
COMP9318 (2009 s1) Tutorial 3

Wei WANG

The University of New South Wales

weiw@cse.unsw.edu.au

① Classification

Q1

Consider the following training dataset and the original decision tree induction algorithm (ID3).

Risk is the class label attribute. The *Height* values have been already discretized into disjoint ranges.

- ① Calculate the information gain if *Gender* is chosen as the test attribute.
- ② Calculate the information gain if *Height* is chosen as the test attribute.
- ③ Draw the final decision tree (without any pruning) for the training dataset.
- ④ Generate all the “IF-THEN” rules from the decision tree.

<i>Gender</i>	<i>Height</i>	<i>Risk</i>
F	(1.5, 1.6]	Low
M	(1.9, 2.0]	High
F	(1.8, 1.9]	Medium
F	(1.8, 1.9]	Medium
F	(1.6, 1.7]	Low
M	(1.8, 1.9]	Medium
F	(1.5, 1.6]	Low
M	(1.6, 1.7]	Low
M	(2.0, ∞]	High
M	(2.0, ∞]	High
F	(1.7, 1.8]	Medium
M	(1.9, 2.0]	Medium
F	(1.8, 1.9]	Medium
F	(1.7, 1.8]	Medium
F	(1.7, 1.8]	Medium

Q2

Consider applying the SPRINT algorithm on the following training dataset

<i>Age</i>	<i>CarType</i>	<i>Risk</i>
23	family	High
17	sports	High
43	sports	High
68	family	Low
32	truck	Low
20	family	High

Answer the following questions:

- ① Write down the attribute lists for attribute *Age* and *CarType*, respectively.

-
- ② Assume the first split criterion is $Age < 27.5$. Write down the attribute lists for the left child node (i.e., corresponding to the partition whose $Age < 27.5$).
 - ③ Assume that the two attribute lists for the root node are stored in relational tables name AL_Age and $AL_CarType$, respectively. We can in fact generate the attribute lists for the child nodes using standard SQL statements. Write down the SQL statements which will generate the attribute lists for the left child node for the split criterion $Age < 27.5$.
 - ④ Write down the final decision tree constructed by the SPRINT algorithm.

Q3

Consider a (simplified) email classification example. Assume the training dataset contains 1000 emails in total, 100 of which are spams.

- ① Calculate the class prior probability distribution. How would you classify a new incoming email?
- ② A friend of you suggests that whether the email contains a \$ char is a good feature to detect spam emails. You look into the training dataset and obtain the following statistics (\$ means emails containing a \$ and $\bar{\$}$ are those not containing any \$).

Class	\$	$\bar{\$}$
SPAM	91	9
NOSPAM	63	837

Describe the (naive) Bayes Classifier you can build on this new piece of “evidence”. How would this classifier predict the class label for a

new incoming email that contains a \$ character?

- ③ Another friend of you suggest looking into the feature of whether the email's length is longer than a fixed threshold (e.g., 500 bytes). You obtain the following results (this feature denoted as L (\bar{L})).

Class	L	\bar{L}
SPAM	40	60
NOSPAM	400	500

How would a naive Bayes classifier predict the class label for a new incoming email that contains a \$ character and is shorter than the threshold?