One coordinate at a time

- Adaboost performs gradient descent on exponential loss
- Adds one coordinate ("weak learner") at each iteration.
- Weak learning in binary classification = slightly better than random guessing.
- Weak learning in regression – unclear.
- Uses example-weights to communicate the gradient direction to the weak learner
- Solves a computational problem

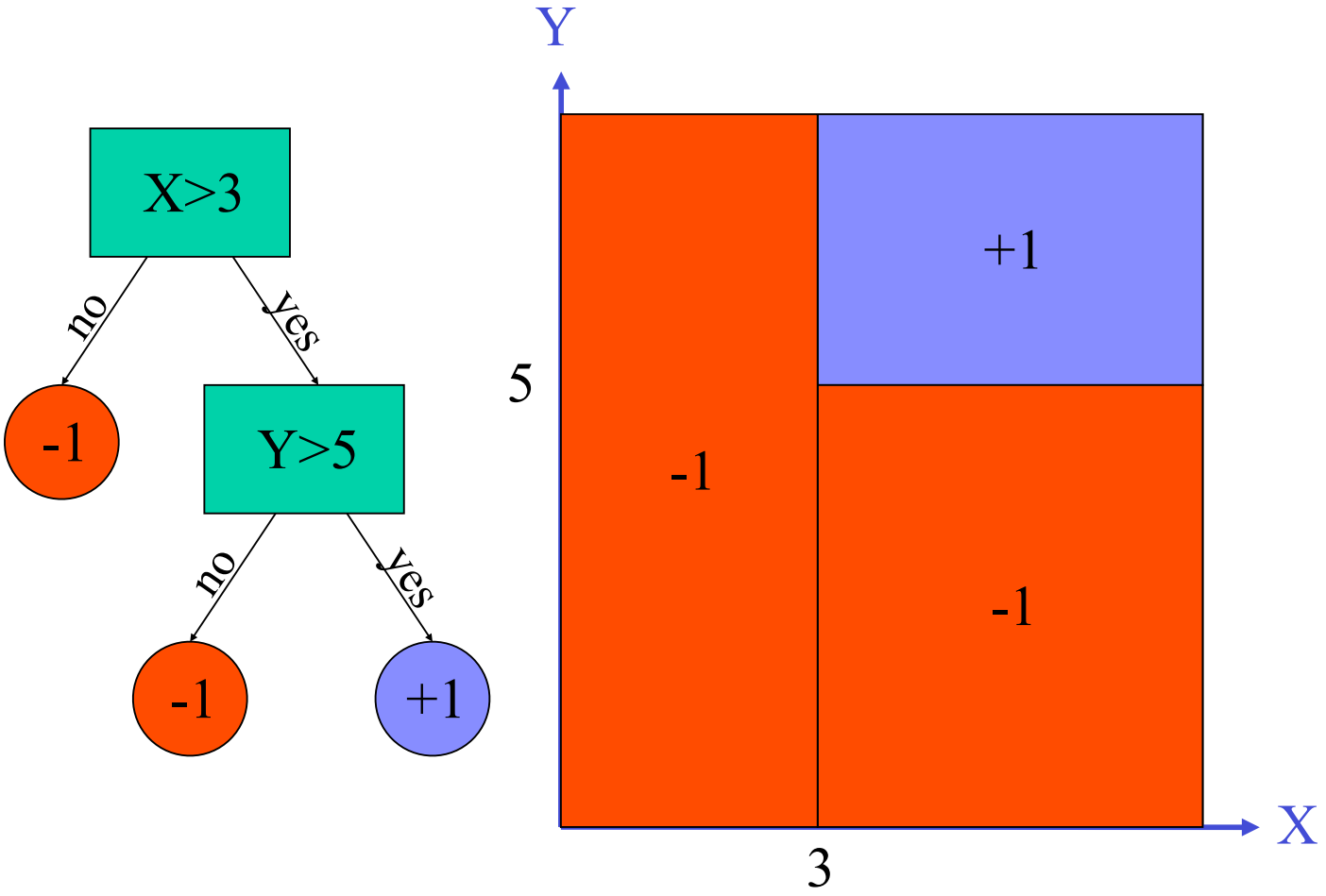
What is a good weak learner?

- The set of weak rules (features) should be flexible enough to be (weakly) correlated with most conceivable relations between feature vector and label.
- Small enough to allow exhaustive search for the minimal weighted training error.
- Small enough to avoid over-fitting.
- Should be able to calculate predicted label very efficiently.
- Rules can be "specialists" – predict only on a small subset of the input space and abstain from predicting on the rest (output 0).

Alternating Trees

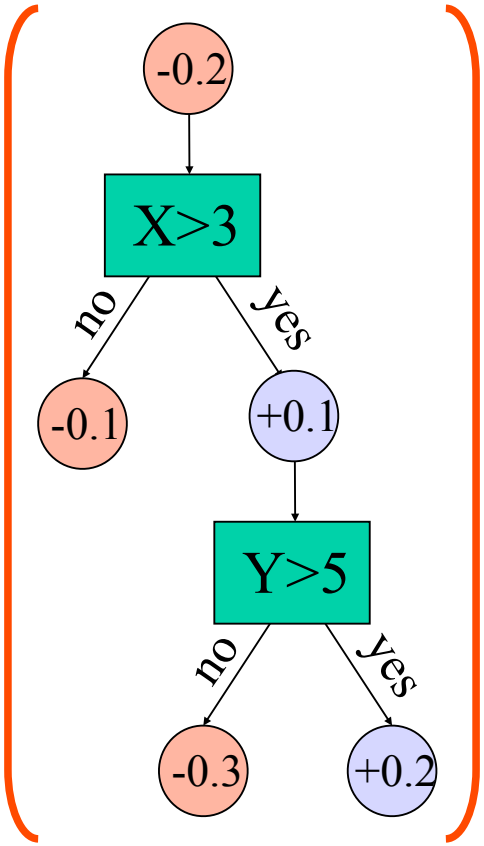
Joint work with Llew Mason

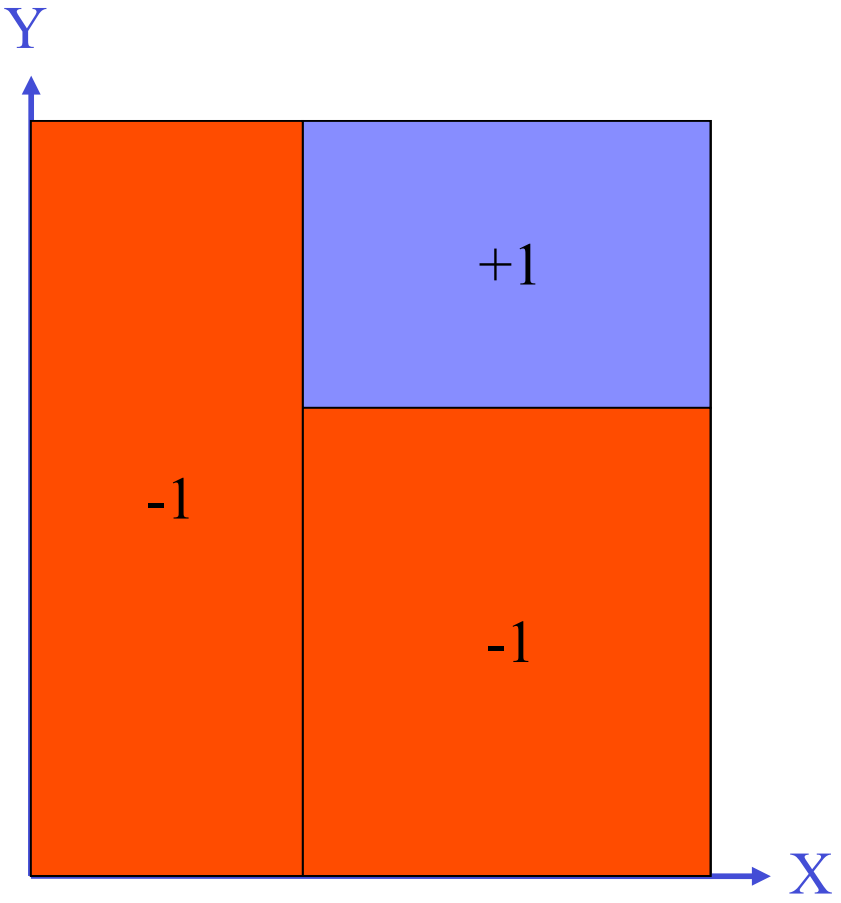
Decision Trees



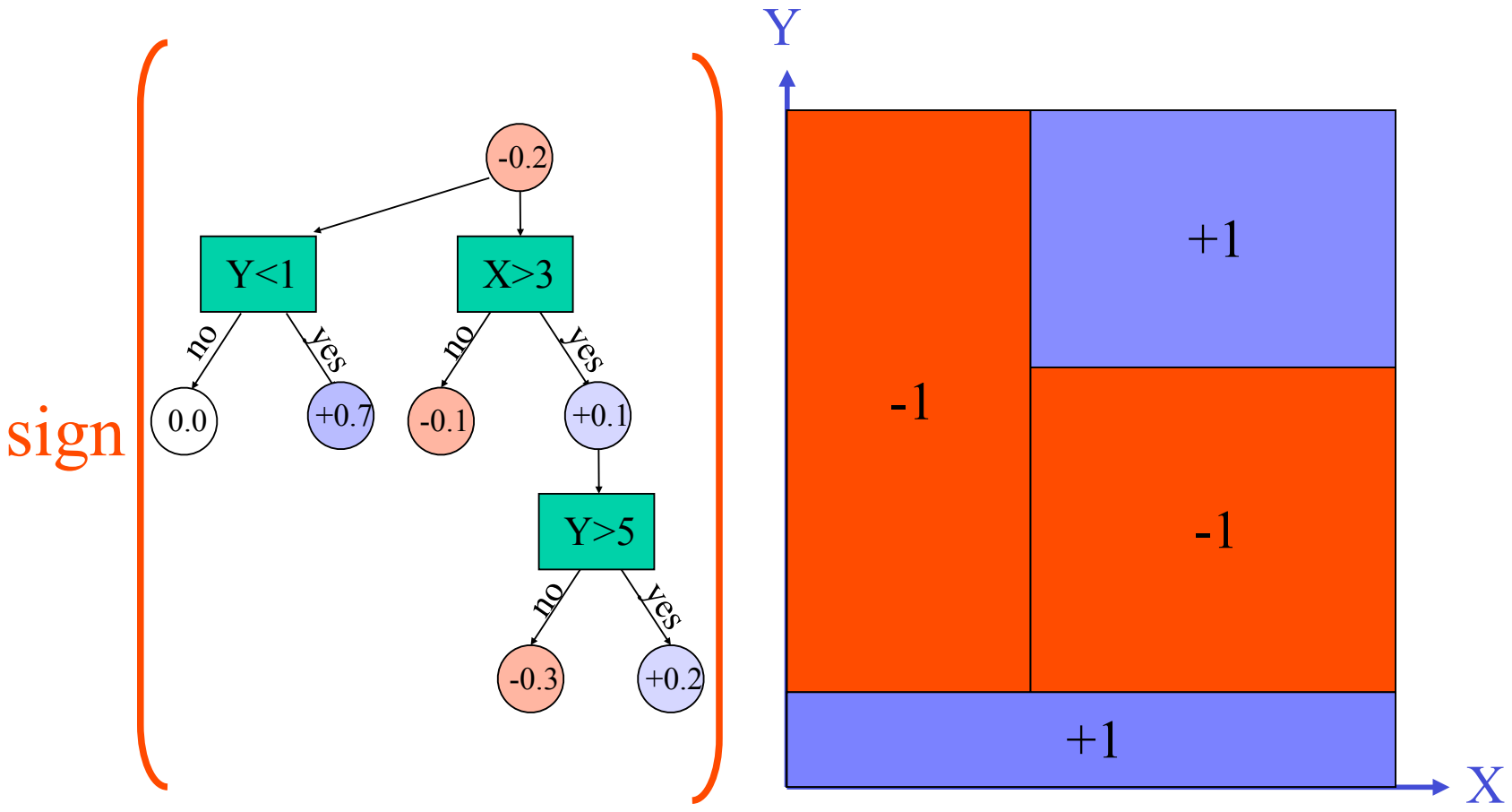
Decision tree as a sum

sign





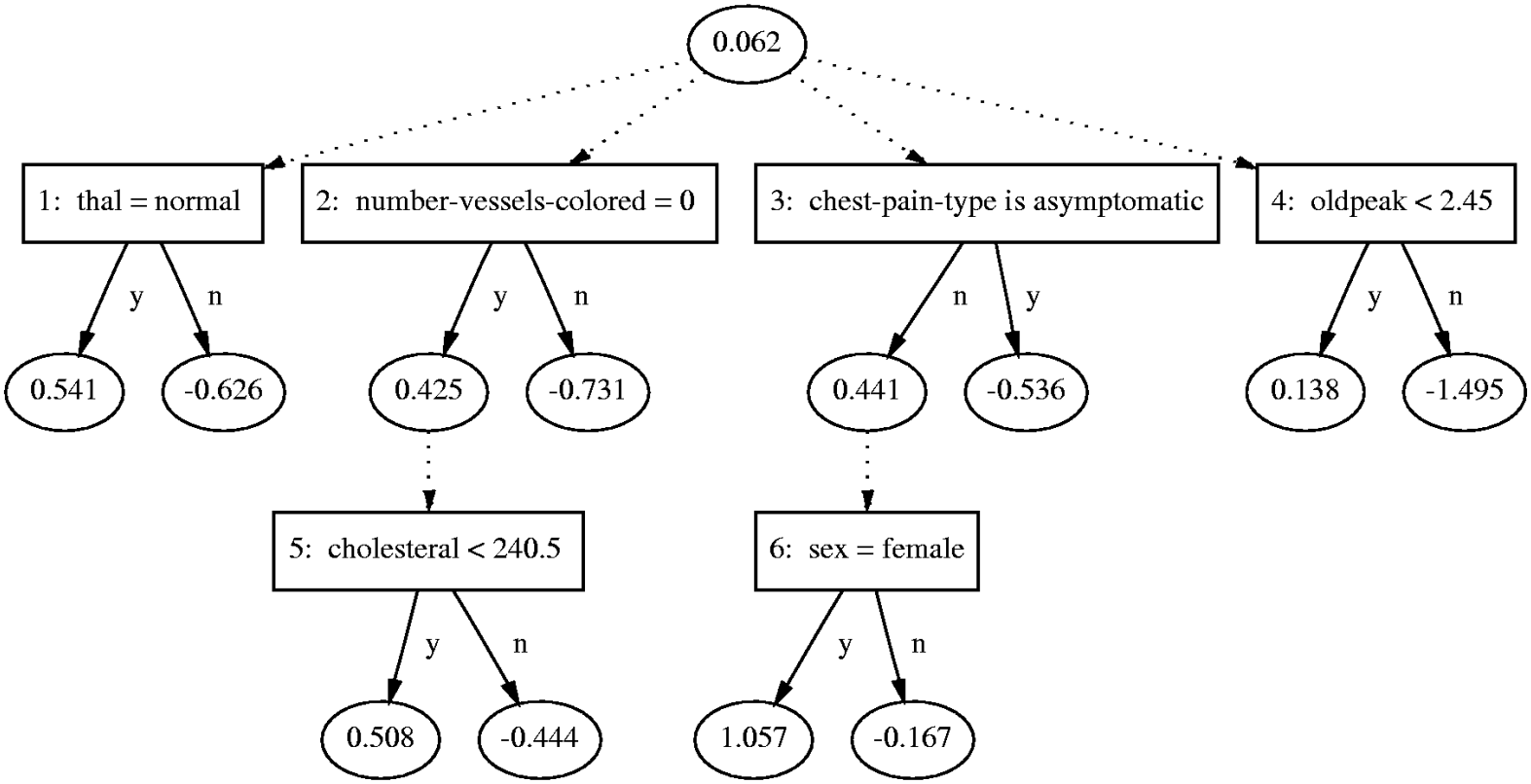
An alternating decision tree



Example: Medical Diagnostics

- `Cleve` dataset from UC Irvine database.
- Heart disease diagnostics (+1=healthy,-1=sick)
- 13 features from tests (real valued and discrete).
- 303 instances.

Adtree for Cleveland heart-disease diagnostics problem



Cross-validated accuracy

Learning algorithm	Number of splits	Average test error	Test error variance
ADtree	6	17.0%	0.6%
C5.0	27	27.2%	0.5%
C5.0 + boosting	446	20.2%	0.5%
Boost Stumps	16	16.5%	0.8%