Machine Learning: Homework 8

Laurent HAYEZ

November 28, 2016

Exercise 1. *Imagine running a medical test for the fictional disease Sthgiw in Llefretniw (population = 1,000), where 40% are infected. The test has a false positive rate (FP/C-) of 5% and no false negative rate (FN/C+). Create the confusion matrix. What's the positive predictive value (probability that it correctly indicates an infection if a person receives a positive test)?*

Now consider the same test applied in Nurrevir (population = 1,000), but where only 2% are infected. Create the confusion matrix and recalculate the probability of actually being infected after one is told that one is infected using the same medical test.

Suddenly we have the new disease Sreklaw. Only one in a million people gets this disease. We develop a new test that gives us 99% of the time the correct result (99 percent of the time, it gives true if the subject is infected, and false if the subject is healthy). We give the test to everybody in Soretsew (population = 1,000,000). How happy are we with our 99% accurate test?

Solution. According to the confusion matrix shown in Table 1, we deduce that the proportion of tests predicted negative which actually are negative is 0.57. Indeed, 60% of the individuals actually are negative (don't have the disease), and 5% of them are tested positive. Hence 3% of the individuals are false positive, and 57% are true negative. Applying the same reasoning for the individuals who have the disease, we obtain that 40% of the individuals are true positive and 0% are false negative.

The positive predictive value is given by $\frac{TP}{TP+FP} = \frac{400}{400+30} \simeq 0.93$.

Doing the same calculation if 2% of the population is infected, we obtain the confusion matrix shown in Table 2. That means that the positive predictive value is $\sim 29\%$.

If we do the same calculations as before for the new disease, we obtain the con-

fusion matrix shown in Table 3. The positive predictive value is $\sim 9.9 \cdot 10^{-5}$, or in other words, the test is completely inaccurate to detect the disease. However the negative predictive value is 0.9999999899, i.e., very accurate to detect that a healthy individual is not affected by the disease. □

Table 1: Confusion matrix for Exercise 1 - part 1

		Tested		
		Positive	Negative	
Actual	Positive	400	0	400
	Negative	30	570	600
		430	570	1000

Table 2: Confusion matrix for Exercise 1 - part 2

		Tested		
		Positive	Negative	
Actual	Positive	20	0	20
	Negative	49	931	980
		69	931	1000

Table 3: Confusion matrix for Exercise 1 - part 3

		Tested		
		Positive	Negative	
Actual	Positive	0.99	0.01	1
	Negative	9'999.99	989'999.01	999'999
		10'000.98	989'999.02	1'000'000

Exercise 2. *Create the ordered rule list. With that, form a single rule and an unordered rule list.*

Classify the following sample according to your rules:

Solution. By observing the table, we see that the rule covering the largest number of information is

R1: IF TV = Yes, THEN Male.

For the next one, we see that if the attributes Paper and Internet are Yes, or the

attribute Magazine is No we always have a woman. The attributed that cover the largest set is Paper. Hence the second rule is

R2: IF Paper = Yes, THEN Female.

Then we only have three samples left. The third rule is

R3: IF Magazine = Yes, THEN Male.

The last observation is a woman, hence we obtain

R4: ELSE Female.

Making this as a single rule, we obtain

```
If (TV == Yes):
    Male
Else if (Paper == Yes):
    Female
Else if (Magazine == Yes):
    Male
Else:
    Female
```

We can also create an unordered rule list:

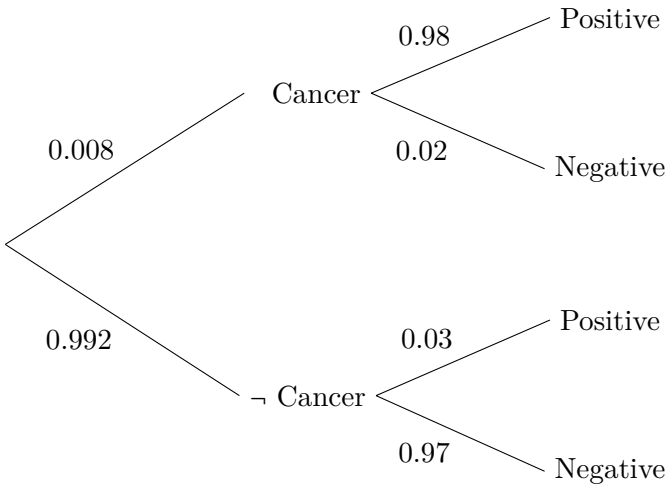
```
R1 : If (TV == Yes) => Male
R2': If (TV == No && Paper == Yes) => Female
R3': If (TV == No && Paper == No && Magazine == Yes) => Male
R4': If (TV == No && Paper == No && Magazine == No) => Female
```

The new sample would be classified according $R3'$, hence it would be classified as a male. Note that this sample is the same as sample 09. □

Exercise 3. *Match up the unordered English statements with their associated probability notations and write the probabilities (calculations with Bayes' Theorem and normalization might be needed). If there is no English statement matching a probability, please write one.*

We know that 0.8% of the people have cancer. If cancer is present, the test returns a correct positive result 98% of the time. It returns a correct negative result 97% of the time if the cancer is not present.

Solution. Let's draw a probability tree to help.



Let A be the event "individual has cancer" and let B be the event "the test is positive".

$$\mathbb{P}(A) = 0.008$$

$$\mathbb{P}(\neg A) = 0.0992$$

$$\mathbb{P}(A \mid B) = \frac{0.008 \cdot 0.98}{0.008 \cdot 0.98 + 0.992 \cdot 0.03} = \frac{49}{235} \simeq 0.21$$

$$\mathbb{P}(A \mid \neg B) = \frac{0.008 \cdot 0.02}{0.008 \cdot 0.02 + 0.992 \cdot 0.97} = \frac{1}{6015} \simeq 1.66 \cdot 10^{-4}.$$

$$\mathbb{P}(\neg A \mid B) = \frac{0.992 \cdot 0.03}{0.992 \cdot 0.03 + 0.008 \cdot 0.98} = \frac{186}{235} \simeq 0.79$$

$$\mathbb{P}(\neg A \mid \neg B) = 1 - \mathbb{P}(A \mid \neg B) \simeq 0.99$$

$$\mathbb{P}(B \mid A) = 0.98$$

$$\mathbb{P}(\neg B \mid A) = 0.02$$

$$\mathbb{P}(B \mid \neg A) = 0.03$$

$$\mathbb{P}(\neg B \mid \neg A) = 0.97.$$

□