

Possible Questions for the Data-Mining Exam

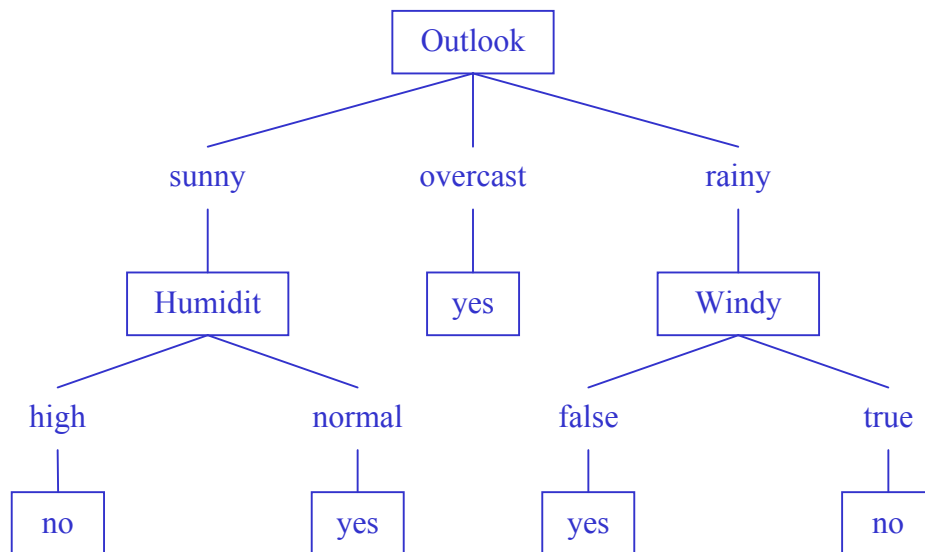
Evgueni Smirnov

Part I. Version Spaces

1. What is the strategy to learn a version space VS that consists of only one description with $\log_2(|VS|)$ instances? Propose a simple algorithm of this strategy for 1-CNF language with discrete attributes.
2. Suppose that the instance i is covered by all the descriptions of the general boundary S . Then explain why the instance is covered by all the descriptions in the version spaces VS as well. Give an example.
3. What is inductive bias of a concept learner? What is the inductive bias of version spaces? What will happen if the conditions of the inductive bias of version spaces do not hold?

Part II. Decision Trees

1. Give all the rules represented by the decision tree given below:



Explain whether there exists an instance covered by at least two resulting rules. Using your answer give advantages and/or disadvantages of the resulting rules (1) when they are interpreted by a human, and (2) when they are interpreted a deductive system.

2. Explain how the *ID3* algorithm can be extended to handle attributes with missing values.

Part III. Decision Rules

1. Consider the following data table:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes

Suppose that the attribute Play is the class attribute. Please give a set of consistent rules for this attribute.

2. Suppose that *A*, *B*, *C*, and *D* are binary attributes. After learning with positive instances only, the resulting rule is:

if *A* and *B* **then** positive Class
else negative Class

Suppose that we have a new negative instance $\langle A, B, C, D \rangle$. Revise the rule to a new rule with exceptions

Part IV. Instance-Based Learning

1. Suppose that the instance language is a 1-CNF languages with 4 Boolean attributes. The training instances are:

$\langle 1,1,0,0 \rangle +$
 $\langle 1,1,0,1 \rangle +$
 $\langle 1,1,1,1 \rangle -$
 $\langle 0,1,1,1 \rangle -$

Your task is to show how the nearest neighbor algorithm works. Suppose that we have an instance $\langle 1,0,0,1 \rangle$ to be classified. Give and explain the resulting classification of the instance for k equal to 1, 2 and 3!

Part V. Bayesian Learning

1. Suppose that we have to build a procedure for post-pruning decision trees that is based on the minimum description length principle. Propose some encoding of the decision trees and data that is suitable for this procedure.

Part VI. Association Rule Learning

1. Consider the table below.

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Please give all the association rules with support greater or equal to 50% and confidence greater than 60%!