	Declared	Undeclared
Second	30	30
Third	32	8



Bayes' Rule

- The method that we have just used is due to the Reverend Thomas Bayes (1701-1761).
- His method solved what was called an "inverse probability" problem: given new data, how can you update chances you had found earlier?
- Though Bayes lived three centuries ago, his method is widely used now in machine learning.

Bayes' Rule

- **Prior probabilities.** Before we knew the chosen student's major declaration status, the chance that the student was a Second Year was 60% and the chance that the student was a Third Year was 40%. These are the *prior* probabilities of the two categories.
- **Likelihoods.** These are the chances of the Major status, given the category of student; thus they can be read off the tree diagram. For example, the likelihood of Declared status given that the student is a Second Year is 0.5.
- **Posterior probabilities.** These are the chances of the two Year categories, *after* we have taken into account information about the Major declaration status. We computed one of these

Bayes' Rule

- The posterior probability that the student is a Third Year, given that the student has Declared, is denoted P(Third Year | Declared) and is calculated as follows.
  - P(Third Year | Declared) =

    $$\frac{\text{(prior probability of Third Year)×(likelihood of Declared given Third Year)}}{\text{total probability of Declared}}$$

    $= (0.4 * 0.8)/(0.6 * 0.5 + 0.4 * 0.8) = 0.51$

  - P(Second Year | Declared) =

    $$\frac{\text{(prior probability of Second Year)×(likelihood of Declared given Second Year)}}{\text{total probability of Declared}}$$

    $= (0.6 * 0.5)/(0.6 * 0.5 + 0.4 * 0.8) = 0.484$

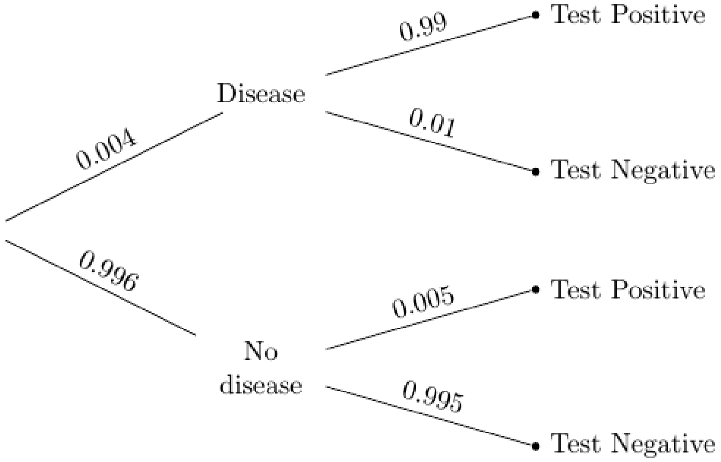
- Notice that both the posterior probabilities have the same denominator: the chance of the new information, which is that the student has Declared.
- Because of this, Bayes' method is sometimes summarized as a statement about proportionality: **posterior ∝ prior × likelihood**
- **Tree diagrams help**

Making Decisions

- A primary use of Bayes' Rule is to make decisions based on incomplete information, incorporating new information as it comes in.
- This section points out the importance of keeping your assumptions in mind as you make decisions.
- Many medical tests for diseases return Positive or Negative results.
- A Positive result means that according to the test, the patient has the disease. A Negative result means the test concludes that the patient doesn't have the disease.
- Medical tests are carefully designed to be very accurate. But few tests are accurate 100% of the time. Almost all tests make errors of two kinds:
  - A **false positive** is an error in which the test concludes Positive but the patient doesn't have the disease.
  - A **false negative** is an error in which the test concludes Negative but the patient does have the disease.

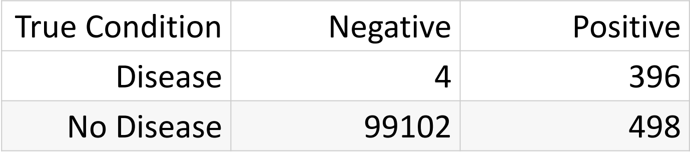
A Test for a Rare Disease

- Suppose there is a large population and a disease that strikes a tiny proportion of the population. The tree diagram below summarizes information about such a disease and about a medical test for it.
- Overall, only 4 in 1000 of the population has the disease. The test is quite accurate: it has a very small false positive rate of 5 in 1000, and a somewhat larger (though still small) false negative rate of 1 in 100.
- Individuals might or might not know whether they have the disease; typically, people get tested to find out whether they have it.
- So **suppose a person is picked at random from the population** and tested. If the test result is Positive, how would you classify them: Disease, or No disease?
- (0.004 * 0.99)/(0.004 * 0.99 + 0.996*0.005) = 0.442
- Given that the person has tested Positive, the chance that he or she has the disease is about 44%. So we will classify them as: No disease.
- This is a strange conclusion. We have a pretty accurate test, and a person who has tested Positive, and our classification is ... that they **don't** have the disease? That doesn't seem to make any sense.
- When faced with a disturbing answer, the first thing to do is to check the calculations. The arithmetic above is correct. Let's see if we can get the same answer in a different way.



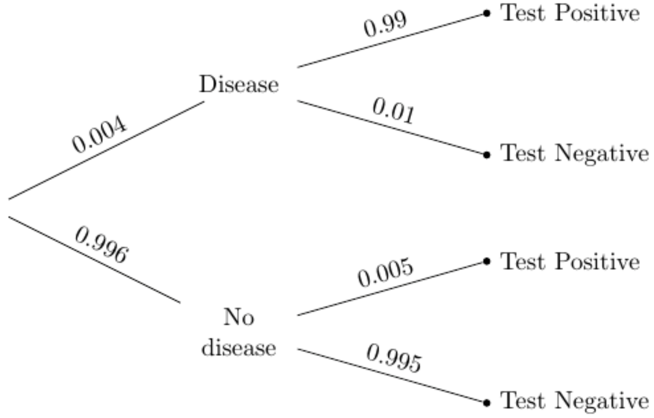
A Test for a Rare Disease

- We will call population with 0.004 as the argument, and then pivot to cross-classify each of the 100,000 people.
- The cells of the table have the right counts. For example, according to the description of the population, 4 in 1000 people have the disease. There are 100,000 people in the table, so 400 should have the disease.
- That's what the table shows: 4 + 396 = 400.



True Condition	Negative	Positive
Disease	4	396
No Disease	99102	498

- Of these 400, 99% get a Positive test result: 0.99 x 400 = 396.
- Among the Positives, the proportion that have the disease is: 396/(396 + 498) = 0.44
- That's the answer we got by using Bayes' Rule. The counts in the Positives column show why it is less than 1/2. Among the Positives, more people **don't** have the disease than do have the disease.
- The reason is that a huge fraction of the population doesn't have the disease in the first place. The tiny fraction of those that falsely test Positive are still greater in number than the people who correctly test Positive. This is easier to visualize in the tree diagram:
    - The proportion of true Positives is a large fraction (0.99) of a tiny fraction (0.004) of the population.
    - The proportion of false Positives is a tiny fraction (0.005) of a large fraction (0.996) of the population.



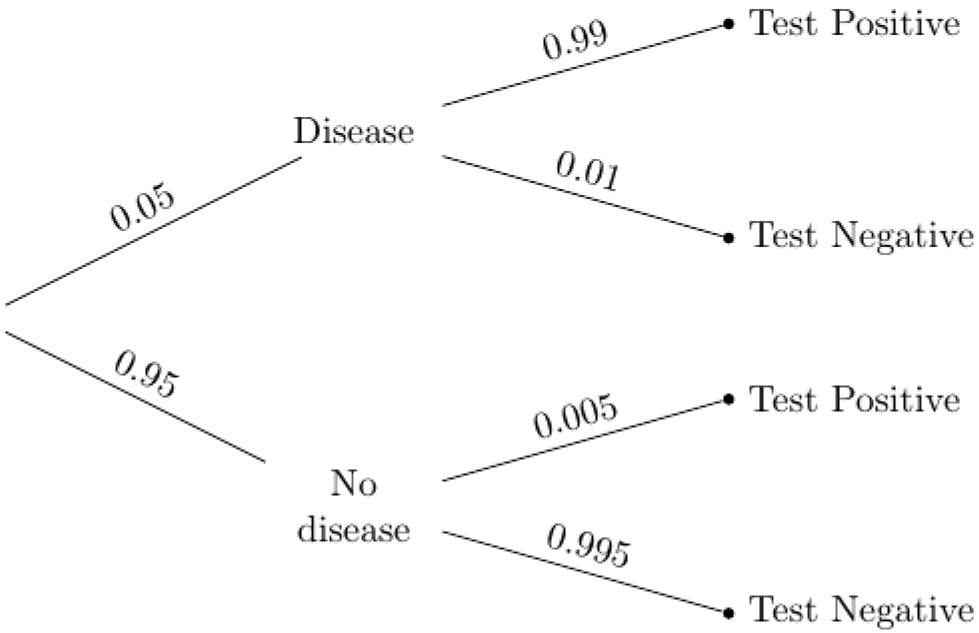
- These two proportions are comparable; the second is a little larger.
- So, given that the randomly chosen person tested positive, we were right to classify them as more likely than not to **not** have the disease

A Subjective Prior

- Being right isn't always satisfying.
- Classifying a Positive patient as not having the disease still seems somehow wrong, for such an accurate test. Since the calculations are right, let's take a look at the basis of our probability calculation: the assumption of randomness.
- Our assumption was that a randomly chosen person was tested and got a Positive result. But this doesn't happen in reality. People go in to get tested because they think they might have the disease, or because their doctor thinks they might have the disease. **People getting tested are not randomly chosen members of the population.**
- That is why our intuition about people getting tested was not fitting well with the answer that we got. We were imagining a realistic situation of a patient going in to get tested because there was some reason for them to do so, whereas the calculation was based on a randomly chosen person being tested.
- So let's redo our calculation under the more realistic assumption that the patient is getting tested because the doctor thinks there's a chance the patient has the disease.
- Here it's important to note that "the doctor thinks there's a chance" means that the chance is the doctor's opinion, not the proportion in the population. It is called a *subjective probability*. In our context of whether or not the patient has the disease, it is also a *subective prior* probability.

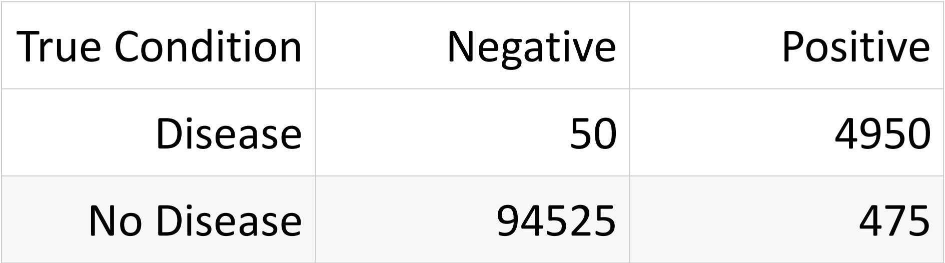
A Subjective Prior

- Suppose the doctor's subjective opinion is that there is a 5% chance that the patient has the disease. Then just the prior probabilities in the tree diagram will change:
- Given that the patient tests Positive, the chance that he or she has the disease is given by Bayes' Rule.
- (0.05 * 0.99)/(0.05 * 0.99 + 0.95 * 0.005) = 0.91
- The effect of changing the prior is stunning. Even though the doctor has a pretty low prior probability (5%) that the patient has the disease, once the patient tests Positive the posterior probability of having the disease shoots up to more than 91%.
- If the patient tests Positive, it would be reasonable for the doctor to proceed as though the patient has the disease.



Confirming the Answer

- Though the doctor's opinion is subjective, we can generate an artificial population in which 5% of the people have the disease and are tested using the same test.
- Then we can count people in different categories to see if the counts are consistent with the answer we got by using Bayes' Rule.
- In this artificially created population of 100,000 people, 5000 people (5%) have the disease, and 99% of them test Positive, leading to 4950 true Positives. Compare this with 475 false Positives: among the Positives, the proportion that have the disease is the same as what we got by Bayes' Rule.
- 4950/(4950 + 475) = 0.9124423963133641
- Take a simple random sample of size 10,000 from the population, and extract the table positive consisting only of those in the sample that had Positive test results.
- Run the two cells a few times and you will see that the proportion of true Positives among the Positives hovers around the value of 0.912 that we calculated by Bayes' Rule.



True Condition	Negative	Positive
Disease	50	4950
No Disease	94525	475