Machine Learning
Exercise Sheet 11

Prof. Dr. Dr. Lars Schmidt-Thieme, Martin Wistuba
Information Systems and Machine Learning Lab
University of Hildesheim

January 21th, 2014
Submission until January 28th, 13.00 to wistuba@ismll.de

Exercise 26: Decision Trees - Missing Values (5 Points)

a) Given is 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

The target variable *PlayTennis* with possible values *yes* and *no* needs to be predicted for different Saturdays depending on the attributes of the respective mornings.
Find a surrogate split for the first primary split using Gini Index as the split quality criterion.

b) Given is the following data:

x_1	x_2	x_3
3	1	Hot
.	.	Hot
4	.	Hot
2	5	Mild
3	3	Cool
.	2	Cool
7	4	Cool
5	1	Mild
3	.	Cool
.	3	Mild

In this dataset some entries are missing. Apply imputation given that

1. x_1 is missing completely at random
2. x_2 is missing at random but it depends on x_3