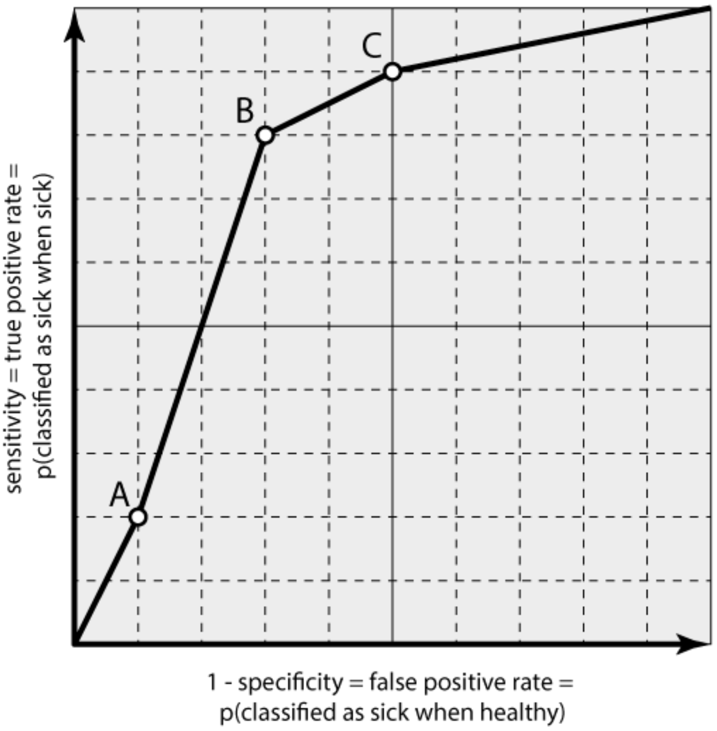
Homework #4: Operating points

Sara is a veterinarian who treats hamsters for Chomsky disease. About 20 % of hamsters she tests have this disease (luckily, it's not serious, it only makes them run backwards on the running wheel).



She can make two kinds of mistakes:

- Failing to detect the disease when present costs her 1000 euros (lawsuits etc.).
- Treating a hamster that is actually healthy costs her 600 euros (lawsuits etc.).

Don't worry about her, in both cases she charges enough to survive.

Her choice whether to administer the cure will be based on the classifier that predicts the probability of the disease from the observed symptoms. The classifier she uses is not perfect, as shown in the ROC curve.

1 What is the false positive rate at each point marked on the curve?

2 What is the true positive rate at each point?

3 What are then the probabilities of making one or the other of the above mistakes at each point?

4 What is the expected cost of mistakes at each point?

5 Sara has accidentally put a sick and a healthy subject (that is, hamster) in the same cage. Now she doesn't know which is which. She is going to diagnose both hamsters and administer the cure to the one which she believes is more likely to be sick. Compute is the probability that she'll pick the ***wrong*** one.

Solution

1. We read false positive rates (FPR) for points A, B and C directly from the graph: they are 0.1, 0.3 and 0.5, respectively.
2. True positive rates (TPR) are also shown in the graph: they are 0.2, 0.8 and 0.9.
3. The probability of treating a healthy hamster, that is, p(+|healthy), equals the false positive rate (she believes the hamster is "positive" when it is actually not). These are 0.1, 0.3 and 0.5, as we read in the first question.

   The probability of not treating a sick one, p(-|sick), is equal to the false negative rate. Consider the point A. If Sara has 100 positives (that is, 100 sick hamsters), she would correctly classify 20 as positive — these are the true positives that the previous question asked about. Hence she would incorrectly classify the remaining 80 as negative. False negative rate is therefore one minus true positive rate; FNR = 1 - TPR. For points A, B and C, false negative rates are 1 - 0.2 = 0.8 and 1 - 0.8 = 0.2 and 1 - 0.9 = 0.1, respectively.

   Some students computed probabilities p(healthy|+) and p(sick|-), and some computed the expected classification error, that is, p(+, healthy) + p(-, sick). The question was indeed ambiguous, so I considered these answers to be correct, too.

   $$p(+, \text{healthy}) + p(-, \text{sick}) = p(+|\text{healthy})\ p(\text{healthy}) + p(-|\text{sick})\ p(\text{sick})$$

   These errors are 24 %, 28 % and 42 %, respectively. The latter is the most interesting: the error rate is high because most hamsters are healthy but the FPR is high.

4. We need to multiply the costs of mistakes by the respective probabilities.

   $$\text{cost} = p(-, \text{sick})\ \text{cost\_FN} + p(+, \text{healthy})\ \text{cost\_FP}$$

   where p(-, sick) = p(-|sick) p(sick), and p(+, health) = p(+ | healthy) p(healthy)

   Costs for points A, B and C are 208, 184 and 260, respectively.

   Some students used p(-|sick) instead of p(-, sick). This ignores the prior probability that the hamster sick. Imagine that only 1 % of hamsters that Sara sees are sick. Or that 99 % of hamsters that she sees are sick. Doesn't this make a difference? In the first case, the FN will be very rare, and so she will rarely pay 1000 euros, while in the second case it will become very common.

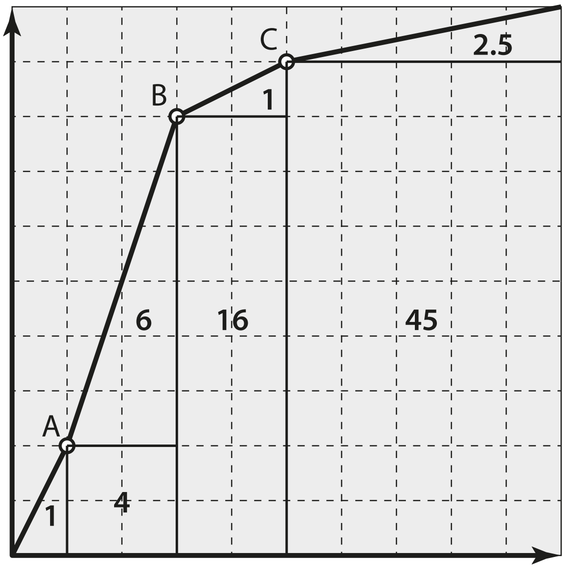
Some students used p(sick|-) instead. This mistake is similar, except that it doesn't ignore the prior probability of being sick, but the prior probability of having a negative test - which in turn depends upon the prior probability of being healthy (and the specificity of the test).

5. The task says that she will "administer the cure to the one which she believes is more likely to be sick". She has to pick one of the two, hence she cannot work with classification lest she can classify both as sick or both as healthy — and then what? The task says "*she believes is more likely*", which translates to "*she assigns a higher probability to*".

   Furthermore, in this task we know that one hamster is sick and one is healthy. This is important! Many students ignored this and compared probabilities of false positives, false negatives and so forth.

   We explained the answer at the lecture and ... well, those who read the paper about ROC curves saw the answer there. (Also: the answer for question 4!)

   So, we know that the area under the ROC curve has a nice probabilistic interpretation. Say that we are given two data instances and we are told that one is positive and the other is negative. We use the classifier to estimate the probabilities of being positive for each instance, and decide that the one with the highest probability is positive. The probability that such a decision is correct equals the AUC of this model.



   The task asks about the opposite probability, the probability of making the wrong decision, which is then 1 - AUC.

   We must compute the AUC from the graph. Conveniently, there are 100 squares, so we just count the number of squares below the curve: the AUC is 0.755, so the probability of picking the wrong hamster is 1 - 0.755 = 0.245.