Exam for Machine Learning

To be solved from 09.00 on Monday 7/5, 2018 to 12.00 on Wednesday 9/5, 2018

Each student must hand in a complete solution to the exam as well as a complete report for all projects even if he or she has been cooperating with other students. Those students who have been working together must state who they have been working with and what is to be regarded as joint work. The exam may be delivered on paper at the info square or alternatively sent as a PDF file to

`eksamen-halden@hiof.no`

The report for all projects should be sent as a single PDF file to the same adress.

The reports for projects one, two and three contribute 65% to the final grade whereas the exam contributes with 35%.

1 The exam (35%)

Clearly explain how each answer has been produced and give good motivations for all calculations. In order to obtain a good grade on an exam problem, all intermediary steps in the machine learning method that you use to solve the problem should be carefully presented.

1.1 Dataset for the exam problems

In this dataset, the task is to predict the risc of lung cancer using three attributes, namely smoking, radon exposure and exercise. The following data has been collected.

Smoking	Radon exposure	Exercise	Risc for lung cancer
normal	little	little	low
little	normal	much	low
normal	much	little	high
much	much	little	high
little	normal	normal	low
normal	normal	little	high
little	little	little	low
little	much	normal	low
normal	little	normal	low
normal	little	much	low
much	normal	little	high
normal	normal	normal	high
little	little	much	low
normal	much	normal	high
much	normal	normal	high
much	much	normal	high
little	much	much	low
normal	normal	much	low
normal	much	much	high
much	little	much	low
much	normal	much	high
much	much	much	high
little	normal	little	low
little	much	little	low
much	little	normal	high
little	little	normal	low

1.2 Exam problems

1. Consider the following simple machine learning model for the data set.

   ```
   if Smoking = little then Risc = low else Risc = high
   ```

   (a) Calculate the classification error for this rule on the data set.

   (b) Find the double sided 95% confidence interval for the error.

   (c) Find the single sided 95% confidence interval for the error.

2. Consider the rule above.

   (a) Calculate the entropy gain that results from using this rule.

   (b) Calculate the Gini impurity that results from using this rule.

3. Use the ID3 algorithm to construct a decision tree that perfectly classifies this data set. Clearly explain all calculations.

4. Manually construct a neural net of perceptrons that perfectly classifies the dataset.

5. Find the lung cancer risc for the following new person using naive Bayes classification.

Smoking	Radon exposure	Exercise	Risc for lung cancer
much	little	little	?

6. Design a genetic algorithm that learns rules that predict risc for a dataset like the one above. Describe and motivate your choice of representation, genetic operators, fitness function and population management as well as other mechanisms you may need to obtain a machine learning methods with good generalizing ability.