

Resit Exam Machine Learning

2010-2011

10 June 2011

This exam is **open book**: you can use Tom Mitchell's "Machine Learning" as well as the lecture slides and any notes you've taken. You can use a calculator.

Answers are allowed in Dutch and English.

Good luck!

Questions

1. Short answers (15 Points)

- (a) (**True or False**): The back-propagation algorithm learns a globally optimal neural network with hidden layers.
- (b) (**True or False**): The error of a hypothesis measured over its training set provides a pessimistically biased estimate of the true error of the hypothesis (i.e., the real-world performance of the hypothesis would be better than the training set performance).
- (c) (**True or False**): Overfitting is more likely when the hypothesis space is small
- (d) (**True or False**): Because gradient descent is guaranteed to find a local optimum, it cannot overfit.
- (e) (**True or False**): When training neural networks, the network weights are updated more often in incremental or stochastic gradient descent than in batch gradient descent.

2. Decision Trees (20 Points)

Summertime is festival time: Pinkpop, Pukkelpop, Parkpop, to name but a few. To help you choose which of the acts to go see at the next festival you're visiting, you will develop a model based on the data in table 1.

- (a) What are the entropies $H(\text{Like}|\text{Style} = \text{Britpop})$ and $H(\text{Like}|\text{Set length} = 30)$?¹

¹Note: if your calculator can't do $\log_2$, use one of the following equations: $\log_2(x) = 1.44 \cdot \ln(x)$ or $\log_2(x) = 3.32 \cdot \log_{10}(x)$. If you didn't bring a calculator, you may give the answer as an expression.

	2019	2018	2017	2016	2015
Revenue	100	100	100	100	100
Operating Profit	20	20	20	20	20
Net Income	15	15	15	15	15
EPS	1.5	1.5	1.5	1.5	1.5
Dividend	5	5	5	5	5
Payout Ratio	33%	33%	33%	33%	33%

1. Which of the following is not a measure of profitability?
2. Which of the following is not a measure of liquidity?
3. Which of the following is not a measure of solvency?
4. Which of the following is not a measure of efficiency?
5. Which of the following is not a measure of risk?
6. Which of the following is not a measure of leverage?
7. Which of the following is not a measure of turnover?
8. Which of the following is not a measure of return?
9. Which of the following is not a measure of cost?
10. Which of the following is not a measure of value?
11. Which of the following is not a measure of growth?
12. Which of the following is not a measure of innovation?
13. Which of the following is not a measure of sustainability?
14. Which of the following is not a measure of social responsibility?
15. Which of the following is not a measure of corporate governance?
16. Which of the following is not a measure of stakeholder engagement?
17. Which of the following is not a measure of brand equity?
18. Which of the following is not a measure of customer loyalty?
19. Which of the following is not a measure of employee satisfaction?
20. Which of the following is not a measure of organizational culture?

5. Instance-Based Learning (15 Points)

- (a) The following data set (Fig. 1) shows classification data with two classes: plus and minus. Which of the two nearest neighbours classifiers has the largest Leave-One-Out Cross-validation error on this data set?
- 1-NN
 - 3-NN

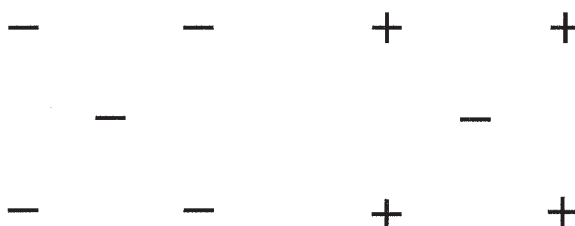


Figure 1: A classification data set

- (b) Suppose that you want to use k-nearest neighbour on a data set that contains a non-numeric attribute 'weight' which can be *light*, *medium* or *heavy*. Could you re-code this attribute to be able to use it? How?
- (c) The same for an attribute 'material' which can have the values *aluminium* or *plastic*.

6. Hypothesis Comparison and Cross Validation (15 Points)

- (a) On a much larger data set than that in table 1, someone has developed a classifier that correctly classifies 700 out of 1000 test examples. Give an estimate of the true error of this classifier, and a 90% confidence interval around that estimate (you may give your confidence interval in the form of an expression).²
- (b) In the above problem, if you increase the size of the test set by a factor of 10, assuming a similar error rate of 30%, what happens to the confidence interval?
- (c) When tuning the parameters of a learning algorithm (e.g., the number of neighbours in kNN) to a given application, why is it necessary to use cross-validation or a test set?

²The table of z-values:

N%:	50%	68%	80%	90%	95%	98%	99%
z_N :	0.67	1.00	1.28	1.64	1.96	2.33	2.58