CS412 office hour

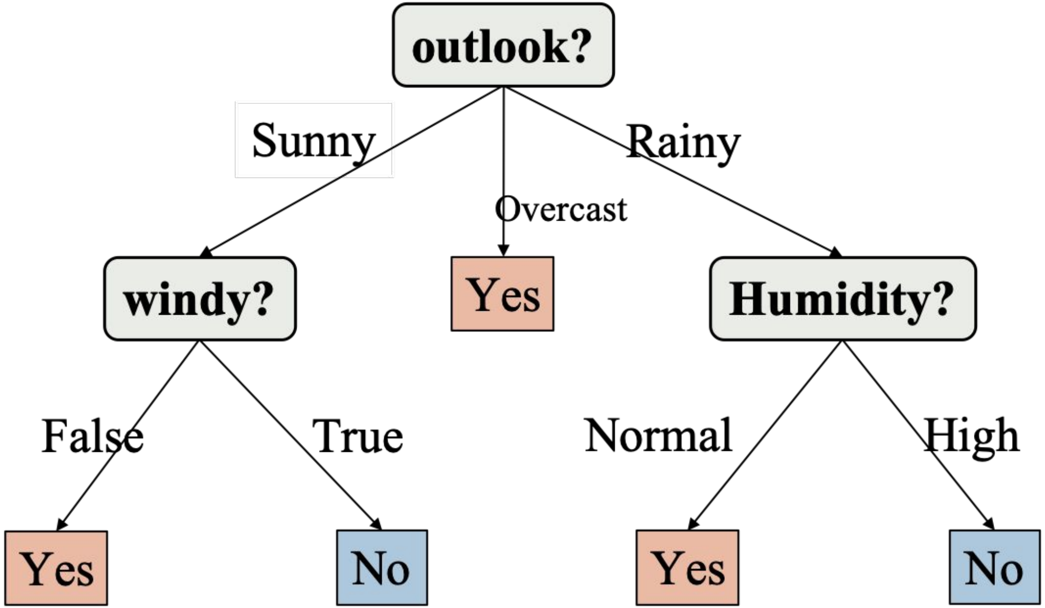
Apr 3, 2019

Today’s Office Hour

- Announcement
    - Exam 2 stats will be released next week
    - We’ll go over exam 2 questions on next Wednesday

- Decision Tree
- QA

Decision Tree



Pros and Cons

- Pros
  - Easy to explain (even for non-expert)
  - Easy to implement (many software)
  - Efficient
  - Can tolerant missing data
  - White box
  - No need to normalize data
  - Non-parametric: No assumption on data distribution, no assumption on attribute independency
  - Can work on various attribute types

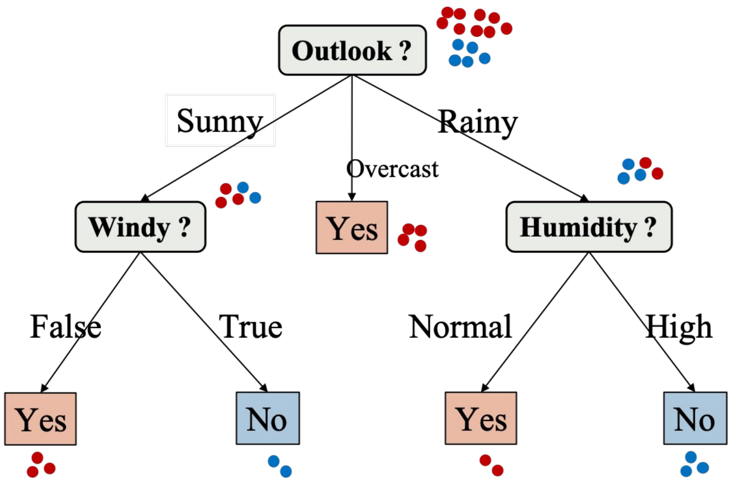
Pros and Cons

- Cons
  - Unstable. Sensitive to noise
  - Accuracy may be not good enough (depending on your data)
  - The optimal splitting is NP. Greedy algorithms are used
  - Overfitting

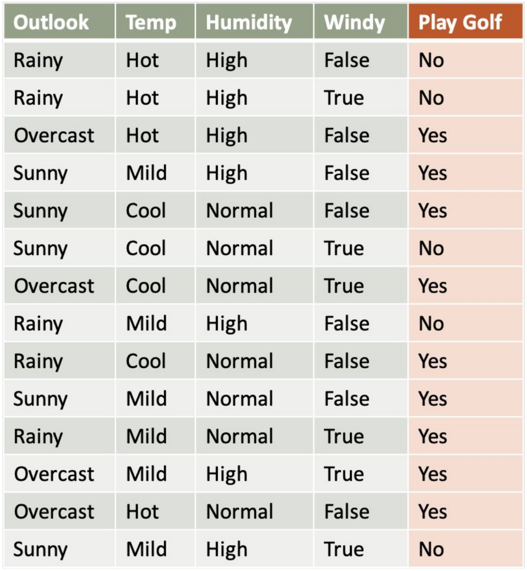
Remark on Decision Tree

- Decision Trees are no longer widely used (stand alone) due to their limitation in performance
- However, they serve as fundamental building blocks of some of the most commonly used classification algorithms
    - Random Forest
    - XGBoost

Building a Decision Tree



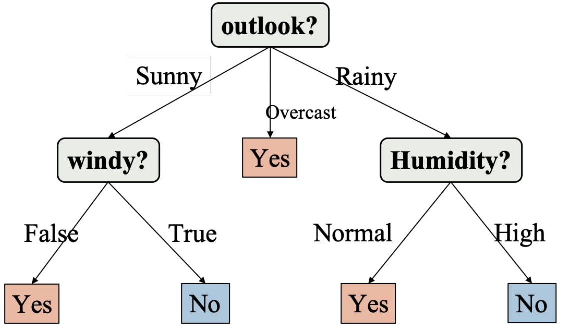
Training data set: Play Golf?

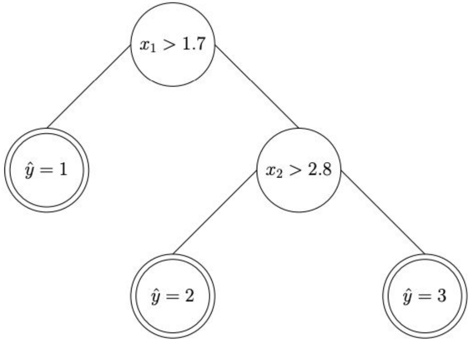


Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Building a Decision Tree

- Optimal: NP hard -> Use greedy approach
- For each node, select a dimension to split, based on some criteria
  - Categorical split
  - Binary split



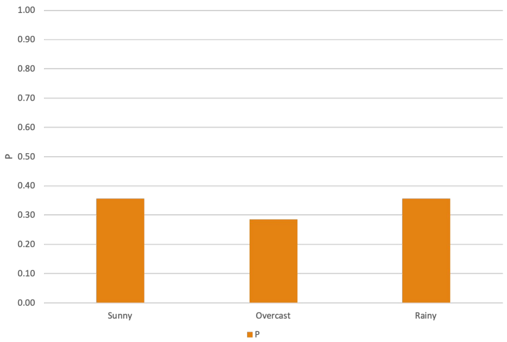


Building a Decision Tree

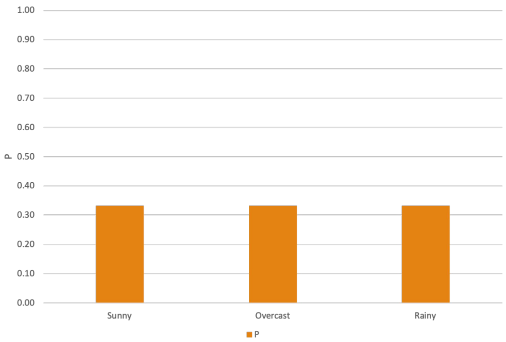
- Greedily select the feature dimension which minimize uncertainty
  - Information gain
  - Gain ratio
  - Gini index (gini impurity)
  - ...
- Procedure: Enumerate all dimensions, compute {gini, info-gain, ...}, select the best one

Entropy

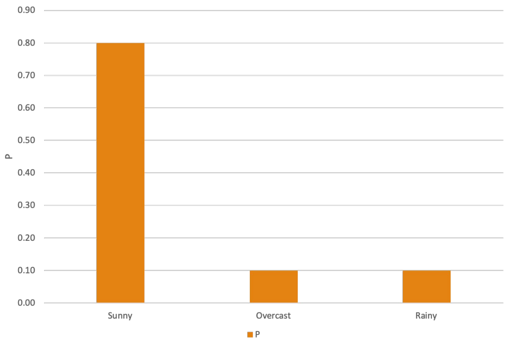
- Uncertainty of a random variable
- The larger the entropy, the more uncertain



entropy=0.475

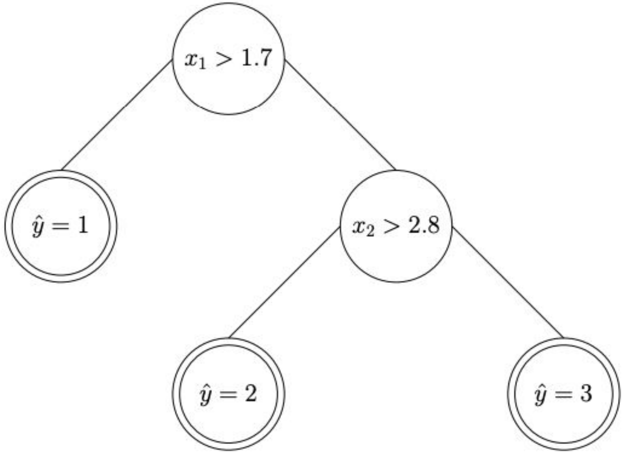


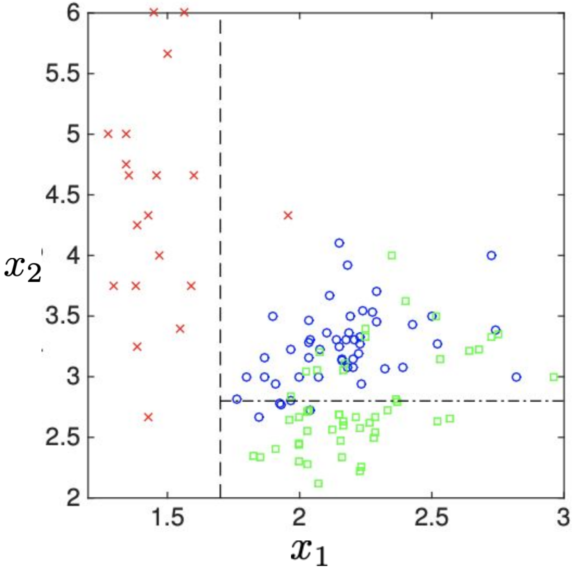
entropy=0.477



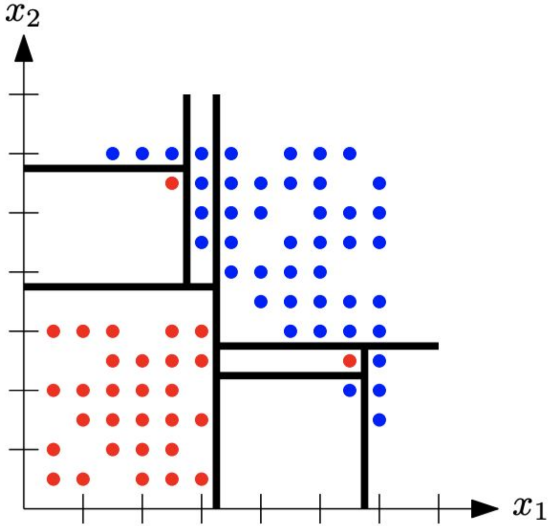
entropy=0.278

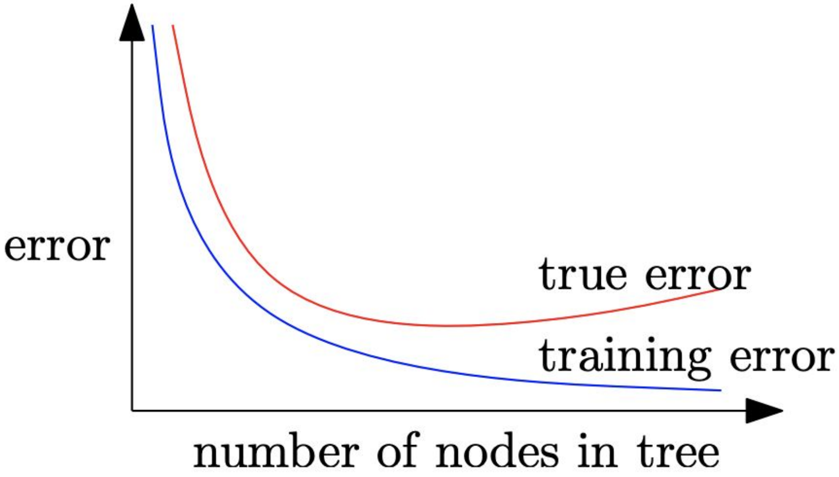
Decision Boundary of Decision Trees





Overfitting

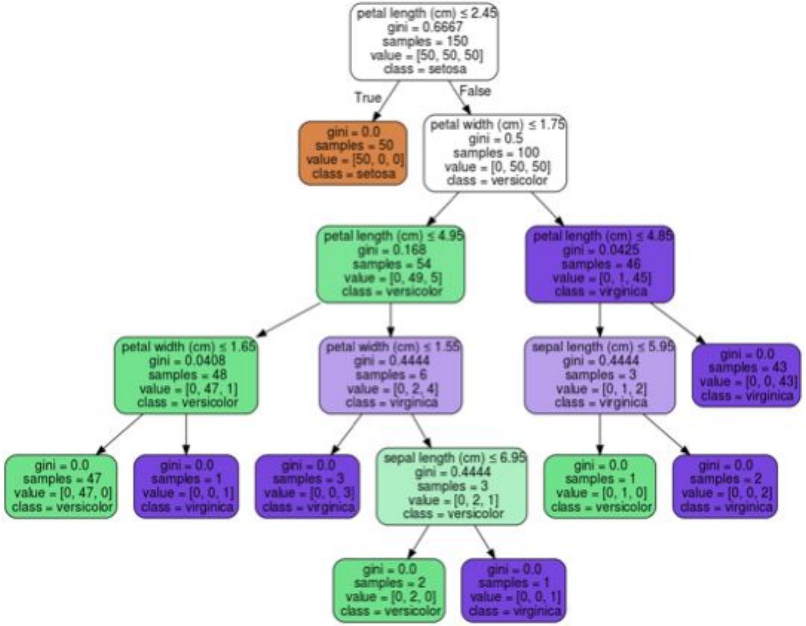




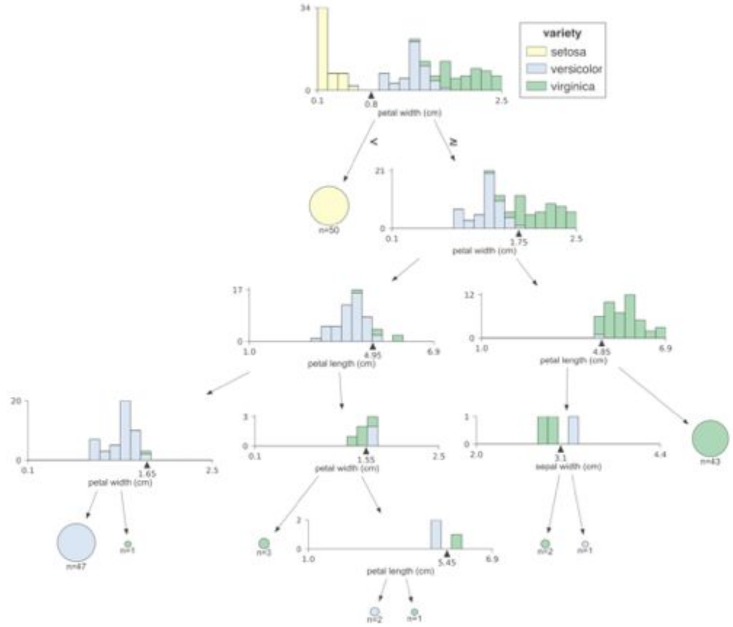
Tree Pruning

- Pre-pruning: Halt tree construction early
- Post-pruning: Remove branches from a “fully grown” tree

Visualization (Scikit-Learn)



Visualization


34
0
0.1
0.8
2.5
petal width (cm)
variety
setosa
versicolor
virginica
<
≥
n=50
21
0
0.1
1.75
2.5
petal width (cm)
17
0
1.0
4.95
6.9
petal length (cm)
12
0
1.0
4.85
6.9
petal length (cm)
20
0
0.1
1.65
2.5
petal width (cm)
3
0
0.1
1.55
2.5
petal width (cm)
1
0
2.0
3.1
4.4
sepal width (cm)
n=43
n=47
n=1
n=3
2
0
1.0
5.45
6.9
petal length (cm)
n=2
n=1
n=2
n=1