COMP3010J Tutorial

Decision Trees

1. Consider the following dataset, which contains examples describing several cases of sunburn:

	Name	**Hair**	**Height**	**Build**	**Lotion**	**Result**
1	Sarah	blonde	average	light	no	sunburned
2	Dana	blonde	tall	average	yes	none
3	Alex	brown	short	average	yes	none
4	Annie	blonde	short	average	no	sunburned
5	Emily	red	average	heavy	no	sunburned
6	Pete	brown	tall	heavy	no	none
7	John	brown	average	heavy	no	none
8	Katie	brown	short	light	yes	none

a) What is the entropy of this dataset with respect to the target class label *Result*?

b) Construct the decision tree that would be built with Information Gain for this dataset. Show your work for selection of the root feature in your tree.

c) Using your decision tree from (b), how would you classify the following example X?

	Hair	**Height**	**Build**	**Lotion**	**Result**
X	blonde	average	heavy	no	???

2. Consider the following dataset that aims to predict the risk of a loan application based on 3 features describing each applicant: credit history, debt, and income. Applications are assigned to 3 different risk classes: low, medium, high.

	Credit History	**Debt**	**Income**	**Risk**
1	bad	low	0to30	high
2	bad	high	30to60	high
3	bad	low	0to30	high
4	unknown	high	30to60	high
5	unknown	high	0to30	high
6	good	high	0to30	high
7	bad	low	over60	medium
8	unknown	low	30to60	medium
9	good	high	30to60	medium
10	unknown	low	over60	low
11	unknown	low	over60	low
12	good	low	over60	low
13	good	high	over60	low
14	good	high	over60	low

a) What is the entropy of this dataset with respect to the target class label *Risk* based on the 14 examples above?

b) Compute the entropy of each of the 3 descriptive features.

c) Which one of the descriptive features would be selected by ID3 at the root of a decision tree? Explain your answer. Show all the steps of the calculations.

d) What is the main problem with the Information Gain criterion for feature selection in decision trees?