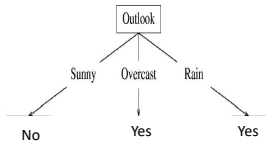
Ensemble Methods

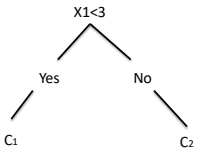
Bagging

1

Decision Stumps

- Depth = 1. Only one question asked. Root + leaves.





2

Bagging

Bootstrap Aggregating

- Create different datasets from original training dataset
- Train same type of model from each dataset (decision stumps in our example)
- Type of each model (called *base learner*) must have *high variance* (significant change to model with small changes in data*)*

3

Creation of Datasets

- Each of the new datasets will have the same size N (which is the size of the original data sets).
- Create k different sets (bootstrap sample), all of size N. (k is a user-chosen parameter)
- Datasets created by sampling with replacement.

4

Bagging Algorithm Training

- For j = 1 to k
  - Create D_j (by sampling with replacement) of size N
  - Train base classifier C_j on the dataset D_j

5

Application Time

- Given a new instance **x**
  1. $\delta(C_j(x)=c) = 1$ if $C_j(x)=c$ (otherwise $\delta(C_j(x)=c)=0$)
  2. For a specific c, $\Sigma_j \delta(C_j(\mathbf{x})=c)$ is the number of base learners which predict class =c
  3. $\text{argmax}_c \Sigma_j \delta(C_j(\mathbf{x})=c)$ is that class c (among all possible output classes) which has highest number of "votes"
- $C^*(\mathbf{x}) = \text{argmax}_c \Sigma_j \delta(C_j(\mathbf{x})=c)$

6

How Different are the Datasets

- Diversity in datasets is key to good performance of the ensemble. (Base learner type is the same)
- The different datasets are created from original by sampling (with replacement)
- They are all of the same size N as original.
- So are the datasets too similar to each other?
- To answer this question, we consider the probability of any arbitrary element is in one of the datasets.

7

Probability of Picking an Instance

- Probability of choosing a specific instance in one attempt (of sampling) = 1/N
- Probability of not choosing a specific instance in one attempt = (1 – (1/N))
- We pick N times
- Probability of not choosing a specific instance for the entire dataset = $(1 – (1/N))^N$
- Probability of choosing a specific instance in a dataset = $1 - (1 – (1/N))^N$

8

How Different are the Datasets

- Probability of choosing a specific instance in a dataset = 1 - $(1 - (1/N))^N$
- This is is approximately 1 – 1/e, as N tends to infinity.
- This is approximately 0.632.
- Unique data points = 63.2% and rest are duplicates.

9

Example

- Original Training Data

x	1	2	3	4	5	6	7	8	9	10
y	1	1	1	0	0	0	0	1	1	1

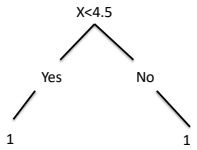
- Base Classifiers: Decision Stump
- Two attributes x<3.5 and x<7.5
- Both have same IG. So if we choose x<3.5
  - If x<3.5 then y=1 else y=0
  - Makes 3 mistakes on training data (8,9,10).
  - Without an ensemble, we won't be able to fit the data.

10

D_1: We choose 1,1,2,2,7,8,9,9,10,10

x	1	1	2	2	7	8	9	9	10	10
y	1	1	1	1	0	1	1	1	1	1

Attributes: x<4.5 (average of 2 and 7) and x<7.5



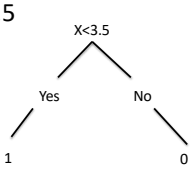
x	1	2	3	4	5	6	7	8	9	10
y	1	1	1	1	1	1	1	1	1	1

11

D_2

x	1	1	2	2	3	4	5	6	7	10
y	1	1	1	1	1	0	0	0	0	1

Attributes x<3.5 and x<8.5



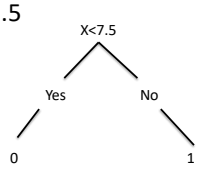
x	1	2	3	4	5	6	7	8	9	10
y	1	1	1	0	0	0	0	0	0	0

12

D_3

x	1	4	4	5	5	7	8	9	9	10
y	1	0	0	0	0	0	1	1	1	1

Attributes x<2.5 and x<7.5



x	1	2	3	4	5	6	7	8	9	10
y	0	0	0	0	0	0	0	1	1	1

13

Ensemble Prediction

Training Set

x	1	2	3	4	5	6	7	8	9	10
y	1	1	1	0	0	0	0	1	1	1

Predictions of Different Models

x	1	2	3	4	5	6	7	8	9	10
O_1	1	1	1	1	1	1	1	1	1	1
O_2	1	1	1	0	0	0	0	0	0	0
O_3	0	0	0	0	0	0	0	1	1	1

Prediction of Ensemble

x	1	2	3	4	5	6	7	8	9	10
O	1	1	1	0	0	0	0	1	1	1

14

Summary

- Create different datasets from the original by sampling with replacement
- Diversity in voting comes from diversity in the datasets
- Usually used with weak learners
- Base learners must have high variance

15