ML 101 hands-on

daniel.c.moura@gmail.com | linkedin.com/in/dmoura

Apr 2018

Machine learning

- Learn from data
- Logic extracted from data
- Logic not provided by the programmer

Machine learning

- **Supervised Learning**
  - Teaching by example
- Unsupervised learning
  - Extracting structure from the data

Supervised Learning

- **f(x) → y**
- *f:* the function we want to learn
- *x*: inputs
- *y*: output

Supervised Learning

f(age, weight, height, gender) → accepted?

age	weight	height	gender	accepted?
20	72	182	male	yes
25	80	160	female	no
22	75	170	male	yes
21	70	185	female	?

Regression vs Classification

- *y* is continuous → **Regression**
- *y* is discrete → Classification
  - *y* can take 2 values → **Binary classification**
  - *y* can take 3+ values → Multi-class classification
    - if order of classes matters → Ordinal classification

Regression vs Classification

WEKA

Designing Experiments

- Three sets
  - **Training**: for building the model
  - **Validation**: for choosing / tuning the model
  - **Test**: for evaluating the model

Splitting the Data

- Holdout
  - e.g. 60%, 20%, 20%
- Repeated sampling
- Cross-validation
  - e.g. 10 fold cross-validation, leave-one-out

Splitting the Data

WEKA

Evaluation - Regression

- Correlation
- Mean absolute error
- (Root) Mean squared error

Visualising regression errors

WEKA

Evaluation - Classification

- Confusion matrix

	Predicted **NO**	Predicted **YES**
Actual **NO**	TN	FP
Actual **YES**	FN	TP

Confusion Matrix Measures

- Accuracy: *how often is the prediction correct?*

accuracy (ACC)

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

- Recall: *when it is actually yes, how often is the prediction yes?*

sensitivity, recall, hit rate, or true positive rate (TPR)

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

- Precision: *when the prediction is yes, how often is it correct?*

precision or positive predictive value (PPV)

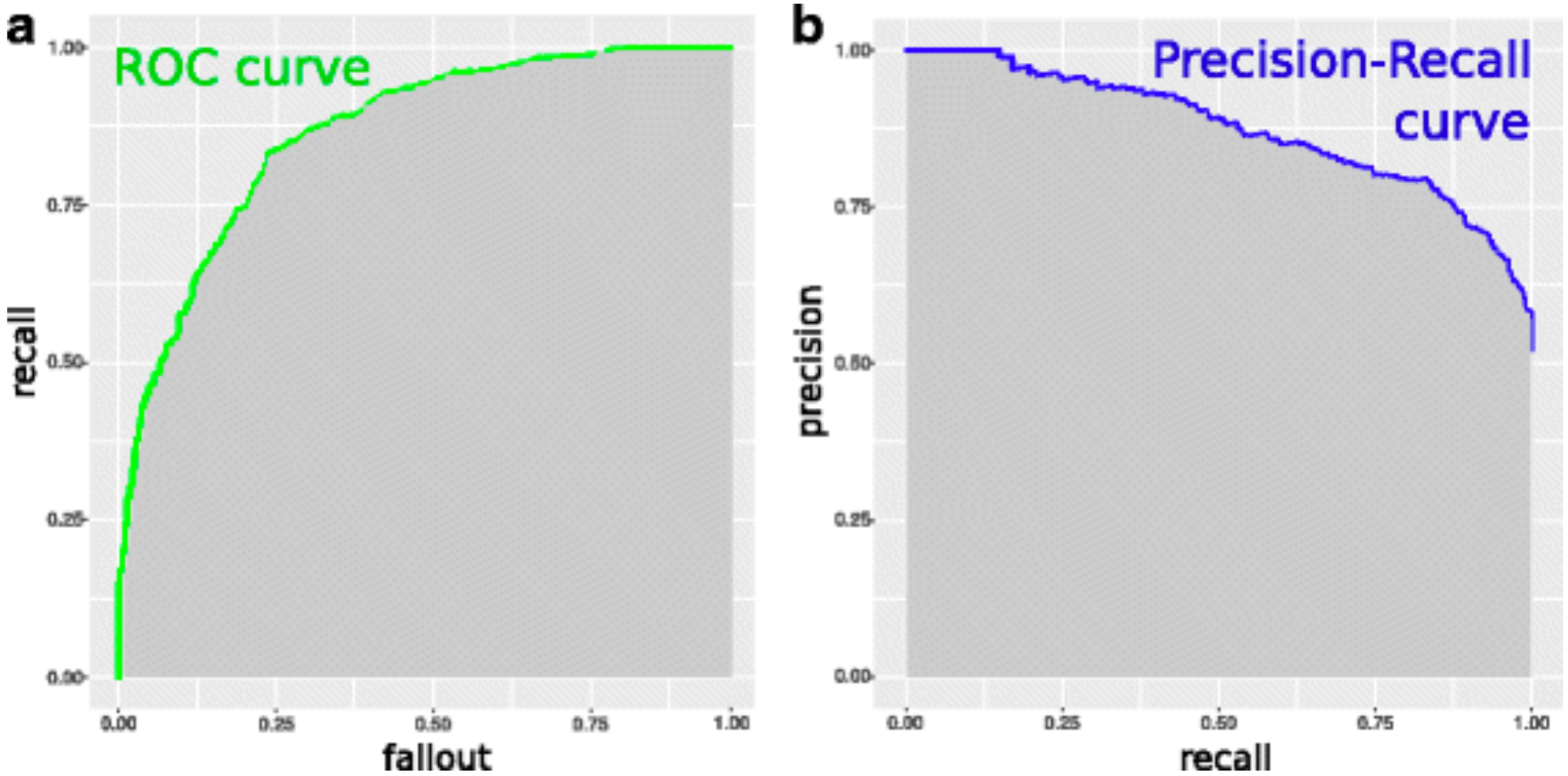
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- F1 score: harmonic mean of precision and recall

- MCC: correlation between actual and predicted

Source: https://en.wikipedia.org/wiki/Confusion_matrix

AUC measures



Source: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3

Evaluation

WEKA

Problems & Pitfalls

- Feature scaling
- Non-linear transformations
- Dimensionality reduction
- Categorical variables
- Radial variables
- Missing data
- Outliers
- Overfitting
- Interpolation
- Extrapolation
- Hyperparameters selection
- Algorithm selection
- Imbalanced classes
- Cost/benefit analysis

Automating Experiments

WEKA

Further reading

- Ten quick tips for machine learning in computational biology, BioData Mining 2017 Dec 8, 10:35, https://doi.org/10.1186/s13040-017-0155-3

- Data Mining: Practical Machine Learning Tools and Techniques, https://www.cs.waikato.ac.nz/ml/weka/book.html

Thanks!

daniel.c.moura@gmail.com