Machine Learning with MATLAB

Kirsty van Ryneveld
MathWorks Consulting

What you will learn

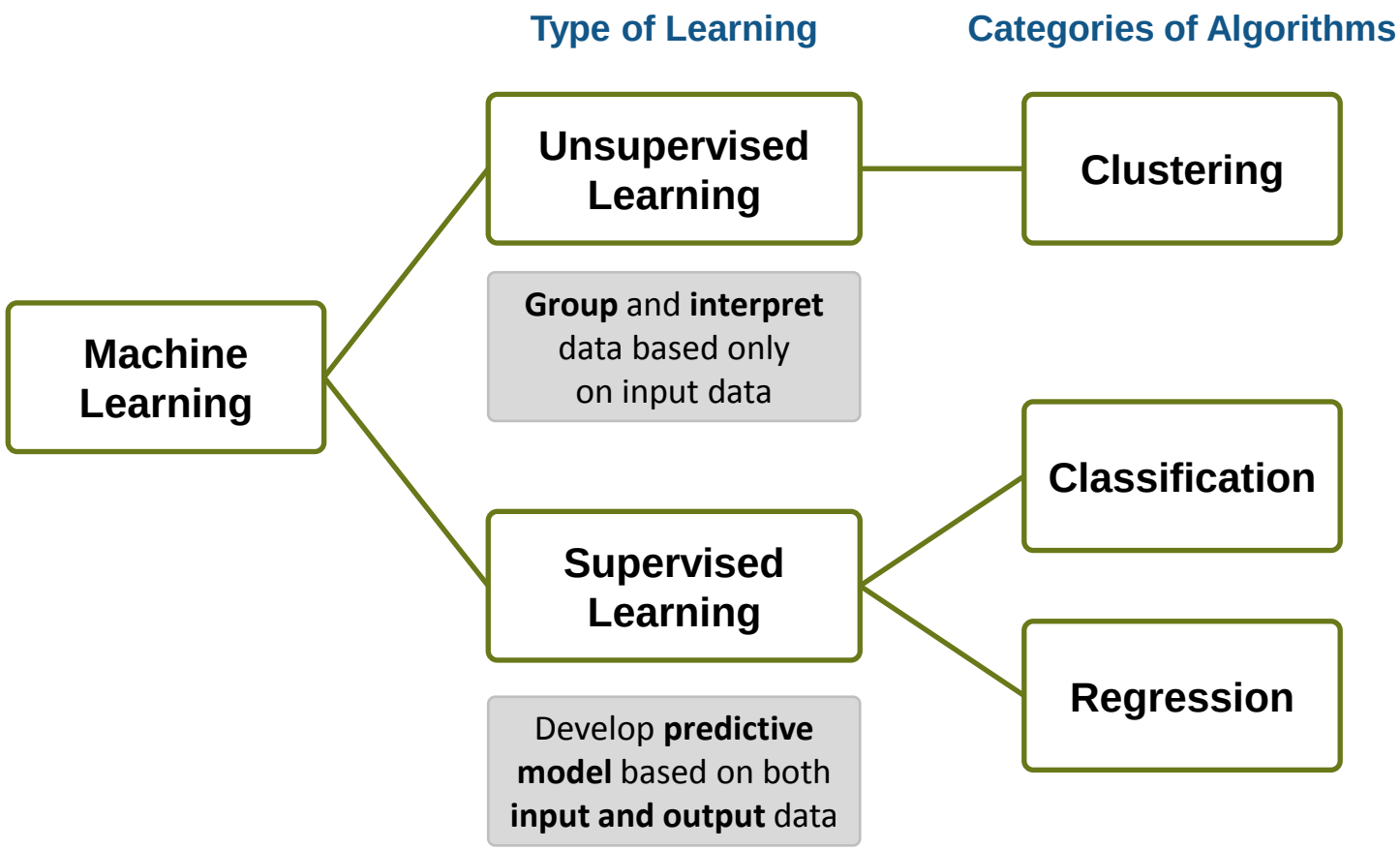
- Machine learning workflow in MATLAB
- Common pitfalls
- Feature selection
- Model evaluation

Worked Example: Loan Data

- Data downloaded from Lending Club
  https://www.lendingclub.com/info/download-data.action

1 loan_amnt	2 term	3 int_rate	4 annual_inc	5 sub_grade	6 total_rec_prncp	7 loan_status
25000	36	11.8900	85000	'B4'	25000	'Fully Paid'
7000	36	10.7100	65000	'B5'	7000	'Fully Paid'
1200	36	13.1100	54000	'C2'	1200	'Fully Paid'
10800	36	13.5700	32000	'C3'	1.0800e+04	'Fully Paid'
7500	36	10.0800	85000	'B3'	5.0250e+03	'Fully Paid'
3000	36	14.2600	80800	'C5'	3000	'Fully Paid'
4000	36	7.8800	148000	'A5'	4000	'Fully Paid'
5600	36	14.9600	45000	'D2'	5.2207e+03	'Charged Off'
3200	36	9.8800	54000	'B1'	3200	'Fully Paid'
4000	36	11.1400	60000	'B1'	2.1704e+03	'Charged Off'
5000	36	11.3400	90000	'C2'	2.6500e+03	'Fully Paid'
2525	36	12.2100	27000	'B5'	2525	'Fully Paid'
10625	36	13.4700	34000	'C4'	10625	'Fully Paid'
2800	60	11.4900	24000	'B4'	2800	'Fully Paid'
7500	36	13.2400	138000	'D3'	7.5000e+03	'Fully Paid'
10000	36	8.5900	56000	'A4'	10000	'Fully Paid'

Types of Machine Learning Algorithms



Classification Problem – Loan Defaults

- Goal
  - Predict which customers will default on their loan
  - When? Before application, post decision, before payment due?

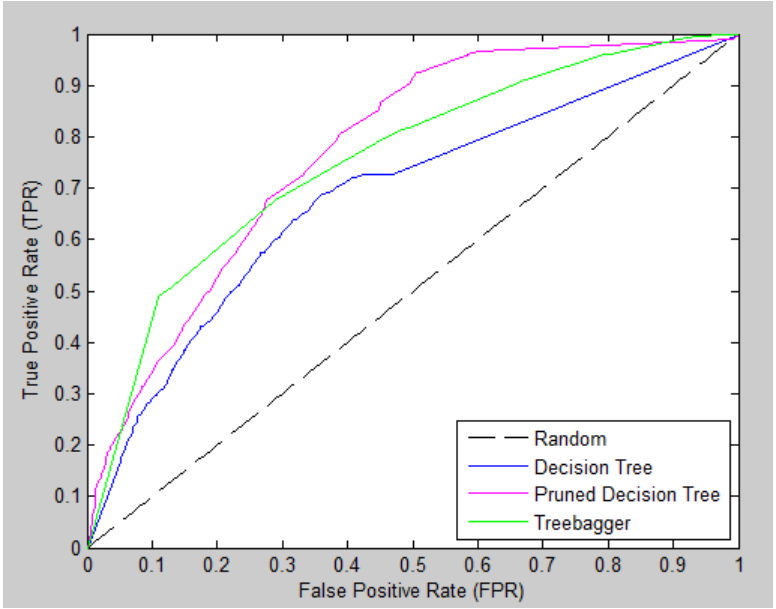
- Approach
  1. Select appropriate features
  2. Train a classifier
  3. Use cross validation to measure accuracy
  4. Evaluate and compare models

Reasons to reject a feature

- Free text input
- Values are all empty
- Data is mostly missing
- Every value is unique (e.g. id)
- Every value is the same
- Two features are directly related
- Leaks information from the future

How do I evaluate a model?

- Cross validation
- Confusion matrix
- ROC curve
- Correlation coefficient
  (for regression)



Performance of Decision Tree:		
	Predicted Paid off	Predicted Defaulted
Actual Paid off	89.02% (12342)	10.98% (1523)
Actual Defaulted	69.74% (1429)	30.26% (620)

Questions?