
WRITTEN EXAMINATION
“Foundations of Machine Learning and Data Mining”
SUMMER TERM 2007
March 26TH, 2008
PROF. DR. RALF MÖLLER

Name, Student Id:

Name of curriculum:

Signature:

IT IS NOT PERMITTED TO WRITE ON THESE PIECES OF PAPER BEFORE THE START SIGNAL. IN ADDITION, WRITING ON THESE PIECES OF PAPER AFTER THE END SIGNAL WILL HAVE THE EFFECT THAT YOU DO NOT PASS THIS EXAM WITHOUT FURTHER WARNINGS. THIS ALSO APPLIES IF YOU WRITE YOUR NAME AND/OR STUDENT ID AFTER THE END SIGNAL.

Please note:

- 1. Do not open the exam sheets until permitted by the supervisor!**
- 2. Put your student identification card as well as your passport on the table.**
- 3. If you are not registered for the exam, then fill out a proviso, which will be given to you by the supervisor.**
- 4. You have 90 minutes for answering the questions. Additional resources are not allowed.**
- 5. There is sufficient space for your solutions on the examination sheets.**
- 6. If you receive additional pieces of paper from the supervisor, please write your name and student Id also on these pages.**
- 7. If you need to leave the examination room, silently inform the supervisor.**
- 8. The exam is closed book. That means that the only things you are allowed to have on your desk or use during the exam are pens and the exam itself.**
- 9. All phones off. A switched on phone is considered cheating.**
- 10. Cheating will cause you to fail this exam.**

1) Why is Data Mining related to Machine Learning? Give an example and explain it. (2 pts)

2) Suppose we have a hypothesis H that describes how to classify an entity. Explain the terms underfitting and overfitting in relation with hypothesis H . You can do this by giving formulas or graphical representations. (2 pts)

3) What are decision trees used for in learning. Give an example. (2 pts)

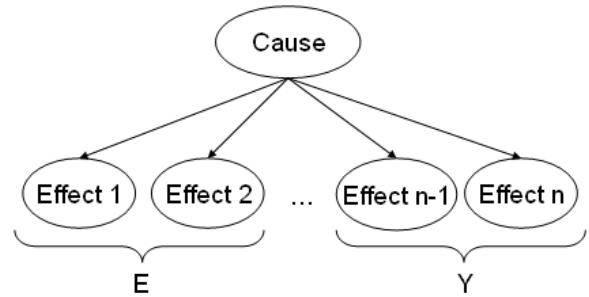
4) Give an algorithm (use some pseudo code) to build up decision trees. Describe possible criteria for choosing an attribute for the next step. (4 pts)

5) Name and explain three weaknesses of decision trees. (3 pts)

6) Describe the general structure and components of a version space. Compare the pros and cons of Version Space based learning and Decision Tree learning (3 characteristics). (5 pts)

7) Describe the terms "absolute independence" and "conditional independence" by giving the corresponding definitions and an example each. (4 pts)

- 8) Specify a general algorithm for answering queries of the form $\mathbf{P}(\text{Cause} | E_1=e_1, E_2=e_2, \dots, E_n=e_n)$, using a naive Bayes approach. You should assume that the evidence may assign values to *any subset* of the effect variables. Therefore, divide the effect variables into those with Evidence, E, and those without evidence, Y (see the following figure) (4 pts)



- 9) Suppose in a hospital there is a monitoring system that keeps track of the blood sugar concentration of patients and alarms the staff if the blood sugar gets to low. The blood sugar gauge measures the blood sugar concentration of a patient.
- Derive and draw a Bayesian network for the problem described above and integrate a failure model for the gauge and the alarm. (2 pts)

- b. Suppose there are just two possible actual and measured blood sugar concentrations, low and normal. The probability that the gauge gives the correct blood sugar concentration is x when it is working, but y when it is faulty. Give the conditional probability table associated with the gauge. (2 pts)

10) Describe the basic idea of a K-Nearest-Neighbour Learner by giving a graphical example. (2 pts)