NTNU
Institutt for datateknikk
og informasjonsvitenskap

Examination paper in

TDT4173 – Machine Learning and Case-Based Reasoning

Thursday December 6th 2012, 09:00 - 13:00

Contact person during the exam: Lars Bungum, Phone: (735) 94475

Language: English
Examination support: D
No written and handwritten examination support materials are permitted.
A specified, simple calculator is permitted.

Grading done by: January 7th.

Read the text carefully. If you find that some problems are not specified at the expected level of detail, please state explicitly what assumptions you make in your answer.

For all questions, brief answers are just as good as longer answers, as long as they provide satisfactory answers.

Question 1: Fundamental issues

a) What does the *inductive learning hypothesis* state? Can machine learning work if we do not assume this hypothesis to be true? Explain your answer.

b) What is *inductive bias*? Describe the two main types of inductive bias. What is meant by the term *overfitting* in the context of inductive learning? Can either type of inductive bias help avoiding overfitting?

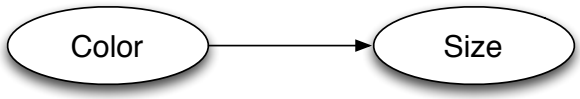
Question 2: Bayesian learning

a) What does *Bayes* theorem state? Describe the formula in your own words. What is the *maximum aposteriori* (MAP) hypothesis, and how can it be determined based on Bayes Theorem? What is the *maximum likelihood* (ML) hypothesis, and under which conditions is it identical to the MAP hypothesis?

b) Assume you have the following dataset, where the two variables *Color* and *Size* are observed:

Color	Size
Red	Big
White	Small
Red	Small
Red	Big
White	Big
Red	Big

You are asked to learn the maximum likelihood parameters for the Bayesian network shown below:



What does it mean to learn the *maximum likelihood parameter*, how should you proceed, and what will the estimators be? (Give the numerical results for the database above.)

c) You get more data for the learning problem described in Question 2b), but the new dataset contains missing values. Which algorithm can you use to learn the maximum likelihood parameters now? Describe the steps of this algorithm.

Question 3: CBR

a) Instance-based and case-based learning methods are called "lazy learners". Why?

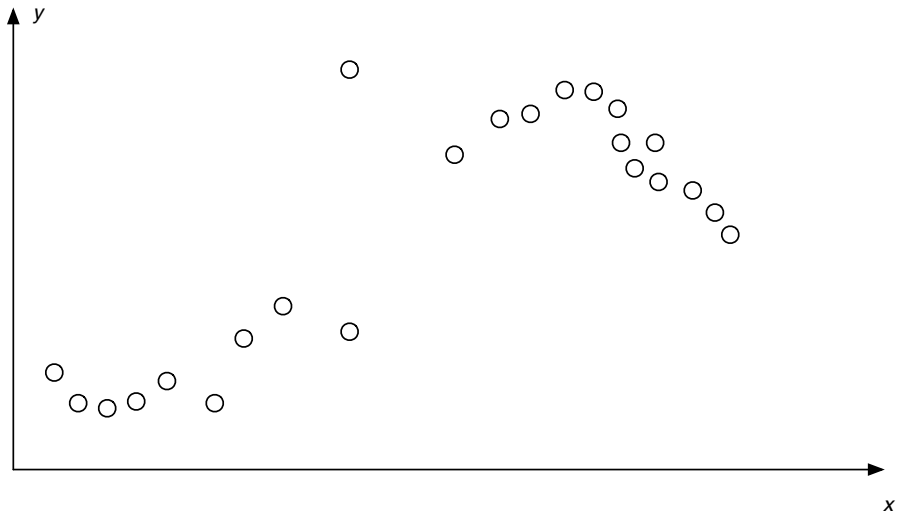
b) Briefly describe the four steps of the CBR-cycle. Discuss how Aamodt amends the classical CBR-cycle with a knowledge intensive explanation engine in his paper *Knowledge-Intensive Case-Based Reasoning in CREEK*.

c) In his taxonomy of similarity mechanisms, Cunningham splits similarity mechanisms into four subcategories. List them and describe their main characteristics.

Question 4: Classification and Regression

a) What is the principal difference between *regression* and *classification*? Give an example of a classification technique and an example of a regression technique. Can classification algorithms be used for regression? Explain your answer, and if your answer is yes: Give an example of a classification algorithm that also works for regression problems.

b) Given data as shown below: How would you proceed when trying to learn a function that maps from the input (x) to the output (y)? Your discussion should include what a reasonable hypothesis space would be, how you parameterize the candidate hypothesis (candidate models) and how you can choose between the different candidates for this dataset. (You are not asked to make any calculations or try to read numerical values off the graph, only to describe ho to proceed.)



Question 5: Mixed questions

a) What is the main idea behind *ensemble learning*? What is a *weak classifier*? Describe the most important differences between *bagging* and *boosting*. In particular, you are asked to discuss how boosting obtains heterogeneity among the learned base classifiers.

b) One of the classification algorithms we have discussed is known as *support vector machines* (SVMs). Describe the main idea of an SVM. Why is an SVM also called a *large margin* classifier?

c) Collaborative filtering methods are often divided into *model-based* methods and *memory-based* methods. Describe each of these method-classes, and discuss the positive and negative aspects of each type.