Machine Learning - Selection of Optimal Classifier

Hanne-Grete Alvheim

May 2022

1 Selection of the appropriate parameter

For this task, I considered the True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). In order to find these, I compared the classified results to the ground truth in the GT.dsv file. What is requested is to not choose the optimal value for this specific set, but the best for other tasks as well. For binary classification, the most critical tasks is to get as few FN's as possible, while also getting as few FP's as possible. For instance in the medical industry it is very important to someone to know if they have a disease, but at the same time a positive result when you are actually negative may result in unnecessary stress. An ROC-curve shows us the highest true positive rate with the lovest false positive rate [1], and can in that way give a good indication of what parameter should be chosen for the optimal result. It plots the false positive rate (FPR) on the x-axis vs the True Positive Rate (recall/sensitivity) on the y-axis. The formulas are respectively:

$$FPR = 1 - \frac{TN}{TN + FP} = 1 - specificity = \frac{FP}{TN + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FP} \tag{2}$$

The ROC curve for classifier 1 is shown in picture 1.

The optimal value of α based on this graph is then α_{22}, which gives quite few false positives and false negatives. The confusion table for α_{22} can be seen in table 1.

This gives this alpha an accuracy of 97%, and a precision = $\frac{TP}{TP+FP}$ of about 0.98, which is quite good.

$\frac{Truevalue}{Classified}$	P	N
P	TP = 48	FP = 1
N	FN = 2	TN = 49

Table 1: Confusion matrix for α_{22}

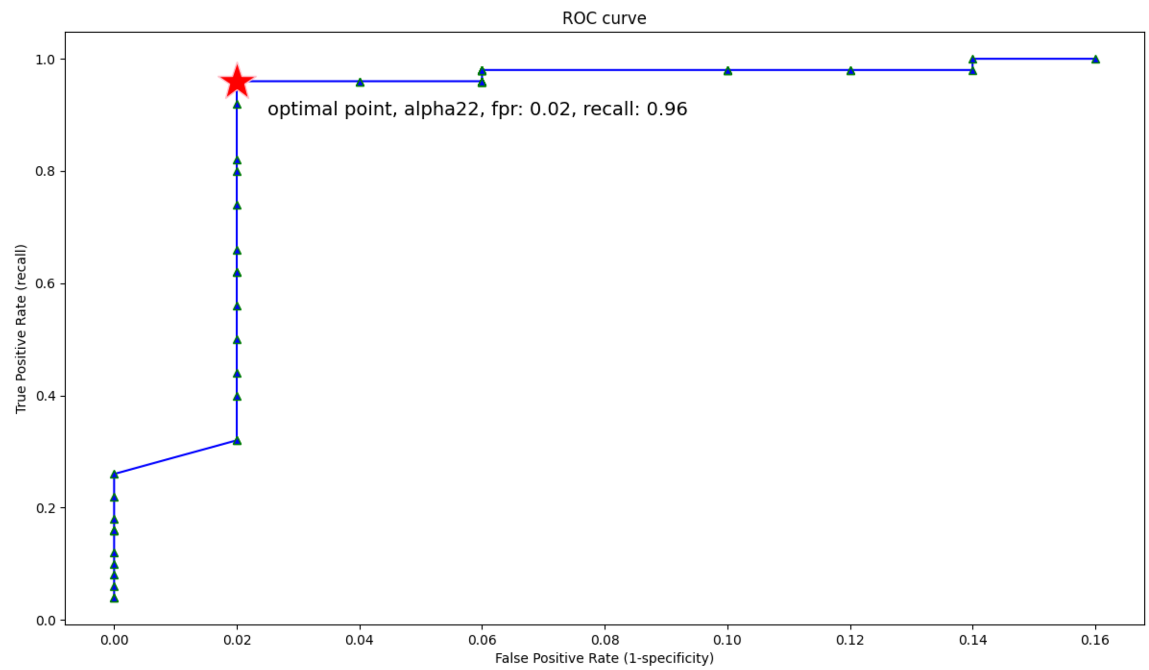

Figure 1: The ROC graph for classifier 1. The optimal point is marked with a red star.

2 Top Secret!

Since it is very important to strive for no false positives as someone can then unlock instead of you, our classifier should have no false positives while also having as few false negatives as possible. This means that we should choose the alpha from the classifier that has 0 false positive rate, while having the highest true positive rate possible. From inspection of the ROC-curves for each classifier this is shown to be Classifier 4 with parameter α_{11}. This gives quite a bad recall, however since the agent always has enough time to unlock the data it does not matter.

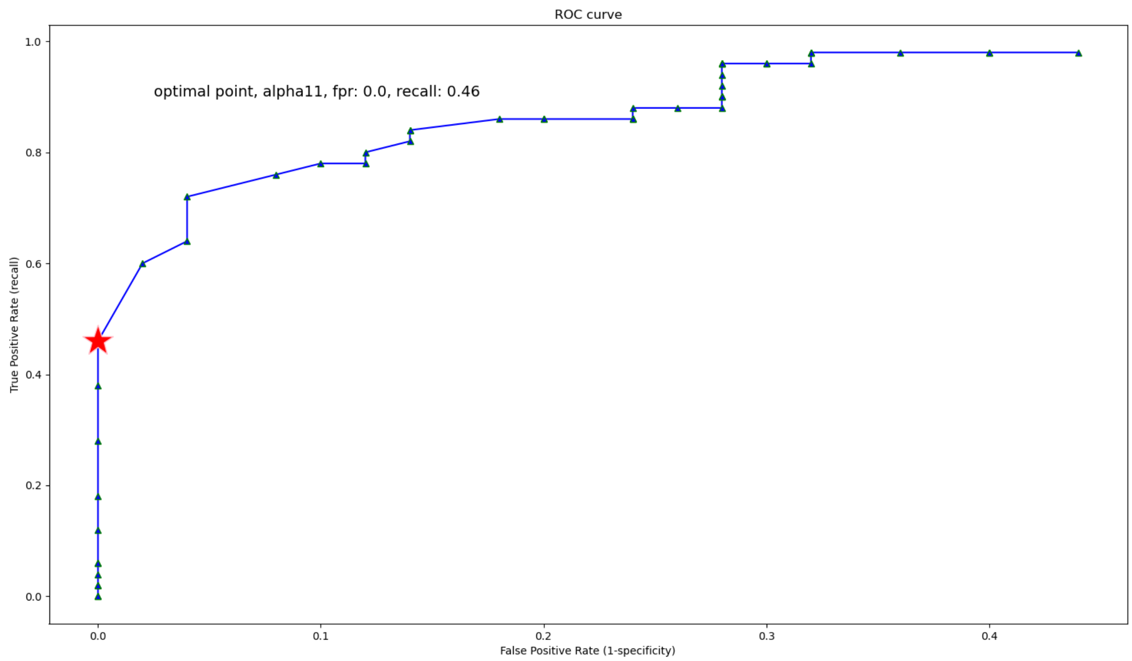

Figure 2: The ROC graph for classifier 4. The optimal point if we want $FPR = 0$ is marked with a red star.

3 Safety First

Not sure if I understand correctly, but I assume that even though the function is made in advance we will still be able to see the ground truth when we compare the two classifiers. Then I can just check if the new classifier has an α-value which gives a $specification = 1$ and a higher recall then the one I chose. Can also check by comparing the two ROC-curves. See the code and especially the function "compare_classifiers" in the python-file.

References

[1] Roc curves – what are they and how are they used? `https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used`. Accessed: 2022-05-10.