COMP135 - Intro to Machine Learning

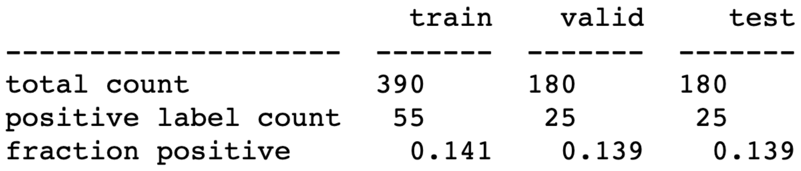
Camelia D. Brumar

October 15, 2020

Contents

1 Problem 1

1.1 Table 1



```
                          train    valid     test
--------------------   -------  -------  -------
total count            390      180      180
positive label count    55       25       25
fraction positive        0.141    0.139    0.139
```

Table 1: Summary of some basic properties of the provided training set, validation set, and test set: total count, positive label count and fraction positive.

1.2 Short answer 1a

> *What accuracy (i.e. baseline_acc from above) does the "predict-0-always" classifier get on the validation set (report to 3 decimal places)? (You should see a pretty good number). Does this mean we should use this classifier?*

Answer: The accuracy of the "predict-0-always" classifier on the validation set is 0.861, which I computed by adding the true negatives and the true positives $(155 + 0)$, and dividing the result by $N = TP + TN + FP + FN = 0 + 155 + 0 + 25 = 180$. It doesn't mean that we should use this classifier even if it does a pretty good job on the validation set. If the purpose of building such classifier is to avoid doing biopsies, since this classifier will always predict that the patient doesn't not have cancer, then it fails the purpose of using it.

1.3 Short answer 1b

> *For the intended application (screening patients before biopsy), describe the possible mistakes the classifier can make in task-specific terms. What costs does each mistake entail (lost time? lost money? life-threatening harm?). How do you recommend evaluating a potential classifier to be mindful of these costs?*

Answer: A classifier (not the "predict-0-always" classifier) might predict that a healthy patient has cancer and a biopsy will be done. In this case there will be a waste of time for the doctors that could have been spent to save sick patients, a loss of money for the patient or the insurance, and of course this biopsies have a chance to go not so well and harming the patient. In the case in which the classifier says that a sick patient is healthy, the consequence is that this patient will not be treated and will most probably die soon.

1.4 Figure 1

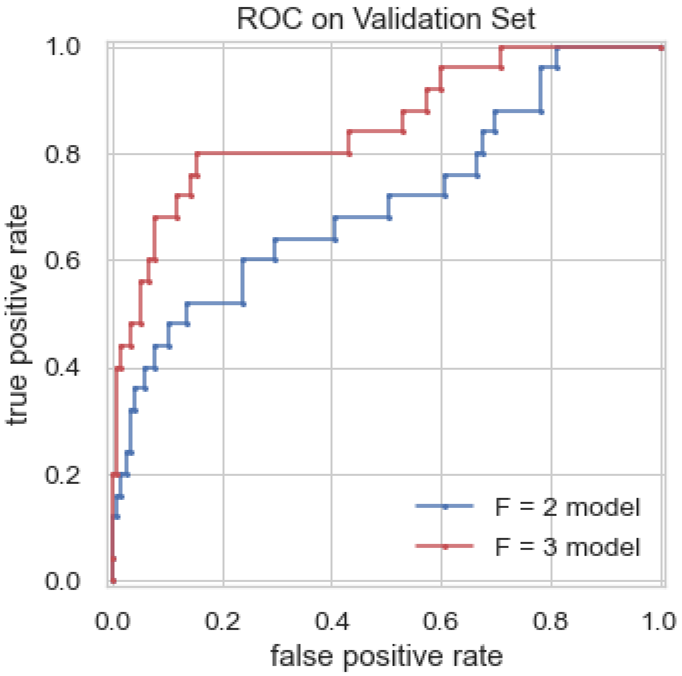


Figure 1: Comparison of the F=2 and F=3 model's performance on the validation set, using ROC curves.

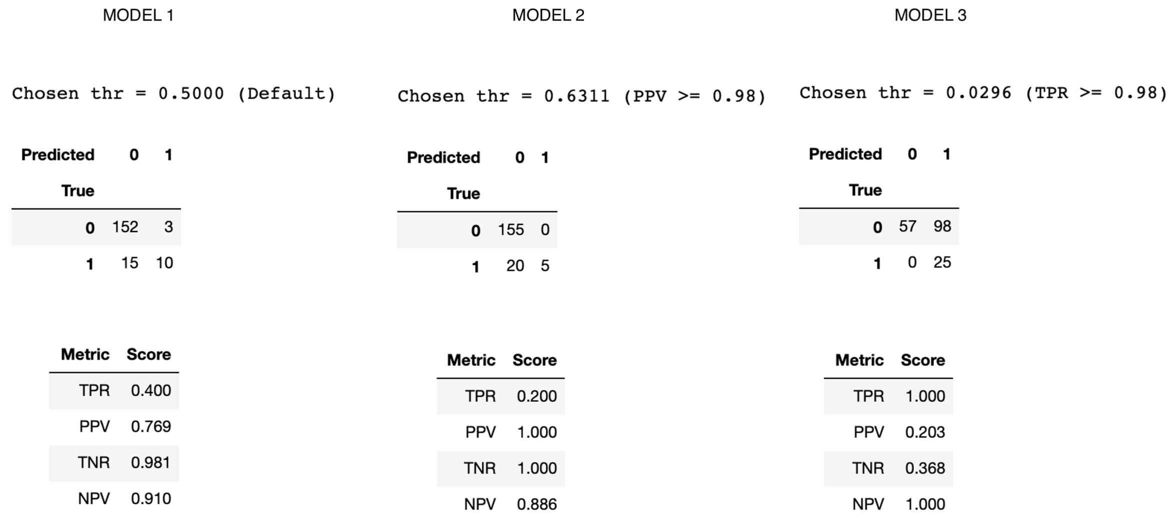
1.5 Short Answer 1c

> *Compare the two models in terms of their ROC curves from Figure 1. Does one dominate the other in terms of overall performance, or are there areas where one model "wins" and others where the other model does? Which model do recommend for the task at hand?*

Answer: The F = 3 model dominates over the F = 2 model for most of the FPR values except maybe for FPR = 0 and after FPR = 0.82, where both models perform almost identically. The model I would recommend is the F = 3 model because its true positive rate in function of the false positive rate is higher than the one for the F = 2 model.

1.6 Figure 2

Confusion matrices and performance metrics on the test set across several thresholds.



MODEL 1

```
Chosen thr = 0.5000 (Default)
```

Predicted	0	1
True		
0	152	3
1	15	10

Metric	Score
TPR	0.400
PPV	0.769
TNR	0.981
NPV	0.910

MODEL 2

```
Chosen thr = 0.6311 (PPV >= 0.98)
```

Predicted	0	1
True		
0	155	0
1	20	5

Metric	Score
TPR	0.200
PPV	1.000
TNR	1.000
NPV	0.886

MODEL 3

```
Chosen thr = 0.0296 (TPR >= 0.98)
```

Predicted	0	1
True		
0	57	98
1	0	25

Metric	Score
TPR	1.000
PPV	0.203
TNR	0.368
NPV	1.000

Figure 2: Summary of the test-set performance of the F=3 logistic regression model across several thresholds.

1.7 Short Answer 1d

> *Compare the 3 confusion matrices in Figure 2. Which model and thresholding strategy best meets our preferences from 1b: avoid life-threatening mistakes at all costs, while also eliminating unnecessary biopsies?*

Answer: To avoid life-threatening mistakes I need our model to avoid false negatives. A good measure of this phenomena is the NPV, which is the probability that a subject is who is called negative will actually be negative. So I want this rate to be as high as possible, and the model that has this property is the third model, the one with thr = 0.0296 ($TPR \geq 0.98$), which has NPV = 1.000.

To avoid unnecessary biopsies, I need our model to avoid false positives. A good measure of this phenomena is the PPV, which is the probability that a subject who is called positive will actually be positive. I want this rate to be as high as possible, and the model that has this property is the second model with thr = 0.6311 ($PPV \leq 0.98$) which has PPV = 1.000.

Since there is no model among the three we have that supports the constraint PPV = NPV = 1.000, it means that I'll have choose a model that has one or both of these rates slightly lower, but which still has a good trade-off of FPs and FNs. I would NOT choose the third model since its PPV is very low, thus there would be many unnecessary biopsies if I would use this model for cancer prediction. Personally, I would choose the first model over the second one because the NPV of the first one (NPV = 0.910) is higher than the NPV of the second one (NPV = 0.886), while the first one also keeps a quite high PPV. I would rather avoid deaths, than avoiding unnecessary biopsies.

Model 1 is the Logistic Regression with all three features (F = 3), using threshold 0.5.

1.8 Short Answer 1e

> *By carefully reading the confusion matrices from Figure 2, estimate how many subjects in the test set are saved from unnecessary biopsies using your selected thresholding strategy. What fraction of current biopsies might be avoided if this classifier was adopted by the hospital?*

Answer: There are 152 subjects that are saved from unnecessary biopsies (the number of true negatives). Assuming that there is a biopsy when the model presents a FP and a TP, the fraction of current biopsies that might be avoided if this classifier was adopted by the hospital is $\frac{FP}{(TP+FP)}$, which is the same as $1 - PPV = 1 - 0.769 = 0.231$.