Receiver-Operator Characteristic

Workshop

Norman Juchler

Introduction

Machine learning terminology in 1 minute

Examples:
x: Shape features
y: Rupture status
Survival / recovery
f: Logistic regression
Threshold classifier

Model or classifier

Model parameters

$$y = f(x|\theta)$$

Response or labels

Predictors or features

$$x \in \mathbb{R}^d$$

Problem dimensionality
d = 1: univariate
d > 1: multivariate

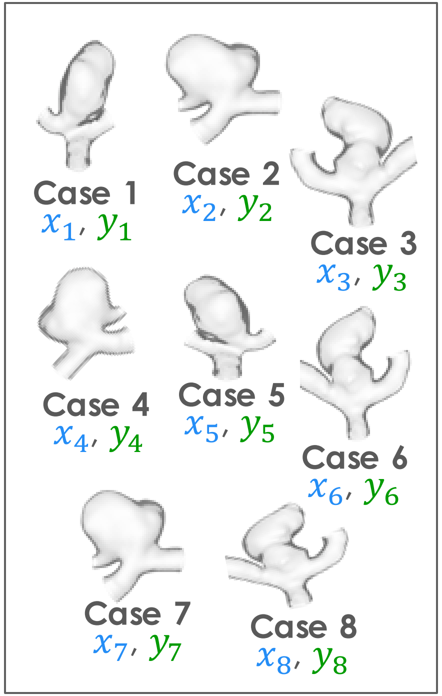
$$\theta \in \mathbb{R}^p$$

$$y \in \mathbb{R}$$

Defines problem type:
Binary classification

Supervised learning: Find optimal parameters θ^* given the training data x_t, y_t
Testing/validation: Compare predictions $y_p = f(x_v|\theta^*)$ with true response y_v

Training and validation scheme

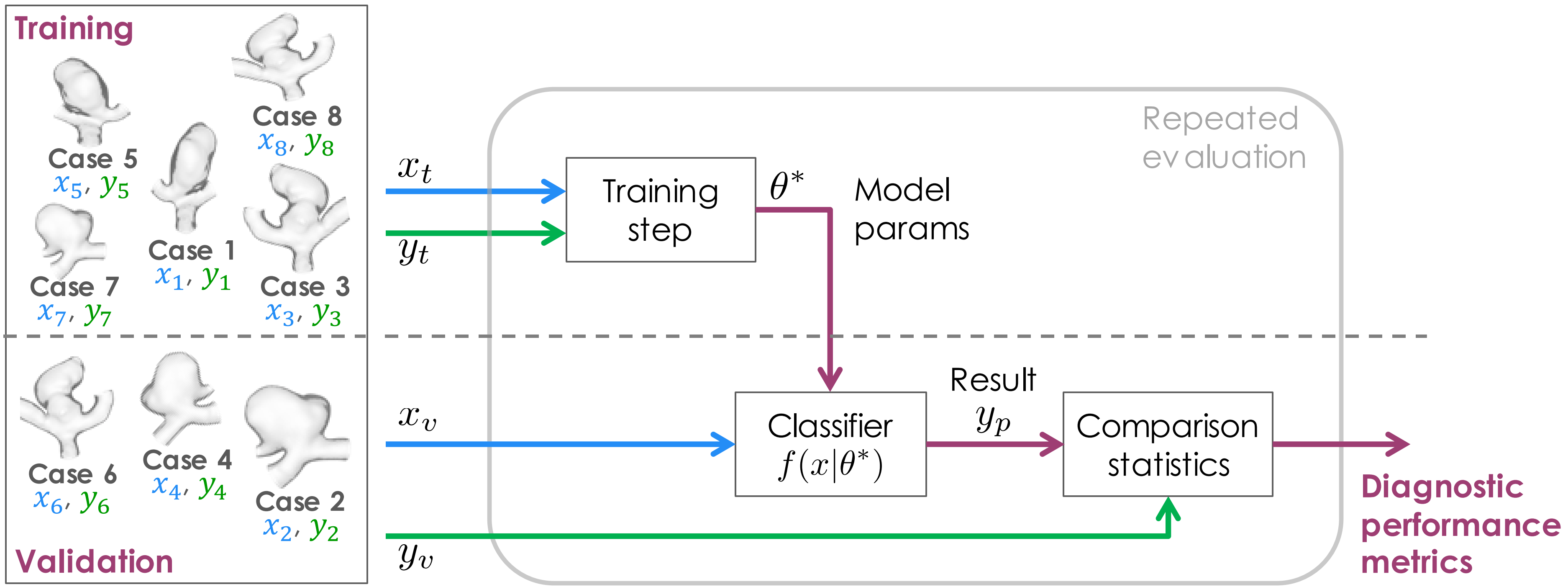


AneuX morphology database (n=750)

x_i: Shape features (aneurysm size, non-sphericity, …)

y_i: Rupture status
0: unruptured
1: ruptured

Training and validation scheme: The benchmark



x_i: Shape features

y_i: Rupture status (binary)

Metrics of diagnostic/predictive accuracy

- **Accuracy**

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}$$

- **Sensitivity** (true positive rate, TPR, **recall**)

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}=\frac{\mathrm{TP}}{\mathrm{P}}$$

- **Specificity** (true negative rate, TNR)

$$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}=\frac{\mathrm{TN}}{\mathrm{N}}$$

- **Precision** (positive predictive value, PPV)

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

		True condition ( )	
	Total population	**P** Condition positive	**N** Condition negative
Prediction (^)	Predicted condition positive	**TP** True positive	**FP** False positive Type I error
	Predicted condition negative	**FN** False negative Type II error	**TN** True negative

Contingency table
(aka confusion matrix)

Metrics of diagnostic/predictive accuracy

- **Accuracy**

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}$$

- **Sensitivity** (true positive rate, TPR, **recall**)

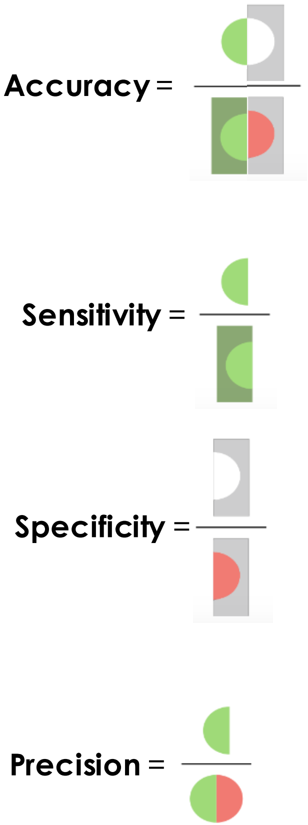
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$$

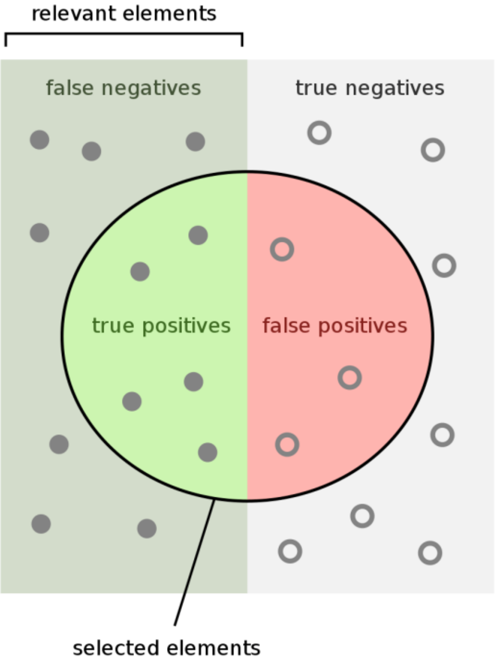
- **Specificity** (true negative rate, TNR)

$$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = \frac{\mathrm{TN}}{\mathrm{N}}$$

- **Precision** (positive predictive value, PPV)

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$





How to handle data imbalance?

- Imbalance: strong difference in class sizes
- Example:
  - Number of healthy patients: 105’056
  - Number of sick patients: 135
- Pitfalls:
  - Misguiding the training objective
  - Optimistic reporting of the diagnostic ability of a model

Dummy/degenerate classifier:

Assign all samples to large class.

- Accuracy: 0.999
- Sensitivity: 1.0
- Specificity: 0.0

Solution:

- Use metrics that are more robust to imbalanced data
- Use more than one metric

Examples:

- ROC-AUC (Area under ROC curve)
- PR-AUC (Area under the Precision-Recall curve)
- Half-class accuracy
- Cohen’s Kappa

Half-class accuracy: $\frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{P}}+\frac{\mathrm{TN}}{\mathrm{N}}\right)$

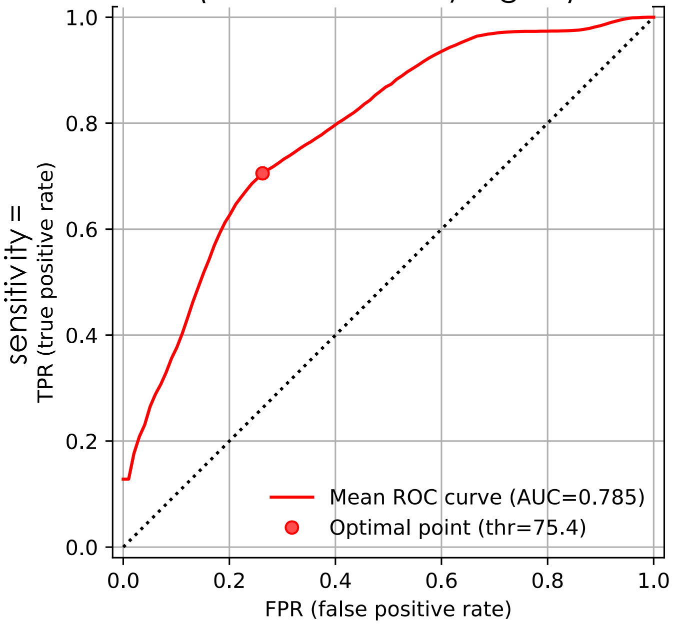
Average of sensitivity and specificity

Receiver-Operating Characteristics (ROC) analysis

- Method to assess the diagnostic/discriminative ability of a **binary classifier** $\hat{y} = f\left(X|\theta\right)$
- Idea: Compute specificity and sensitivity for varying θ
  ROC curve is parametrized by θ

- **Area under ROC curve (AUC)**
  - Measures how well a model discriminates between two classes
  - AUC=1.0: perfect classifier
    AUC=0.5: random classifier
  - Alternative interpretation: Probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one $P\left(X_{|y=1} > X_{|y=0}\right)$
  - Proof: not too complicated. See for example here.

Exemplary ROC curve
(with uncertainty region)



= (1- specificity)

Receiver-Operating Characteristics (ROC) analysis

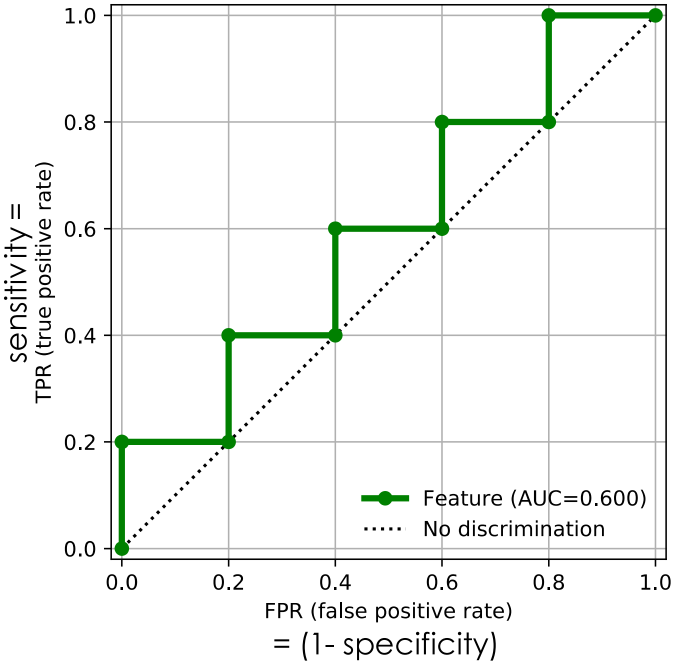
- Method to assess the diagnostic/discriminative ability of a binary classifier $\hat{y} = f\,(X|\theta)$

- Idea: compute specificity and sensitivity for varying θ

- Example: Threshold classifier $\hat{y} = \begin{cases} 0, \text{if } x < \theta \\ 1, \text{if } x \geq \theta \end{cases}$

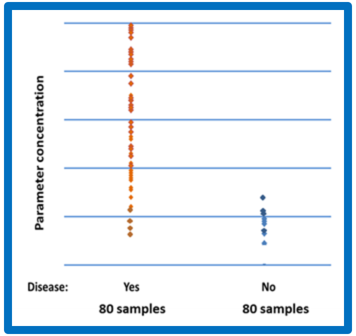
x:	0.1	1.1	2.6	3.3	4.9	5.2	6.5	7.3	8.5	9.3	
y:	0	0	0	0	0	0	1	1	1	1	
$\hat{y}\vert_{\theta=0}$:	1	1	1	1	1	1	1	1	1	1	**A**
$\hat{y}\vert_{\theta=5}$:	0	0	0	0	0	1	1	1	1	1	**B**
$\hat{y}\vert_{\theta=10}$:	0	0	0	0	0	0	0	0	0	0	**C**

Example: Random classifier

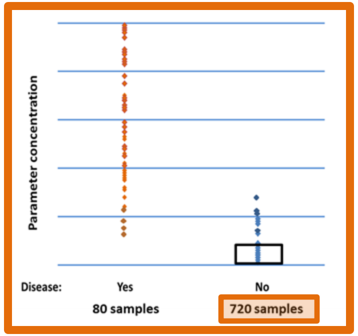


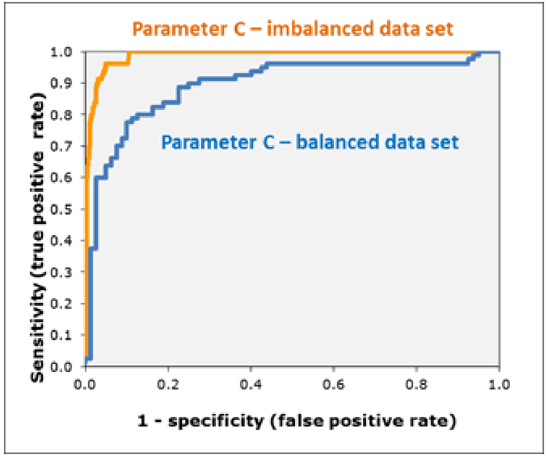
ROC-AUC is not perfectly robust to data imbalance

Balanced

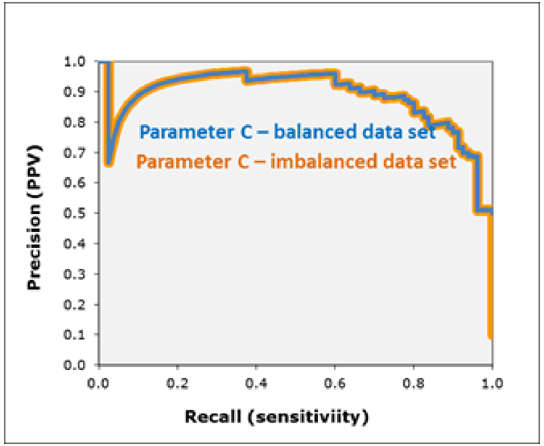


Imbalanced





Solution: report **precision** and **recall**



- **Accuracy**

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}$$

- **Sensitivity** (true positive rate, TPR, **recall**)

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$$

- **Specificity** (true negative rate, TNR)

Reporting guidelines help to write a sound paper

- Resources
  - https://www.equator-network.org/
  - Stuff by Douglas G. Altman

- Relevant in the context of diagnostic tools:
  - STARD: Standards for Reporting Diagnostic Accuracy
  - TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	
Quality improvement studies	SQUIRE	
Economic evaluations	CHEERS	

Complete reporting is crucial!

Univariate models (internal validation, cut *dome*)							
Category	**Predictor**	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Kappa**
Shape	NSI, non-sphericity	**0.80±0.05**	0.73±0.04	0.75±0.08	0.72±0.05	0.50±0.05	0.41±0.08
ZMI	norm. energy Z_6^{surf}	**0.80±0.05**	0.74±0.04	0.75±0.08	0.74±0.06	0.52±0.06	0.43±0.09
ZMI	norm. energy Z_3^{surf}	**0.78±0.04**	0.73±0.04	0.61±0.09	0.78±0.05	0.51±0.06	0.36±0.09
Writhe	$\overline{W}_{mean}^{L_1}$	**0.78±0.04**	0.72±0.04	0.71±0.09	0.72±0.05	0.49±0.05	0.37±0.07
Shape	UI, undulation	**0.77±0.05**	0.74±0.04	0.61±0.10	0.79±0.05	0.52±0.06	0.38±0.09
Curvature	GLN	**0.75±0.05**	0.71±0.04	0.59±0.08	0.76±0.05	0.48±0.06	0.32±0.08
Curvature	MLN	**0.75±0.05**	0.69±0.04	0.63±0.08	0.71±0.05	0.45±0.05	0.31±0.08
Shape	AR, aspect ratio	**0.75±0.05**	0.70±0.04	0.61±0.11	0.74±0.05	0.46±0.05	0.32±0.09
ZMI	$ZMI_{3,1}^{\mathrm{surf}}$	**0.74±0.05**	0.66±0.04	0.71±0.09	0.64±0.06	0.42±0.04	0.29±0.07
ZMI	$ZMI_{5,1}^{\mathrm{surf}}$	**0.72±0.05**	0.66±0.05	0.68±0.09	0.66±0.06	0.43±0.05	0.28±0.09
Writhe	$W_{\mathrm{mean}}^{L_2}$	**0.72±0.05**	0.70±0.04	0.58±0.10	0.74±0.05	0.46±0.06	0.30±0.09
Size	aSz	**0.64±0.05**	0.65±0.04	0.46±0.10	0.72±0.06	0.38±0.06	0.16±0.09

- **Accuracy**

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}$$

- **Sensitivity** (true positive rate, TPR)

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$$

- **Specificity** (true negative rate, TNR)

$$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = \frac{\mathrm{TN}}{\mathrm{N}}$$

- **Precision** (positive predictive value, PPV)

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$