Master Program in *Data Science and Business Informatics*

# Statistics for Data Science

## Lesson 30 - Classifier performances in R

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

# Tests and confidence intervals for classifier performance

## The Caret package

```
Define sets of model parameter values to evaluate
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples
        [Optional] Pre–process the data
        Fit the model on the remainder
        Predict the hold–out samples
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

For resampling methods, see Lesson 28

**See R script**

# Binary classifier performance metrics

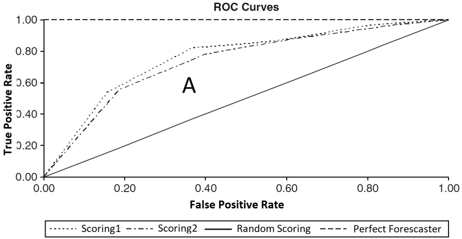
**Confusion matrix** (in R packages, it is transposed)

| | | **Predicted condition** | | | |
|---|---|---|---|---|---|
| | Total population $= \mathrm{P} + \mathrm{N}$ | **Positive (PP)** | **Negative (PN)** | Informedness, bookmaker informedness (BM) $= \mathrm{TPR} + \mathrm{TNR} - 1$ | Prevalence threshold (PT) $= \frac{\sqrt{\mathrm{TPR} \times \mathrm{FPR}} - \mathrm{FPR}}{\mathrm{TPR} - \mathrm{FPR}}$ |
| **Actual condition** | **Positive (P)** | **True positive (TP), hit** | **False negative (FN), type II error, miss, underestimation** | True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\mathrm{TP}}{\mathrm{P}} = 1 - \mathrm{FNR}$ | False negative rate (FNR), miss rate $= \frac{\mathrm{FN}}{\mathrm{P}} = 1 - \mathrm{TPR}$ |
| | **Negative (N)** | **False positive (FP), type I error, false alarm, overestimation** | **True negative (TN), correct rejection** | False positive rate (FPR), probability of false alarm, fall-out $= \frac{\mathrm{FP}}{\mathrm{N}} = 1 - \mathrm{TNR}$ | True negative rate (TNR), specificity (SPC), selectivity $= \frac{\mathrm{TN}}{\mathrm{N}} = 1 - \mathrm{FPR}$ |
| | Prevalence $= \frac{\mathrm{P}}{\mathrm{P} + \mathrm{N}}$ | Positive predictive value (PPV), precision $= \frac{\mathrm{TP}}{\mathrm{PP}} = 1 - \mathrm{FDR}$ | False omission rate (FOR) $= \frac{\mathrm{FN}}{\mathrm{PN}} = 1 - \mathrm{NPV}$ | Positive likelihood ratio (LR+) $= \frac{\mathrm{TPR}}{\mathrm{FPR}}$ | Negative likelihood ratio (LR−) $= \frac{\mathrm{FNR}}{\mathrm{TNR}}$ |
| | Accuracy (ACC) $= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$ | False discovery rate (FDR) $= \frac{\mathrm{FP}}{\mathrm{PP}} = 1 - \mathrm{PPV}$ | Negative predictive value (NPV) $= \frac{\mathrm{TN}}{\mathrm{PN}} = 1 - \mathrm{FOR}$ | Markedness (MK), deltaP (Δp) $= \mathrm{PPV} + \mathrm{NPV} - 1$ | Diagnostic odds ratio (DOR) $= \frac{\mathrm{LR+}}{\mathrm{LR-}}$ |
| | Balanced accuracy (BA) $= \frac{\mathrm{TPR} + \mathrm{TNR}}{2}$ | $\mathrm{F}_1$ score $= \frac{2\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$ | Fowlkes–Mallows index (FM) $= \sqrt{\mathrm{PPV} \times \mathrm{TPR}}$ | Matthews correlation coefficient (MCC) $= \sqrt{\mathrm{TPR} \times \mathrm{TNR} \times \mathrm{PPV} \times \mathrm{NPV}} - \sqrt{\mathrm{FNR} \times \mathrm{FPR} \times \mathrm{FOR} \times \mathrm{FDR}}$ | Threat score (TS), critical success index (CSI), Jaccard index $= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$ |

Metrics computed on a test set are intended to estimate some parameter over the general distribution.

- $X = (W, C) \sim F$, i.e., $F$ is the (unknown) multivariate distribution of predictive features and class
- Accuracy $ACC$ of a classifier $y_\theta^+$ is a point estimate of $E_F[\mathbb{1}_{y_\theta^+(W)=C}] = P_F(y_\theta^+(W) = C)$

# Probabilistic binary classifier performance metrics




- Binary classifier score $s_\theta(w) \in [0,1]$ where $s_\theta(w)$ estimates $\eta(w) = P_{\theta_{TRUE}}(C=1|W=w)$
- ROC Curve *[Cfr. also Lesson 16]*
  - $TPR(p) = P(s_\theta(w) \geq p | C=1)$ and $FPR(p) = P(s_\theta(w) | C=0)$
  - ROC Curve is the scatter plot $TPR(p)$ over $FPR(p)$ for $p$ ranging from 1 down to 0
  - AUC-ROC is the area below the curve **What does AUC-ROC estimate?**
- Squared error loss or $L_2$ loss or Brier score: $\frac{1}{n}\sum_i (s_\theta(w_i) - c_i)^2$
- Classifier is calibrated if $P(C=1|s_\theta(w)=p)=p$ **classifier-calibration.github.io**
  - Binary Expected Calibration Error (binary-ECE): $\sum_b \frac{|B_b|}{n} |Y_b - S_b|$
    - $B_b$ is the set of $i$'s in the $b^{th}$ bin, $Y_b = |\{i| \ i \in B_b, c_i = 1\}|/|B_b|$ , $S_b = (\sum_{i \in B_b} s_\theta(w_i))/|B_b|$