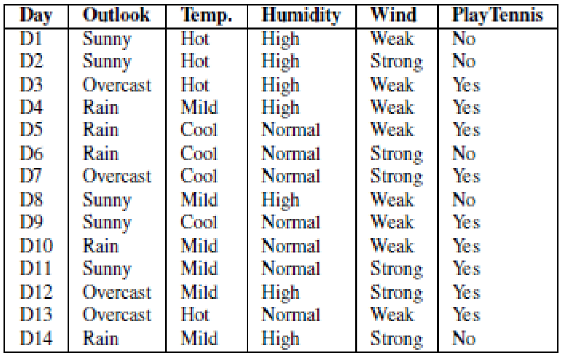
Machine Learning – WS'12
Exercise-8

Prof. Dr. Dr. Lars Schmidt-Thieme, Umer Khan
Information Systems and Machine Learning Lab (ISMLL),
University of Hildesheim

Decision Trees

Problem-1: (Learning Decision Trees)

Consider the following training data:



Day	Outlook	Temp.	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

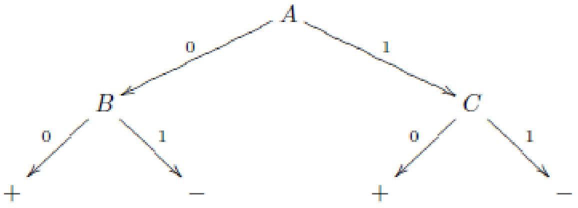
a) Predict the target variable *Play Tennis* with the values *YES* and *NO* for various Saturdays, depending on attributes for each morning. Create a binary decision tree based on the method presented in the lecture ("Greedy strategy"). For each node consider the possible splits. Use the Gini index as a quality criterion for the split.

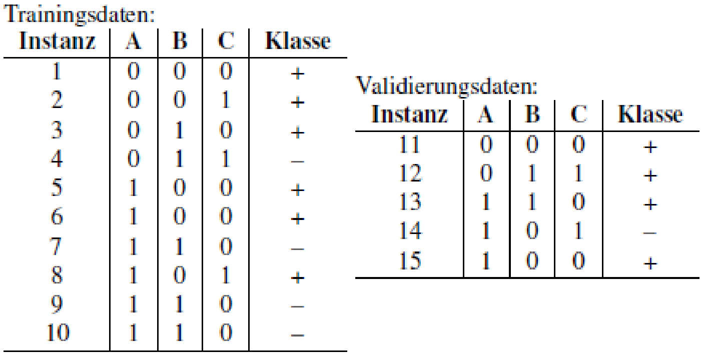
Problem 2:

Consider the given decision tree as well as the training and validation data set.

a) Using the validation data set, estimate the generalization error of the tree using optimistic and pessimistic error approach.

b) Can this tree be pruned? If yes, how? Explain.





Trainingsdaten:

Instanz	A	B	C	Klasse
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	−
5	1	0	0	+
6	1	0	0	+
7	1	1	0	−
8	1	0	1	+
9	1	1	0	−
10	1	1	0	−

Validierungsdaten:

Instanz	A	B	C	Klasse
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	−
15	1	0	0	+

Problem 3:

For each of the following Boolean function in the decision tree:

1. $A \wedge B$
2. $A \vee B$
3. $A \oplus (B \vee C)$
4. $(A \vee B) \wedge (C \vee D)$

In each decision node, is only one variable needs to be queried?