Extending Decision Trees

Alice Gao

Lecture 10

Based on work by K. Leyton-Brown, K. Larson, and P. van Beek

Outline

Learning Goals

By the end of the lecture, you should be able to

- Construct decision trees with real-valued features.
- Construct a decision tree for noisy data to avoid over-fitting.
- Choose the best maximum depth of a decision tree by K-fold cross-validation.

Jeeves the valet - training set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Jeeves the valet - test set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Mild	High	Strong	No
2	Rain	Hot	Normal	Strong	No
3	Rain	Cool	High	Strong	No
4	Overcast	Hot	High	Strong	Yes
5	Overcast	Cool	Normal	Weak	Yes
6	Rain	Hot	High	Weak	Yes
7	Overcast	Mild	Normal	Weak	Yes
8	Overcast	Cool	High	Weak	Yes
9	Rain	Cool	High	Weak	Yes
10	Rain	Mild	Normal	Strong	No
11	Overcast	Mild	High	Weak	Yes
12	Sunny	Mild	Normal	Weak	Yes
13	Sunny	Cool	High	Strong	No
14	Sunny	Cool	High	Weak	No

Extending Decision Trees

1. Real-valued features
2. Noise and over-fitting

Jeeves dataset with real-valued temperatures

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	29.4	High	Weak	No
2	Sunny	26.6	High	Strong	No
3	Overcast	28.3	High	Weak	Yes
4	Rain	21.1	High	Weak	Yes
5	Rain	20.0	Normal	Weak	Yes
6	Rain	18.3	Normal	Strong	No
7	Overcast	17.7	Normal	Strong	Yes
8	Sunny	22.2	High	Weak	No
9	Sunny	20.6	Normal	Weak	Yes
10	Rain	23.9	Normal	Weak	Yes
11	Sunny	23.9	Normal	Strong	Yes
12	Overcast	22.2	High	Strong	Yes
13	Overcast	27.2	Normal	Weak	Yes
14	Rain	21.7	High	Strong	No

Jeeves dataset ordered by temperatures

Day	Outlook	Temp	Humidity	Wind	Tennis?
7	Overcast	17.7	Normal	Strong	Yes
6	Rain	18.3	Normal	Strong	No
5	Rain	20.0	Normal	Weak	Yes
9	Sunny	20.6	Normal	Weak	Yes
4	Rain	21.1	High	Weak	Yes
14	Rain	21.7	High	Strong	No
8	Sunny	22.2	High	Weak	No
12	Overcast	22.2	High	Strong	Yes
10	Rain	23.9	Normal	Weak	Yes
11	Sunny	23.9	Normal	Strong	Yes
2	Sunny	26.6	High	Strong	No
13	Overcast	27.2	Normal	Weak	Yes
3	Overcast	28.3	High	Weak	Yes
1	Sunny	29.4	High	Weak	No

Handling a real-valued feature

- Discretize it.
- Dynamically choose a split point.

Choosing a split point for a real-valued feature

1. Sort the instances according to the real-valued feature
2. Possible split points are values that are midway between two different values.
3. Suppose that the feature changes from X to Y. Should we consider $(X + Y)/2$ as a possible split point?
4. Let L_X be all the labels for the examples where the feature takes the value X.
5. Let L_Y be all the labels for the examples where the feature takes the value Y.
6. If there exists a label $a \in L_X$ and a label $b \in L_Y$ such that $a \neq b$, then we will consider $(X + Y)/2$ as a possible split point.
7. Determine the expected information gain for each possible split point and choose the split point with the largest gain.

CQ: Testing a discrete feature

CQ: Suppose that feature X has **discrete** values (e.g. Temp is Cool, Mild, or Hot.) On any path from the root to a leaf, how many times can we test feature X?

(A) 0 times

(B) 1 time

(C) > 1 time

(D) Two of (A), (B), and (C) are correct.

(E) All of (A), (B), and (C) are correct.

CQ: Testing a real-valued feature

CQ: Assume that we will do binary tests at each node in a decision tree.
Suppose that feature X has **real** values (e.g. Temp ranges from 17.7 to 29.4.) On any path from the root to a leaf, how many times can we test feature X?

(A) 0 times

(B) 1 time

(C) >1 time

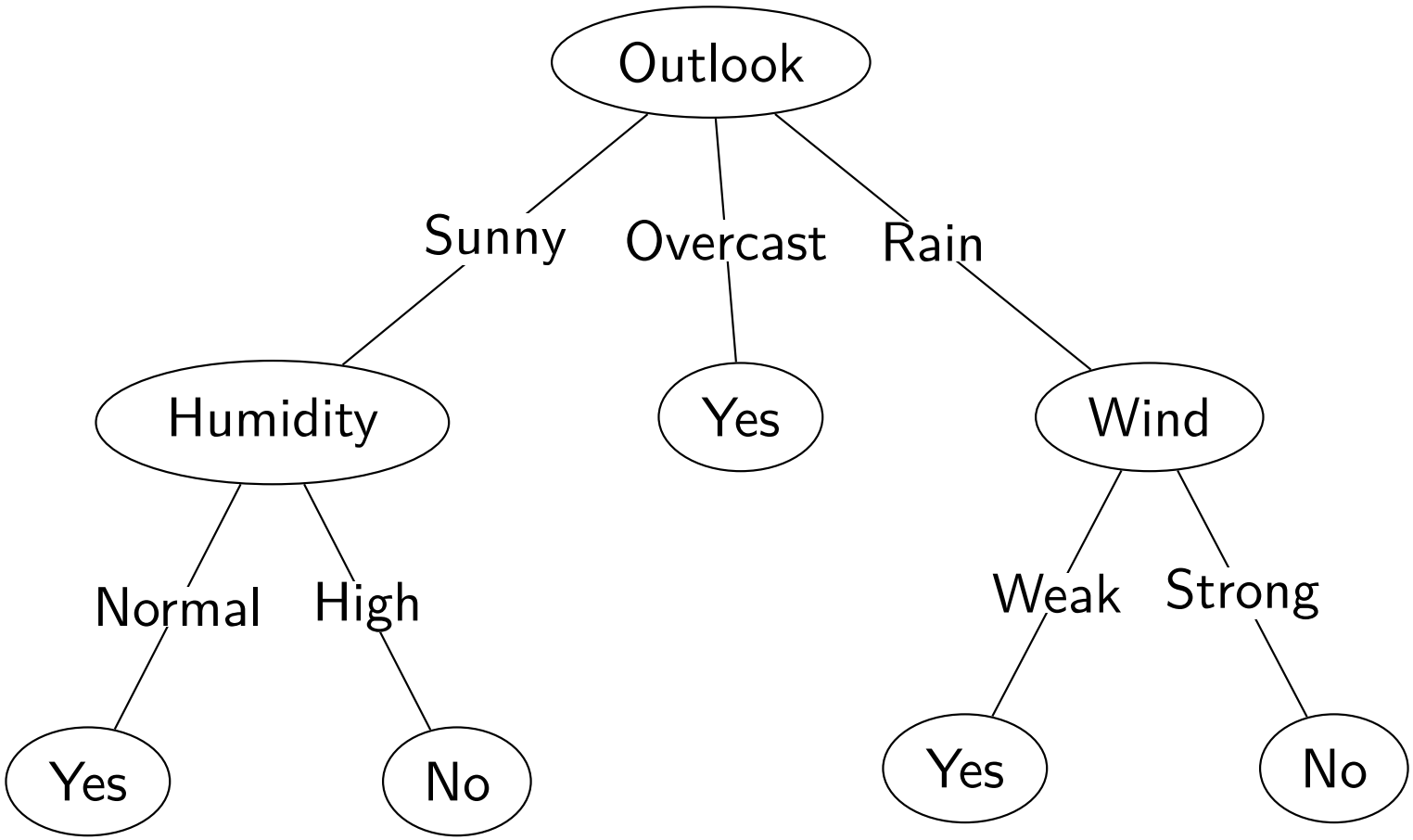
(D) Two of (A), (B), and (C) are correct.

(E) All of (A), (B), and (C) are correct.

Jeeves the valet - training set

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Decision tree generated by ID3

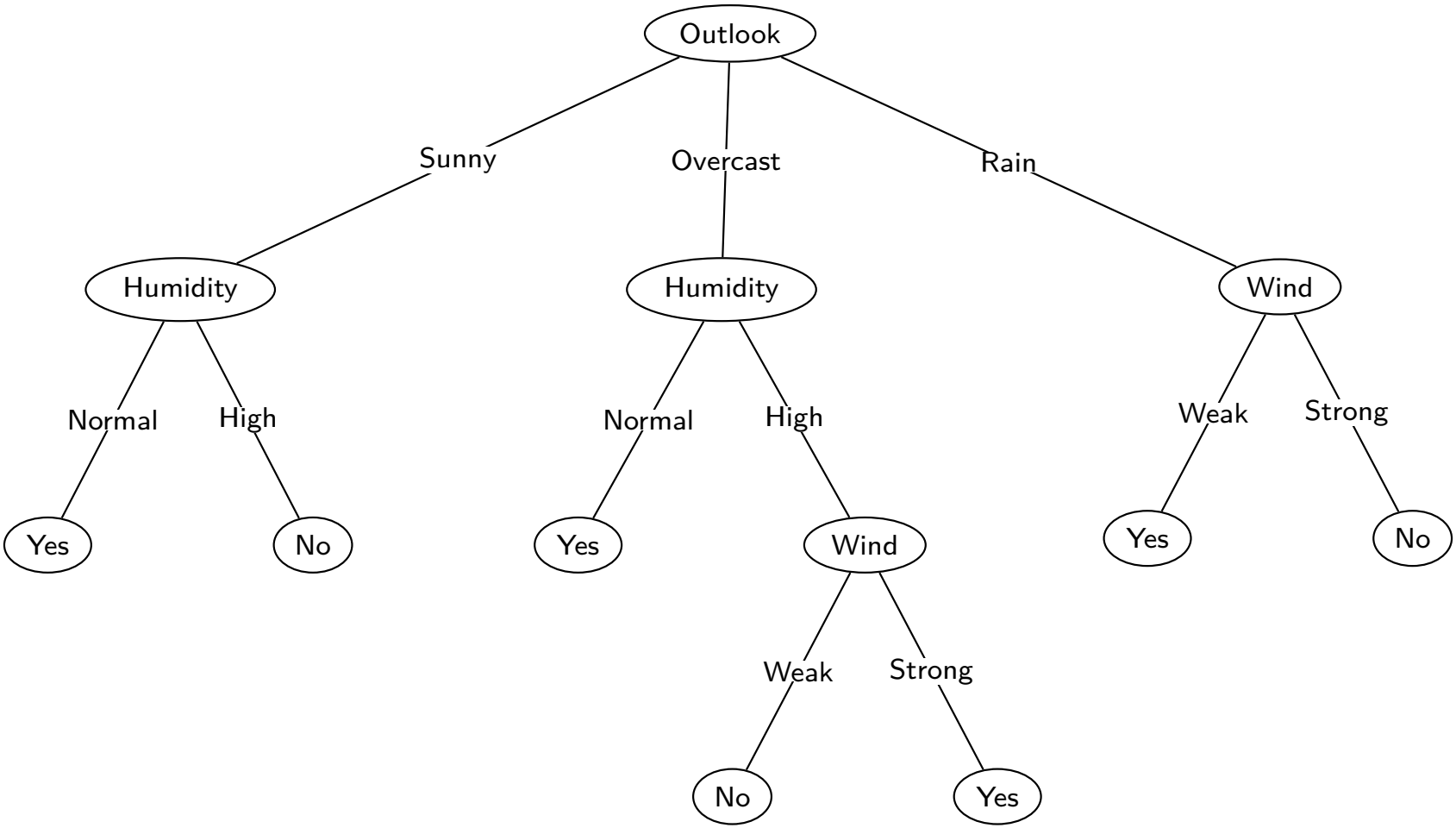


Test error is 0/14.

Jeeves training set is corrupted

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	**No**
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Decision tree for the corrupted data set



Test error is 2/14.

Dealing with noisy data

Problem: When the data is noisy, the ID3 algorithm grows the tree until the tree perfectly classifies the training examples. Over-fitting occurs.
However, a smaller tree is likely to generalize to unseen data better.

- Grow the tree to a pre-specified maximum depth.
- Enforce a minimum number of examples at a leaf node.
- Post-prune the tree using a validation set.

Growing the tree to a maximum depth

- Randomly split the entire dataset into a training set and a validation set. (For example, 2/3 is the training set and 1/3 is the validation set.)
- For each pre-specified maximum depth, generate a tree with the maximum depth on the training set.
- Calculate the prediction accuracy of the generated tree on the validation set.
- Choose the maximum depth which results in the tree with the highest prediction accuracy.

K-fold cross-validation

Suppose that $K = 5$.

1. For each pre-specified maximum depth, do steps 2 to 6.
2. Split the data into 5 equal subsets.
3. Perform 5 rounds of learning.
4. In each round, 1/5 of the data is used as the validation set and 4/5 of the data is used as the training set.
5. Over the 5 rounds, generate 5 different trees and determine their prediction accuracies on the 5 different data sets.
6. Calculate the average prediction accuracy on the validation sets.
7. Choose the maximum depth that results in the highest prediction accuracy on the validation sets.

Revisiting the Learning Goals

By the end of the lecture, you should be able to

- Construct decision trees with real-valued features.
- Construct a decision tree for noisy data to avoid over-fitting.
- Choose the best maximum depth of a decision tree by K-fold cross-validation.