LEC 7: Decision Trees and Random Forests

Mar 18, 2020

Quiz

https://forms.gle/pgc1WUBDJJvyRkrK7

Presentation:

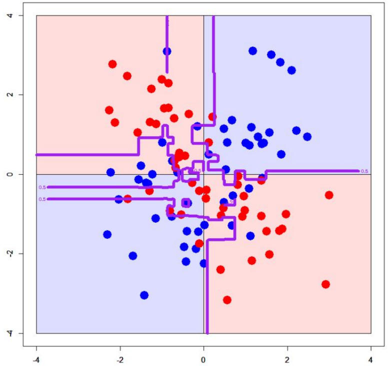
https://docs.google.com/presentation/d/1OD1000y0EhAxPLF7iGvdCWzcUrmhzhxDL83zsT3akps/edit#slide=id.g70ef318bc0_0_5

Some models for classification

1. Supervised - training data with labels provided
    a. Logistic regression and Maximum Likelihood Estimation
    b. Support Vector Machines
    c. Decision Trees and Random Forest
    d. K-Nearest Neighbors
    e. Neural Networks
2. Unsupervised - training data does not require labels
    a. K-Means
    b. Expectation Maximization

Motivation for Decision Trees

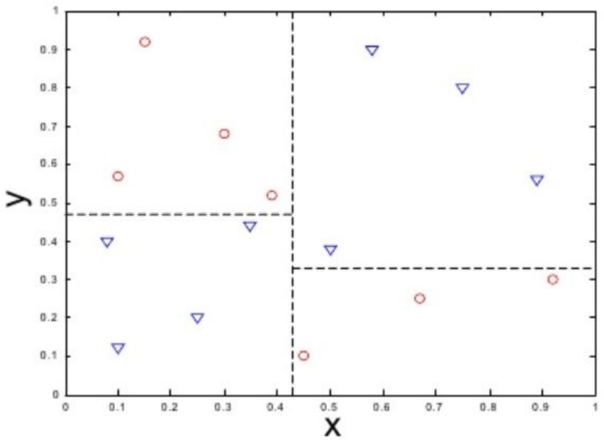
1. Model non-linear, complex and non-contiguous boundaries
2. Works well with categorical data
3. Interpretability: we can see the decisions/splits the algorithm made
4. Can return classification probability (SVMs cannot)



Model: Decision Tree

- **Model:** Decision Tree
- **Target result:** Decision flow that outputs 0 when the predicted class is Class 0 and 1 when Class 1
- **Minimize:** Dissimilarity in true class within a predicted class

An intuitive example

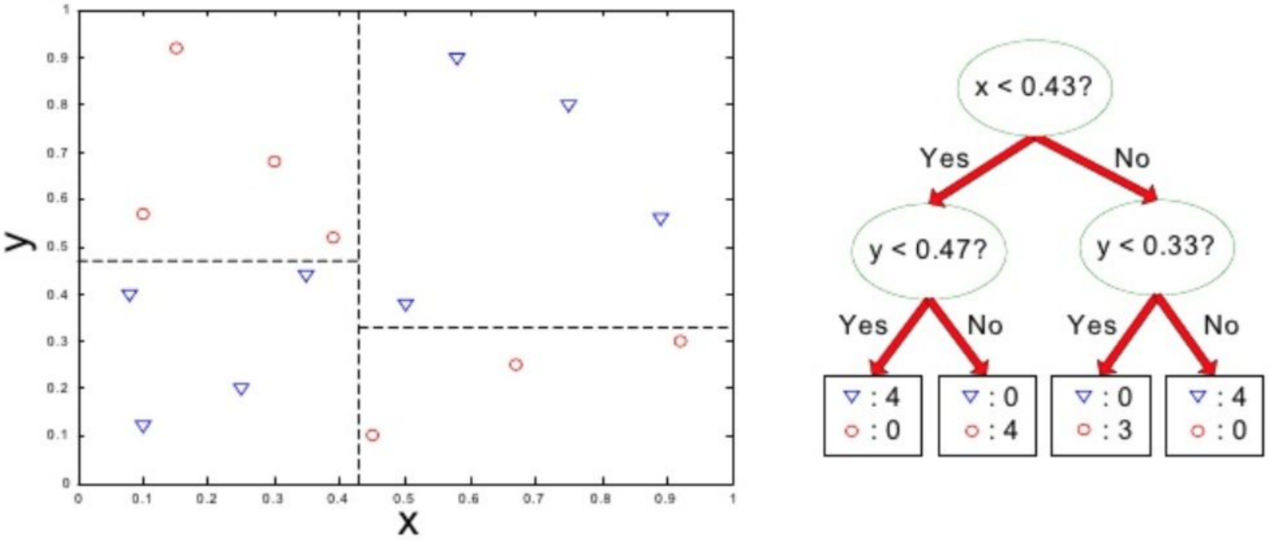


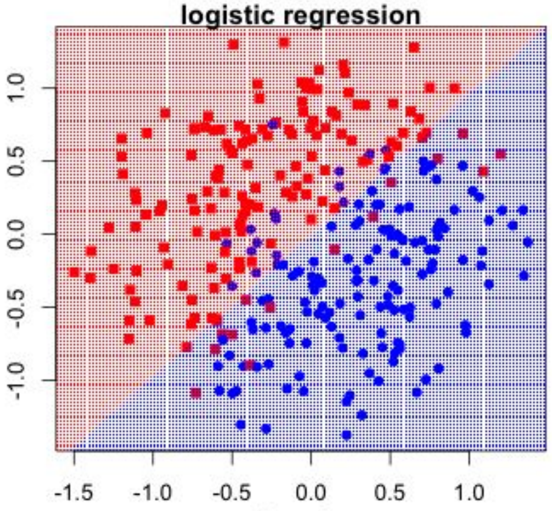
Answer:

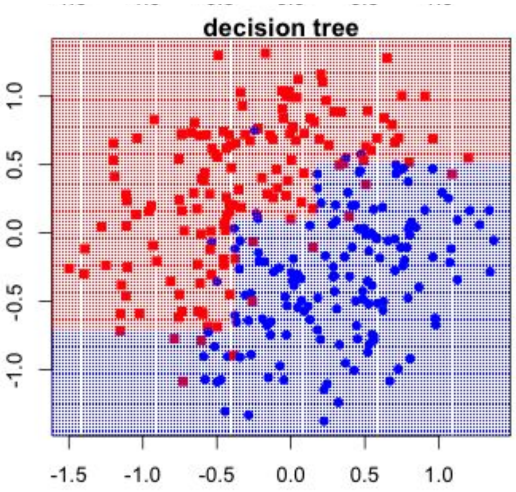
How would you define the decision boundaries for classification?

Decision trees make linear separations along the axes of the features

Notice the boundaries are vertical/horizontal cuts forming rectangular regions



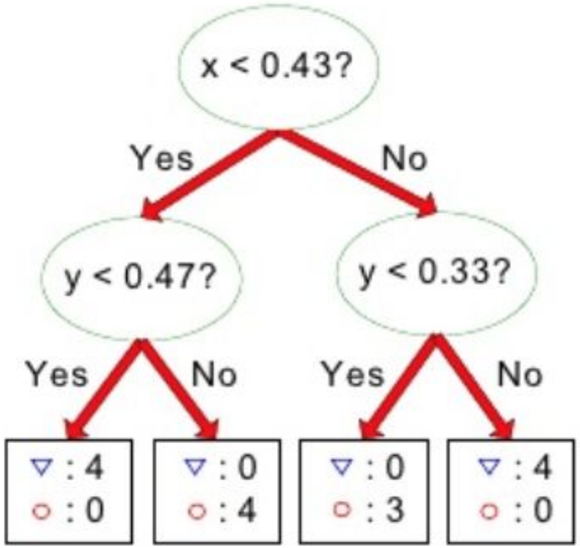

logistic regression
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


decision tree
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

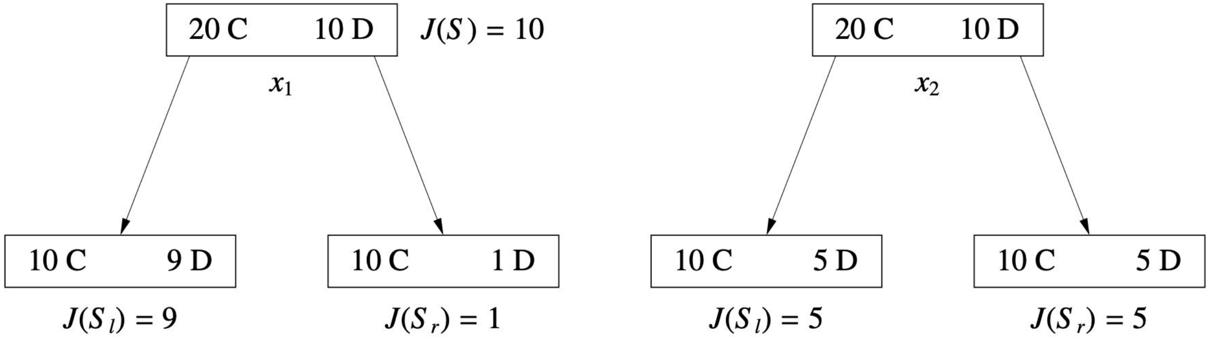
Trees

Vocabulary for decision trees

- **Splitting feature** (x, y)
- **Splitting value** (numerical)
- **Branches**
- **Leaves** contain the prediction
  For the region demarcated by the leaf
- **Parent node**
- **Child node**



Node splitting

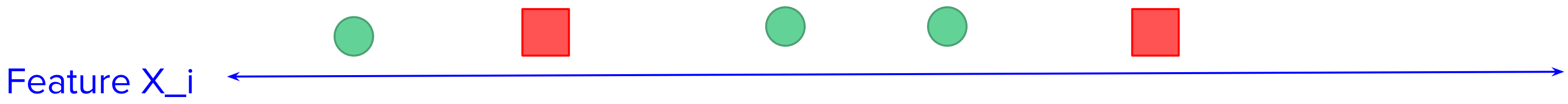


Which is the better split?

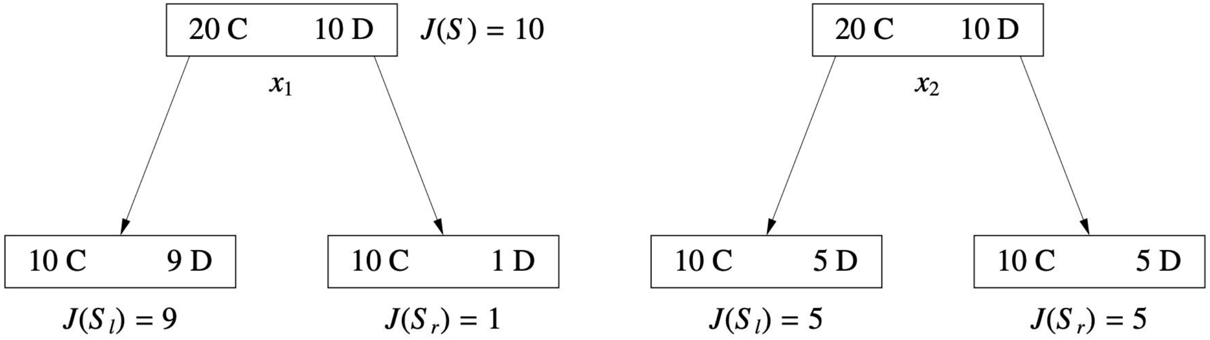
How do we determine the split feature and value to choose next?

Greedy heuristic - what does greedy mean?

- Ideally: If the node is **pure** (only contains one class), then return current state
  - Make this the terminating condition to not split the tree any further
- If the node is not **pure**:
  - Go through all the features x_i in (x, y, ...)
  - For each feature try all the discrete splits into 2 nodes in the range of x_i
  - Test by doing the split: how much is the **"similarity"** within child nodes improved?
- Find the best split and continue this on the child nodes



Measures of “similarity”



Why is cost function Min J = J(S_l) + J(S_r) not a good cost function?

Entropy

For a single data point: Entropy is defined as the surprise of a data point x with label A being in Class A

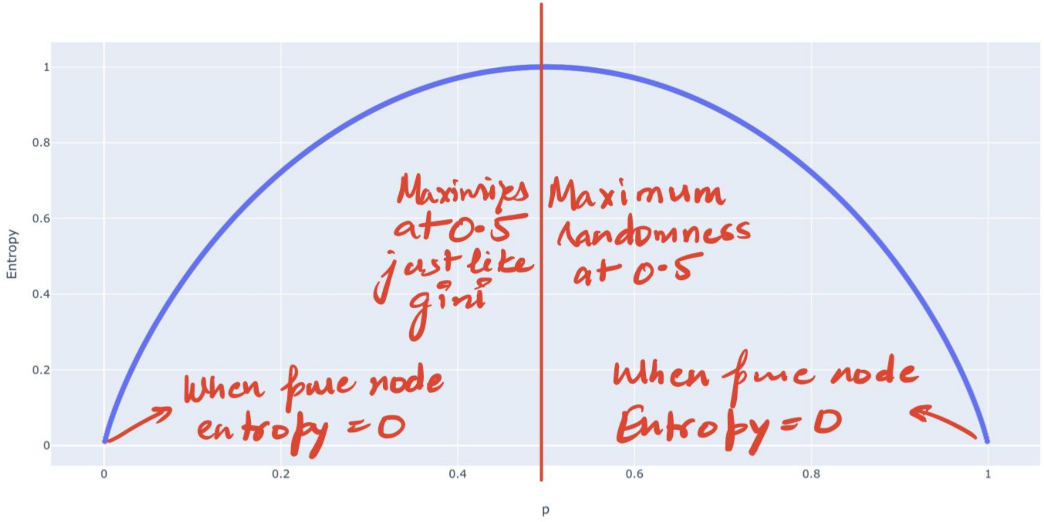
Let p_c be the proportion of points in set S that are in class C.

$$H(S) = -\sum_{\mathrm{C}} p_{\mathrm{C}} \log_2 p_{\mathrm{C}}$$

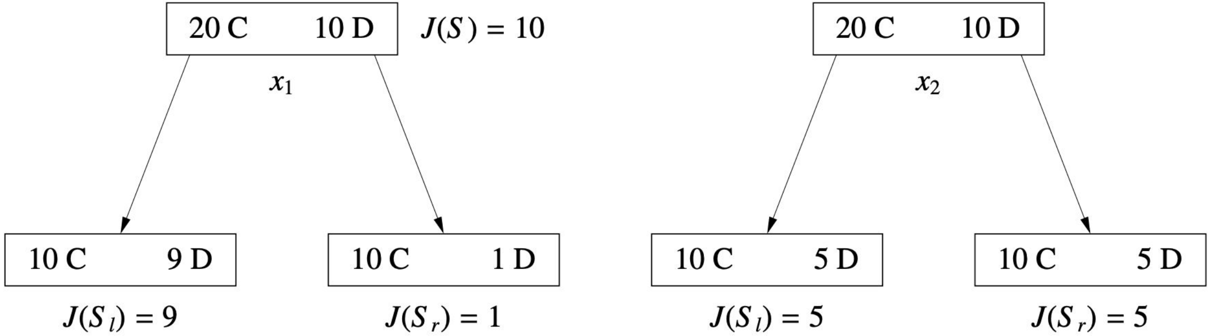
If all points in set S belong to Class A? H(S) = 1 log2(1)= 0
1 is the probability of S bring in class A

Half class A, half class B? H(S) = -0.5 log2(0.5) - 0.5 log2(0.5) = 1

Why entropy function can be minimized



Entropy example



Therefore, best split is the split that lowers entropy the most (take weighted avg of the entropies of the nodes)

H(S1) = $\frac{19}{30}\left(-\frac{9}{19}\log_2\left(\frac{9}{19}\right)-\frac{10}{19}\log_2\left(\frac{10}{19}\right)\right)+\frac{11}{30}\left(-\frac{10}{11}\log_2\left(\frac{10}{11}\right)-\frac{1}{11}\log_2\left(\frac{1}{11}\right)\right)$ = 0.79

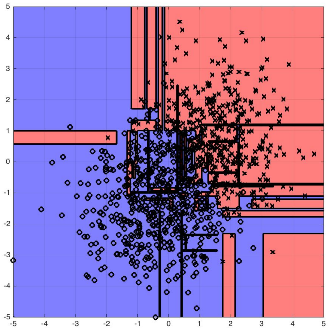
H(S2) = ?

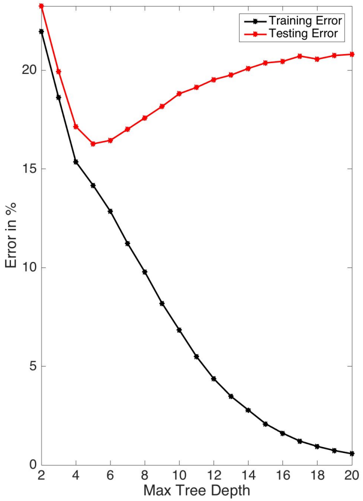
Overfitting

Decision trees can overfit if they become too deep

For this chart, what is the best tree depth?

Solution:
Pruning: try to remove branches from the bottom and see if testing error improves





Ensemble methods: Random Forests

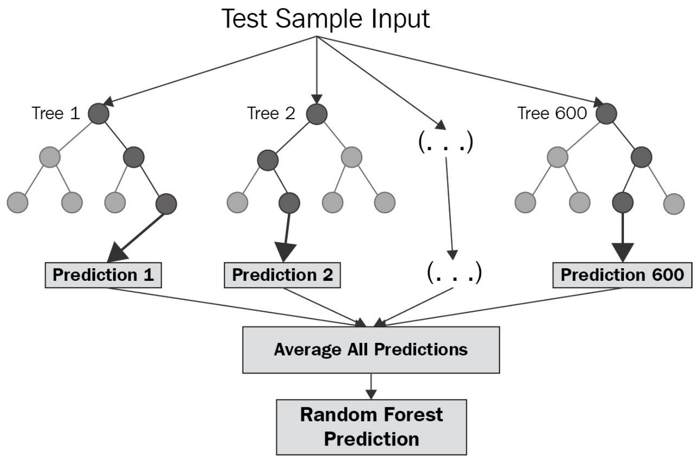
Finds multiple rules - majority wins: effect of drowning out mistakes

Problem: The first split in decision trees has an outsize impact on performance

Solution: At each split, take random sample of **m** features (out of **d**)

If feature x_1 is a super strong predictor, only a fraction of the trees can choose that predictor as the first split. The split tends to “decorrelate” the trees.

When testing a data point, return the majority vote of the trees



Business case study: Netflix

Netflix Prize *Began October 2006*

- Supervised learning task
  - Training data is a set of users and ratings (1,2,3,4,5 stars) those users have given to movies.
  - Construct a classifier that given a user and an unrated movie, correctly classifies that movie as either 1, 2, 3, 4, or 5 stars

- $1 million prize for a 10% improvement over Netflix's current movie recommender

Introduction to Machine Learning and Data Mining, Carla Brodley

Ensemble methods are the best performers…

NETFLIX

Netflix Prize COMPLETED

Home Rules Leaderboard Update

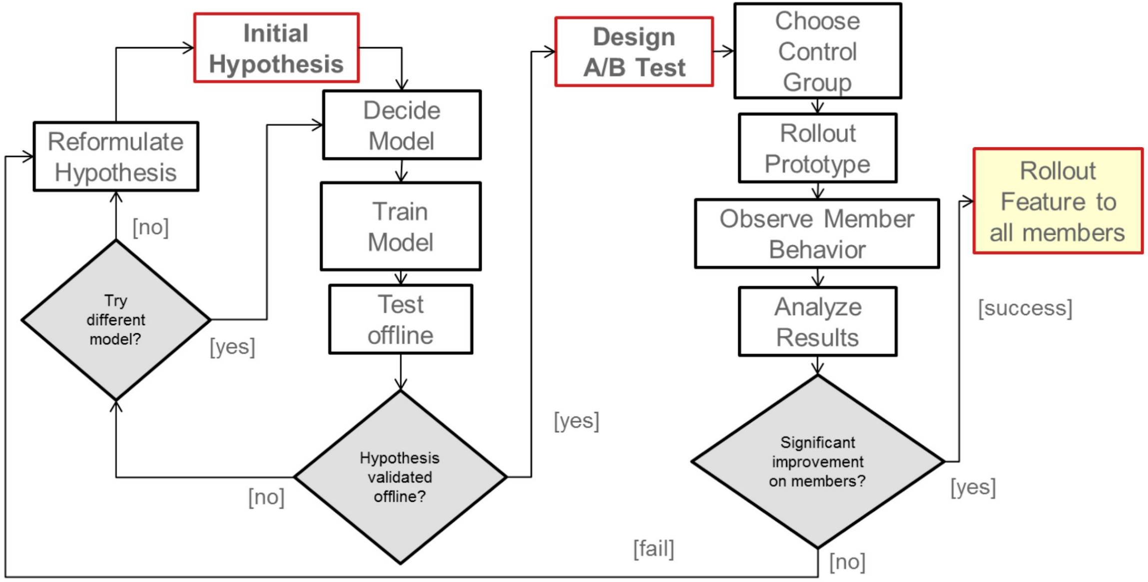
Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22
14	Gravity	0.8643	9.26	2009-04-22 18:31:32
15	Ces	0.8651	9.18	2009-06-21 19:24:53
16	Invisible Ideas	0.8653	9.15	2009-07-15 15:53:04
17	Just a guy in a garage	0.8662	9.06	2009-05-24 10:02:54
18	J Dennis Su	0.8666	9.02	2009-03-07 17:16:17
19	Craig Carmichael	0.8666	9.02	2009-07-25 16:00:54
20	acmehill	0.8668	9.00	2009-03-21 16:20:50
Progress Prize 2007 - RMSE = 0.8723 - Winning Team: KorBell				
Cinematch score - RMSE = 0.9525				

Introduction to Machine Learning and Data Mining, Carla Brodley


Initial Hypothesis
Reformulate Hypothesis
Decide Model
Train Model
Test offline
Try different model?
[no]
[yes]
Hypothesis validated offline?
[no]
[yes]
Design A/B Test
Choose Control Group
Rollout Prototype
Observe Member Behavior
Analyze Results
Significant improvement on members?
[yes]
[success]
[no]
[fail]
Rollout Feature to all members

Feedback

https://forms.gle/Uv3YfeGejQqnFXv39

https://tinyurl.com/tw7u8nd