Chapter 8

Classification Performance Metrics

Goal: Quantify the performance of a binary classifier on a test dataset.

Definitions:
- $\mathbf{y}$ - target values vector of the test database, $\mathbf{y} \in \mathbb{R}^M$
- $\hat{\mathbf{y}}$ - predicted value, $\hat{\mathbf{y}} \in \mathbb{R}^M$, output of some classifier $\hat{\mathbf{y}} = f_{\boldsymbol{\theta}}(\mathbf{X})$.

Typically, in binary classification, $y_i \in \{0, 1\}$.

8.1 Definitions

Goal: Classification between two groups (only).

Basic terminology:
- '1' – positive group or result
- '0' – negative group or result
- Y – actual class
- $\hat{Y}$ – predicted class

Positive/negative terminology is rather arbitrary. Typically, the result of interest is termed positive.

8.2 Confusion matrix

Goal: Summarize classification results of a test set of a particular database.

The summarization is in the form of a 2D non-normalized histogram of $(Y, \hat{Y})$.
The (test) database has M values, among them:
- TP + FN positive values
- FP + TN negative values

This is the most common way to summarize the performance of a particular classifier on a particular dataset. It can be easily extended for multi-class classifiers.

8.3 Performance Characteristics

Goal: Characterization is useful to compare classifiers and/or performance on different datasets.

8.3.1 Accuracy

Goal: The most intuitive metric, fraction of $Y = \hat{Y}$, among all the classification results, $\Pr\left(Y = \hat{Y}\right)$.

		Predicted values: **Positive**, $\hat{Y} = 1$	**Predicted** values: **Negative**, $\hat{Y} = 0$
Actual values	**Positive**, $Y = 1$	**TP** True Positive $Y = 1, \hat{Y} = 1$	**FN** False Negative $Y = 1, \hat{Y} = 0$
	Negative, $Y = 0$	**FP** False Positive $Y = 0, \hat{Y} = 1$	**TN** True Negative $Y = 0, \hat{Y} = 0$

Figure 8.1: Confusion matrix. Note, sometimes, transposed representation is used.

$$\begin{aligned} \text{Accuracy} &= \frac{\text{correct predictions}}{\text{total predictions}} \\ &= \frac{TP + TN}{TP + NT + FP + FN} \end{aligned} \tag{8.1}$$

Example 8.1: Covid antibody (fast non-PCR) test performance. The example includes test statistics of 239 participants [3], as presented below.

		Predicted: Yes	Predicted: No
Actual	Yes	141	67
	No	0	31

The resulting accuracy is

$$\text{Accuracy} = \frac{141 + 31}{239} = 0.7196652 \approx 72.0\% \tag{8.2}$$

Term	Radar Interpretation
Accuracy	Percentage of all correctly identified as planes or not planes
Precision	Among all classified as planes, the portion that is correctly classified as planes
Recall sensitivity	Among all existing planes, portion of correctly classified as planes
Specificity	Among all classified as non-planes, portion of correctly classified as non-planes

Table 8.1: Radar interpretation of the classification metrics.

In the example, FN=67 is a bad performance, and FP=0 is probably something good. However, accuracy does not reflect the discrepancy between these two. Additional metrics are used to quantify these aspects.

8.3.2 Precision

Goal: Proportion of positive classification that is actually correct, $\Pr\left(Y=1|\hat{Y}=1\right)$.

From the probability theory,

$$\Pr\left(Y=1|\hat{Y}=1\right)=\frac{\Pr\left(Y=1,\hat{Y}=1\right)}{\Pr(Y=1)} \tag{8.3}$$

$$\Pr(Y=1)=\Pr\left(Y=1,\hat{Y}=0\right)+\Pr\left(Y=1,\hat{Y}=1\right) \tag{8.4}$$

$$\text{Precision}=\frac{TP}{FP+TP}=\frac{\text{Correctly predicted 1's}}{\text{All predicted 1's}} \tag{8.5}$$

TP = Correctly predicted 1's
TP + FP = All predicted 1's

Example 8.1: Back to the previous example,

$$\text{Precision}=\frac{141}{0+141}=1=100\% \tag{8.6}$$

The high value of the precision is due to the low FP. From the medical point of view, all positive results are actually positive. Whoever was identified by this test as Covid-positive is really positive.

8.3.3 Recall (sensitivity)

Goal: Proportion of positives identified correctly, $\Pr\left(\hat{Y}=1|Y=1\right)$.

From the probability theory,

$$\Pr\left(\hat{Y}=1|Y=1\right)=\frac{\Pr\left(Y=1,\hat{Y}=1\right)}{\Pr\left(\hat{Y}=1\right)} \tag{8.7}$$

$$\Pr\left(\hat{Y}=1\right)=\Pr\left(\hat{Y}=1,Y=0\right)+\Pr\left(\hat{Y}=1,Y=1\right) \tag{8.8}$$

$$\text{Recall}=\frac{TP}{TP+FN}=\frac{\text{Correctly predicted 1's}}{\text{Actual 1's}} \tag{8.9}$$

Medical meaning: portion of correctly classified ill among all the ill.

Example 8.1: Back to the previous example,

$$\text{Recall}=\frac{141}{141+67}=0.678=67.8\% \tag{8.10}$$

The low value of the recall is due to the high FN. From the medical point of view, among all the positive results, only 67.8% are actually positive.

8.3.4 Specificity

Goal: Proportion of negatives identified correctly, $\Pr\left(\hat{Y}=0|Y=0\right)$.

$$\text{Specificity}=\frac{TN}{FP+TN}=\frac{\text{Correctly predicted 0's}}{\text{Actual 0's}} \tag{8.11}$$

Medical meaning: portion of classified healthy among all the healthy.

Example 8.1: Back to the previous example,

$$\text{Specificity}=\frac{31}{0+31}=1=100\% \tag{8.12}$$

From the medical point of view, all negative results are really negative.

8.3.5 F_1-score

Goal: Combination of precision and recall.

The harmonic mean between precision and recall,

$$F_1=\frac{2}{\frac{1}{recall}+\frac{1}{precision}}=\frac{TP}{TP+\frac{1}{2}\left(FP+FN\right)} \tag{8.13}$$

Example 8.1: Back to the previous example,

$$F_1=\frac{141}{141+\frac{1}{2}\left(0+67\right)}=0.808=80.8\% \tag{8.14}$$

8.4 Imbalanced Dataset

Imbalanced dataset: Dataset with significant differences between the numbers of labels of each class. The following examples present a few problems related to imbalanced datasets.

Example 8.2: Let's take a dataset with 1000 samples:

- 990 samples labeled '0'

- 10 samples labeled '1'

What are the performance metrics of the classifier that always predicts $\hat{Y} = 0$?

Solution: The resulting confusion matrix is

		Predicted: Yes	Predicted: No
Actual	Yes	0	10
	No	0	990

and the resulting quantities are

$$\begin{aligned}
\text{Accuracy} &= \frac{990}{1000} = 0.99 = 99\% \\
\text{Precision} &= \frac{TP}{FP+TP} = \frac{0}{0+0} = \text{Undefined} \\
\text{Recall} &= \frac{TP}{TP+FN} = \frac{0}{0+10} = 0 \\
\text{Specificity} &= \frac{TN}{FN+TN} = \frac{990}{1000} = 0.99 = 99\% \\
F_1 &= \frac{TP}{TP+\frac{1}{2}(FP+FN)} = \frac{0}{\cdots} = 0
\end{aligned} \tag{8.15}$$

- Note, accuracy is insufficient metrics!
- Note, while the convention is to label '1' for the most important class outcome, sometimes it is interchangeable.

Goal: Small dataset problem.

Example 8.3: We have the dataset from the previous example. This time, let's assume that the theoretical performance of the classifier on class '1' is $p = 0.8$. What is the probability that the classifier will classify only 6 samples or less correctly, from 10 measurements?

Solution: The probabilities follow the distribution. $X \sim Bin(n = 10, p = 0.8)$ with a question $\Pr(X \leq 6) =?$. The numerical solution is

$$\Pr(X \leq 6) = \Pr(X = 0) + \cdots + \Pr(X = 6) \approx 12.09\%$$

Moreover, $\Pr(X = 10) = 10.74\%$.

The analysis of the issue illustrated in the example is called *confidence* analysis. While the discussion of confidence intervals is out of the scope of this document, this example emphasizes the problem of a small dataset, particularly in imbalanced data.

Anomaly detection: Sub-field of imbalanced problem: anomaly detection.

8.5 Decision threshold

8.5.1 Receiver Operating Characteristics (RoC)

With the probabilistic loss function, the classifier output is the probability of

$$\Pr(\hat{y} = 1) = f_{\boldsymbol{\theta}}(x). \tag{8.16}$$

The binary decision for $\hat{\mathbf{y}}$ is to compare $f_{\boldsymbol{\theta}}(x)$ with some predefined threshold,

$$\hat{\mathbf{y}} = \begin{cases} 1 & f_{\boldsymbol{\theta}}(x) \geq \mathsf{thr} \\ 0 & f_{\boldsymbol{\theta}}(x) < \mathsf{thr} \end{cases} \tag{8.17}$$

with the default value of $\mathsf{thr} = 0.5$. The change of thr may significantly influence the resulting confusion matrix.

Goal: To quantify the trade-off between confusion matrix elements as a function of thr. The used quantities are:

- True Positive Rate (**TPR**) is a synonym for recall.
- False Positive Rate (**FPR**) is defined by

$$FPR = \frac{FP}{FP+TN} = 1 - \text{specificity} \tag{8.18}$$

RoC is a legacy term from the field of detector theory and communication system theory.

8.5.2 Area under curve (AUC)

Goal: Quantify threshold-independent performance.

AUC: AUC is the area under the RoC curve.
Range: Borderline cases are a coin toss with $\mathsf{AUC} = 0.5$ and an ideal classifier with $\mathsf{AUC} = 1$. All other classifiers fall in the range $0.5 \leq \mathsf{AUC} \leq 1$.

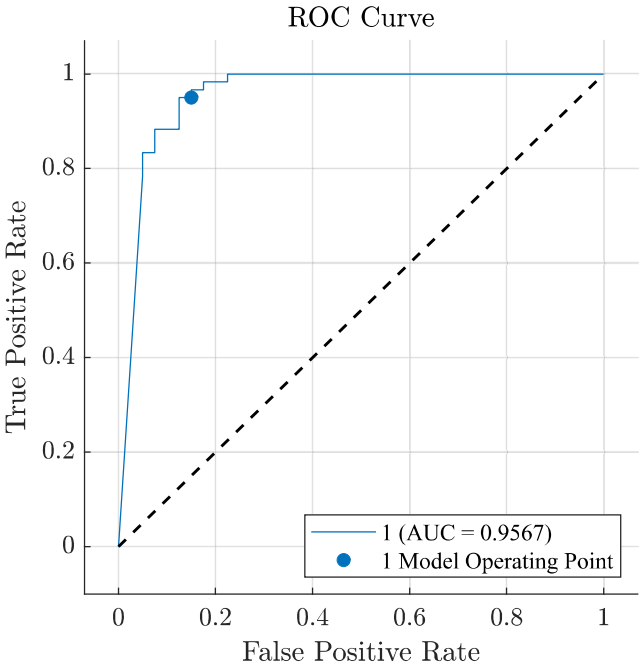

Figure 8.2: RoC of logistic regression example in Fig. 7.1. The model operation point is $\mathsf{thr} = 0.5$.

Properties:

- AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.
- AUC is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

- Scale invariance is not always desirable for a performance assessment.
- Classification-threshold invariance is not always desirable. Sometimes some trade-off between false negatives vs. false positives is required. For example, when doing email spam detection, you likely want to prioritize minimizing false positives (even if that results in a significant increase of false negatives). AUC isn't a useful metric for this type of optimization.