Week 5

Lecture

- Decision Tree
    - Entropy and information gain
    - Pruning
    - Dealing with numeric attributes and highly branching attributes.

- Ensemble methods
    - Bagging / boosting / gradient boosting
    - Random forest

Tutorial

Task 1: Build a decision tree for iris data.
Step 1: Load iris data and split into train/test sets.
Step 2: Build decision tree classifier from sklearn
Step 3: pre-prune the tree

Task 2: Ensemble methods for the moons data.
Step 1: Load the moons dataset and split into train/test sets.
Step 2: Implement bagging method/Random Forests/AdaBoost/Gradient Boosting.

TODO:
1. Prune the tree with max_depth = 2 and compare with un-pruned trees. [Task 1: DT]
2. Advantages of pruning? [Task 1: DT]
3. Compare decision tree with linear regression and KNN. [Task 1: DT]
4. Try different number of trees. [Task 2: RF]
5. Disadvantages of random forests compared to a single decision tree. [Task 2: RF]

Decision Tree

$$T1 = H(S) = I(\frac{9}{14}, \frac{5}{14}) = 0.940\ bits$$

$$T2 = H(S \mid outlook) = \frac{5}{14} \cdot H(S_1) + \frac{4}{14} \cdot H(S_2) + \frac{5}{14} \cdot H(S_3)$$

$$H(S \mid outlook = sunny) = I(\frac{2}{5}, \frac{3}{5}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971 bits$$

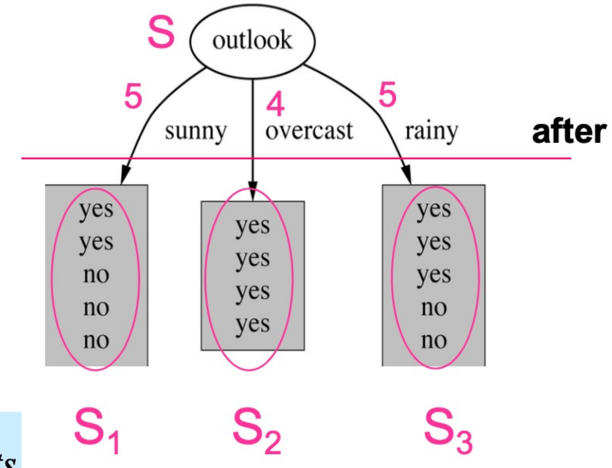
$$H(S \mid outlook = overcast) = I(\frac{4}{4}, \frac{0}{4}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0 bits$$

$$H(S \mid outlook = rainy) = I(\frac{3}{5}, \frac{2}{5}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971 bits$$

$$H(S \mid outlook) = \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0.971 = 0.693 bits$$

Gain(S|outlook)=H(S)-H(S|outlook)= 0.940-0.693=0.247 bits

Before splitting: 9 yes & 5 no



Gain(S|outlook)=H(S)-H(S|outlook)= 0.940-0.693=0.247 bits

- Similarly, the information gain for the other three attributes is:

Gain(S|temperature)=0.029 bits
Gain(S|humidity)=0.152 bits
Gain(S|windy)=0.048 bits

- => we select outlook as it has the highest information gain

Ensemble Method
Bagging vs. Boosting

Similarities

- Use voting (for classification) and averaging (for prediction) to combine the outputs of the individual learners
- Combine classifiers of the same type, typically trees – e.g. decision stumps or decision trees

Differences

- Creating base classifiers:
  - Bagging – separately
  - Boosting – iteratively – the new ones are encouraged to become experts for the misclassified examples by the previous base learners (complementary expertise)
- Combination method
  - Bagging – equal weighs to all base learners
  - Boosting (AdaBoost) – different weights based on the performance on training data

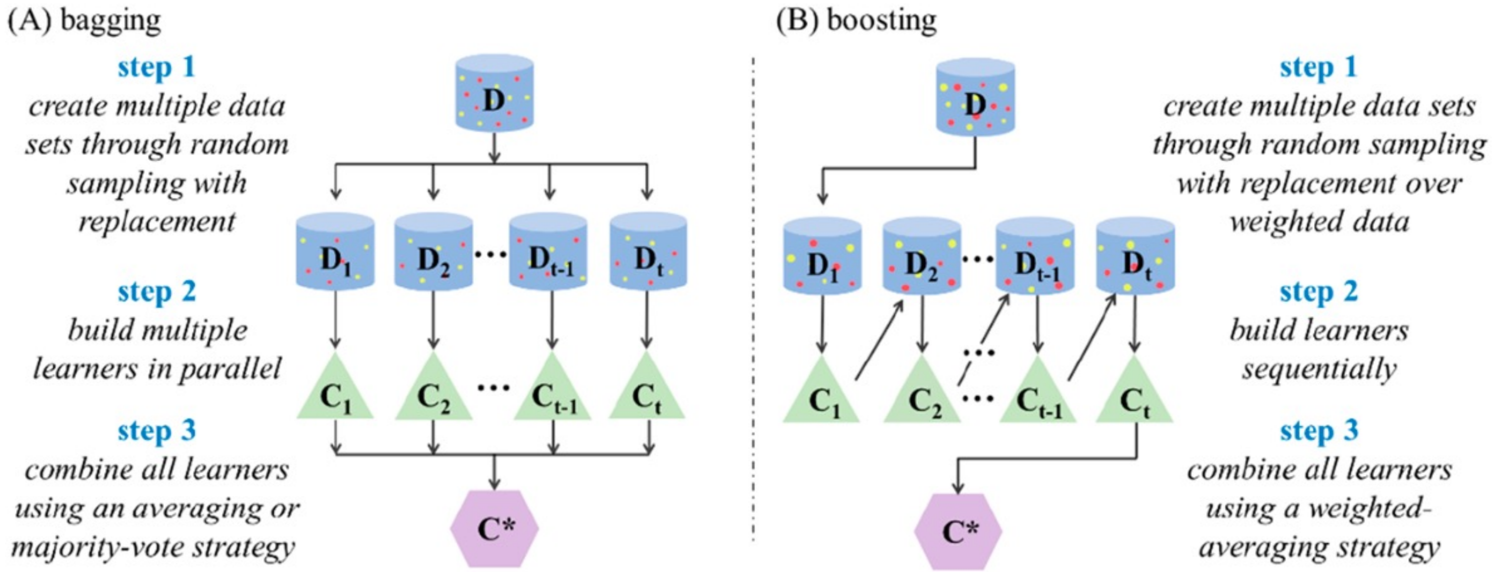


Figure 6. Illustrations of (A) bagging and (B) boosting ensemble algorithms.

Yang, Xin & Wang, Yifei & Byrne, Ryan & Schneider, Gisbert & Yang, Shengyong. (2019). Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chemical Reviews. 119. 10.1021/acs.chemrev.8b00728.