UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Lesson 7.

Performance metrics

DATA SCIENCE AND
AUTOMATION COURSE

MASTER DEGREE SMART
TECHNOLOGY ENGINEERING

TEACHER
Mirko Mazzoleni

PLACE
University of Bergamo

Outline

1. Metrics

2. Precision and recall

3. Receiver Operating Characteristic (ROC) curves

Outline

1. **Metrics**
2. Precision and recall
3. Receiver Operating Characteristic (ROC) curves

Metrics

It is extremely important to use **quantitative metrics** for evaluating a machine learning model

- Until now, we relied on the **cost function value** for regression and classification

- Other metrics can be used to **better evaluate** and understand the model

- **For classification**
  - ✓ Accuracy/Precision/Recall/F1-score, ROC curves,…
- **For regression**
  - ✓ Normalized RMSE, Normalized Mean Absolute Error (NMAE),…

Classification case: metrics for skewed classes

Disease dichotomic classification example

Train logistic regression model $h(\boldsymbol{x})$, with $y = 1$ if disease, $y = 0$ otherwise.

Find that you got 1% error on test set (99% correct diagnoses)

Only 0.50% of patients **actually have** disease

The $y = 1$ class has very few examples with respect to the $y = 0$ class

If I use a predictor that **predicts always the $\mathbf{0}$ class**, I get 99.5% of accuracy!!

For **skewed classes,** the accuracy metric can be deceptive

Outline

1. Metrics

2. **Precision and recall**

3. Receiver Operating Characteristic (ROC) curves

Precision and recall

Suppose that $y = 1$ in presence of a **rare class** that we want to detect

Precision *(How much we are precise in the detection)*

Of all patients where we predicted $y = 1$, what fraction actually has the disease?

$$\frac{\text{True Positive}}{\#\ \text{Predicted Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall *(How much we are good at detecting)*

Of all patients that actually have the disease, what fraction did we correctly detect as having the disease?

$$\frac{\text{True Positive}}{\#\ \text{Actual Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Confusion matrix

Actual class

Predicted class	1 (p)	0 (n)
1 (Y)	True positive (TP)	False positive (FP)
0 (N)	False negative (FN)	True negative (TN)

Trading off precision and recall

Logistic regression: $0 \leq h(\boldsymbol{x}) \leq 1$

- Predict 1 if $h(\boldsymbol{x}) \geq 0.5$
- Predict 0 if $h(\boldsymbol{x}) < 0.5$

These thresholds can be different from 0.5!

→ *At different thresholds, correspond different confusion matrices!*

Suppose we want to predict $y = 1$ (disease) only if very confident

- Increase threshold → Higher precision, lower recall

Suppose we want to avoid missing too many cases of disease (avoid false negatives).

- Decrease threshold → Higher recall, lower precision

F1-score

It is usually better to compare models by means of one number only. The $\mathrm{F1-score}$ can be used to combine precision and recall

	Precision(P)	Recall (R)	Average	F_1 Score
Algorithm 1	0.5	0.4	0.45	0.444
Algorithm 2	0.7	0.1	0.4	0.175
Algorithm 3	0.02	1.0	0.51	0.0392

The best is Algorithm 1

Algorithm 3 predict always 1

Average says not correctly that Algorithm 3 is the best

$$\mathrm{Average} = \frac{\mathrm{P} + \mathrm{R}}{2} \qquad \mathrm{F_1 score} = 2\frac{\mathrm{PR}}{\mathrm{P} + \mathrm{R}}$$

- $\mathrm{P} = 0$ or $\mathrm{R} = 0 \Rightarrow \mathrm{F_1 score} = 0$
- $\mathrm{P} = 1$ and $\mathrm{R} = 1 \Rightarrow \mathrm{F_1 score} = 1$

Summaries of the confusion matrix

Different metrics can be computed from the confusion matrix, depending on the class of interest *(https://en.wikipedia.org/wiki/Precision_and_recall)*

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive,
Power</td><td>False positive,
Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative,
Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td colspan="2"></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $\frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$</td></tr>
<tr><td colspan="2"></td><td>False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Outline

1. Metrics

2. Precision and recall

3. **Receiver Operating Characteristic (ROC) curves**

Ranking instead of classifying

Classifiers such as logistic regression can output a **probability** of belonging to a class (or something similar).

- We can use this to **rank** the different istances and take actions on the cases at top of the list

- We may have a **budget**, so we have to target most promising individuals

- Ranking enables to use different techniques for **visualizing** model performance

Ranking instead of classifying

Instance description	True class	Score
……………	1	0,99
……………	1	0,98
……………	0	0,96
……………	0	0,90
……………	1	0,88
……………	1	0,87
……………	0	0,85
……………	1	0,80
……………	0	0,70

	p	n
Y	0	0
N	100	100

	p	n
Y	1	0
N	99	100

	p	n
Y	2	0
N	98	100

	p	n
Y	2	1
N	98	99

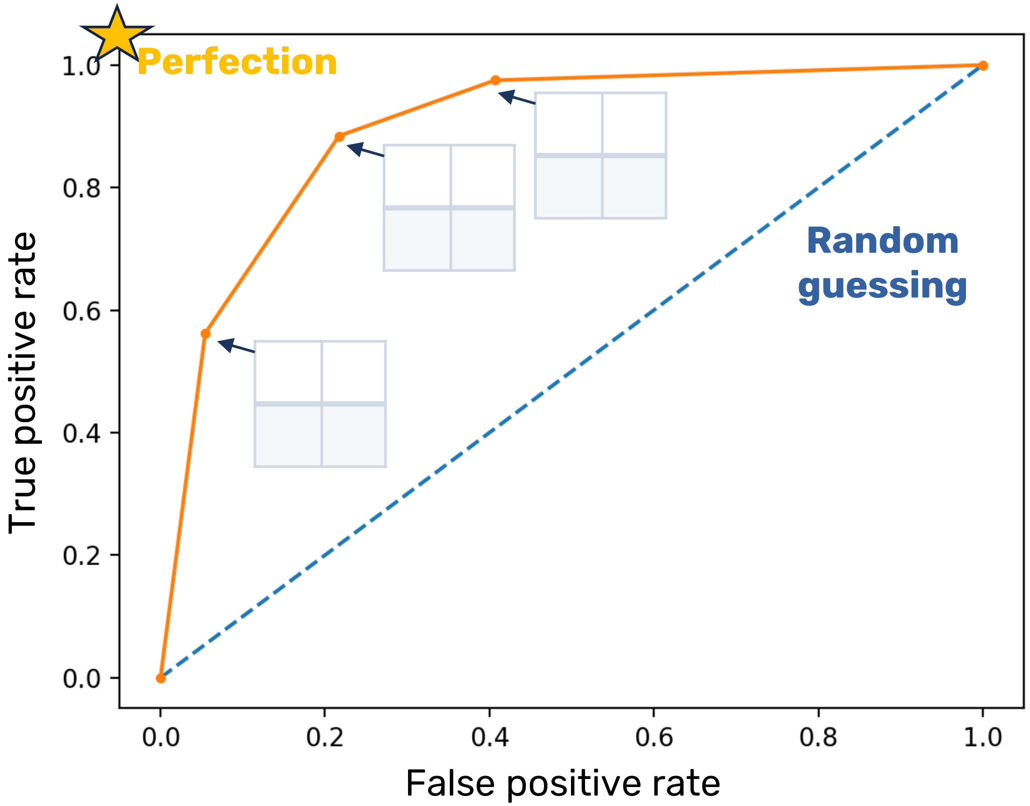
	p	n
Y	6	4
N	94	96

Different confusion matrices by changing the **threshold**

Adapated from [1]

ROC curves

ROC curves are a very general way to **represent and compare** the performance of different models (on a binary classification task)



Observations

- $(0,0)$: predict always negative
- $(1,1)$: predict always positive
- Diagonal line: random classifier
- Below diagonal line: worse than random classifier
- Different classifiers can be compared
- Area Under the Curve (AUC): probability that a randomly chosen positive instance will be ranked ahead of randomly chosen negative instance