Machine Learning 2014-2015—Extra Resit

With Answers

22 April 2015

Answers in bold italics. The exam is **open book**: you can use Flach's *Machine Learning* (as a print of the PDF or as a proper book) as well as the lecture slides and any notes you've taken. You can use a calculator as long as it doesn't provide (internet) connectivity.

The exam grade determines 50% of your final grade for the course, the other half is determined by your project grade. The exam grade must at least be 5.5 to pass the course.

Good luck!

Question	*Points*
1. Short answers	15
2. Problem Type	5
3. Decision Trees	15
4. Neural Networks and Regression	20
5. Support Vector Machines	15
6. Clustering	15
7. Performance Measures and Generalisation	15
Total	**100**

Questions

1. **Short answers** no justification required *(3 Points for each question)*

   (a) (**True** or **False**): Decision trees are a typical example of unsupervised learning.
   False

   (b) (**True** or **False**): Neural networks without hidden nodes can learn anything that more general neural networks can, it just takes longer.
   False

   (c) (**True** or **False**): Kernel functions allow nearest neighbour algorithms to take all datapoints into account instead of only the k nearest.
   True

   (d) (**True** or **False**): If a model overfits, its performance on a validation set is typically worse than on the training set.
   True

   (e) (**True** or **False**): Regression models can only use numerical features.
   true

2. **Problem Type** *(5 points)*

   For each of the following learning problems, please indicate whether it is a prediction, regression or classification problem. (An explanation is not required.)

(a) A biologist has given different amounts of food to different rats in his laboratory. He has recorded the weight of each rat after two months. Now he wants to learn how the weight of the rats depends on the amount of food they get.

(b) Each spring a farmer counts the number of newborn sheep. Based on his counts of the previous years he wants to estimate the number of newborn sheep in the coming year.

(c) A computer program tries to determine whether a newspaper article is about politics based on the number of times the article contains the following words/phrases: ‘law’, ‘sports’, ‘newspaper’, ‘hockey’, ‘elections’, ‘human rights’ and ‘party’.

Answers:

(a) regression

(b) prediction

(c) classification

3. **Decision Trees**

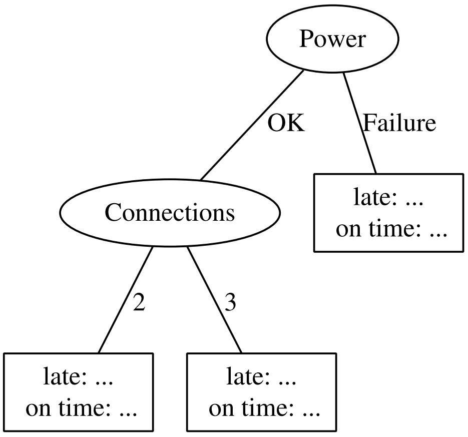
(a) *(5 points)* Based on recent experiences, you decide to build a decision tree to predict if you’ll arrive on time for an exam based on the weather, if there is a power outage and the number of connections you have to make with public transport. You collect the data in table 1.

Weather	Power	Connections	On time
Fine	Failure	2	On time
Fine	OK	2	On time
Fine	OK	2	On time
Cloudy	OK	2	On time
Cloudy	OK	2	Late
Rain	OK	2	Late
Rain	OK	2	Late
Rain	OK	2	Late
Rain	Failure	3	Late

Table 1: On time or not?

Which attribute would the ID3 algorithm choose to use for the root of the tree (assume no pruning)? Show your calculations.

(b) *(5 points)* Using a different technique, someone developed the decision tree below. Draw the corresponding probability estimation tree (basing the probability estimates on the data in table 1).



(c) *(5 points)* When pruning decision trees, the decision to prune a node is based on the accuracy of the resulting tree on a validation set. What would go wrong if we used the train set instead of this validation set?

4. **Neural Networks and Regression**

   In this question, we consider a neural network with two input nodes x_1 and x_2, two hidden nodes h_1 and h_2 and one output node o. The nodes have sigmoid activation functions. The weights for the connections are as follows:

Connection	Weight
$x_1 \rightarrow h_1$	0.8
$x_1 \rightarrow h_2$	-0.4
$x_2 \rightarrow h_1$	0.1
$x_2 \rightarrow h_2$	-0.5
$h_1 \rightarrow o$	0.6
$h_2 \rightarrow o$	0.2

Table 2: Neural Network Weights

   (a) *(5 points)* What is the output of the network if the input is $(0.3, 0.3)$? Show your calculations.

   $o = 0.6051$

   (b) *(5 points)* Let's assume that for this input, the target value $t = 0.0$ and the learning rate $\eta = 0.1$. Compute the new weight for the connection $h_1 \rightarrow o$ using backpropagation. Show your calculations.

   $h_1 \rightarrow o = 0.5918$

   (c) *(5 points)* When training a neural network, what can happen if the learning rate η is set too high?

   i. The network becomes overfitted to the solution it is trained with.

   ii. The network does not stabilise and the algorithm loops indefinitely

   iii. The network's output values are no longer guaranteed to be between 0 and 1

   4(c)ii

   (d) *(5 points)* Describe in your own words what reaching a local optimum (e.g. in gradient descent) means.

5. **Support Vector Machines**

   In this question, we'll consider a dataset for two classes A and B, each with three points.
   Class A: $\{A_1 = (1, 1); A_2 = (-1, 3); A_3 = (2, 6)\}$ and
   Class B: $\{B_1 = (-1, -2); B_2 = (1, -3); B_3 = (-5, -7)\}$

   (a) *(5 points)* Identify the support vectors for a linear Support Vector Machine on this data (Hint: draw a plot of the data).

   $A_1; B_1; B_2$

   (b) *(5 points)* What is the optimal separating line for this data? You may show your answer as an expression or draw a diagram.

   The line midway between the support vectors

   (c) *(5 points)* The kernel trick allows SVMs to define a non-linear separation between classes. Which of the following statements is true (more than one may apply)?

   i. The kernel function replaces the dot-product to calculate the distance between points.
   true

ii. The kernel trick requires an explicit transformation to a higher-dimensional feature space.
false

iii. A kernel can be a polynomial function.
true

6. **Clustering**

(a) *(5 points)* Name two essential differences between hierarchical and k-means clustering.

(b) *(5 points)* Describe, in your own words, the initialisation procedure of the k-means algorithm.

(c) *(5 points)* Consider the data set with 5 cases labelled A,B,C,D and E in figure 1. Draw the dendrogram that would result from a single-link agglomerative hierarchical clustering of that data (assuming Euclidean distance).

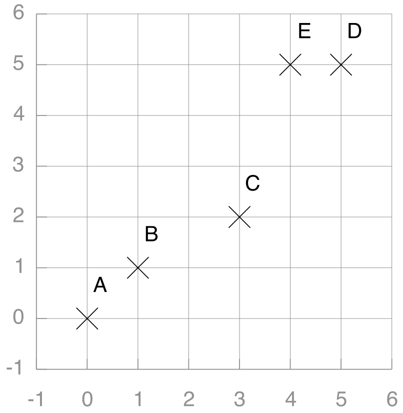


Figure 1: A clustering data set

7. **Performance Measures and Generalisation**

(a) *(5 points)* Given a binary classification problem with dataset D and two learning algorithms L_1 and L_2, describe how you would decide which algorithm to use.

(b) *(5 points)* Fitting a 14th degree polynomial to 15 data points using least squares regression will result in overfitting. For each of the following two cases, argue whether they would help against overfitting:

   i. Suppose that instead of minimising the sum of the squares of the errors, we would minimise the sum of the absolute values of the errors.

   ii. Suppose that instead of 15 data points we had 100 000.

(c) *(5 points)* For a binary classification problem, a classifier is evaluated on a testset of 1000 cases. There are 700 positive and 300 negative examples in the dataset. It correctly identifies 586 positive cases and 234 negative cases.

Give an estimate of the true accuracy of this classifier, and a 95% confidence interval around that estimate (you may give your confidence interval in the form of an expression). Note: $Z_{95} = 1.96$.

Machine Learning 2014-2015—Resit Exam

With Answers

9 June 2015, 18:30 – 21:15

Answers in bold italics. The exam is **open book**: you can use Flach's *Machine Learning* (as a print of the PDF or as a proper book) as well as the lecture slides and any notes you've taken. You can use a calculator as long as it doesn't provide (internet) connectivity.

The exam grade determines 50% of your final grade for the course, the other half is determined by your project grade. The exam grade must at least be 5.5 to pass the course.

Good luck!

Question	*Points*
1. Short answers	15
2. Decision Trees	20
3. Bayesian Learning	15
4. Neural Networks	15
5. Instance-based Learning	15
6. Classification and Performance Measures	20
Total	**100**

Questions

1. **Short answers** no justification required *(3 Points for each question)*

   (a) (**True** or **False**): Support Vector Machines do not need labelled data.

   false

   (b) (**True** or **False**): Decision trees can handle continuous attributes (e.g., age or income).

   true

   (c) (**True** or **False**): Neural networks with a sigmoid activation function can be used for classification problems.

   true

   (d) (**True** or **False**): Cross-validation can be used to set a learning algorithm's parameters (e.g., learning rate or number of hidden nodes in neural networks).

   true

   (e) (**True** or **False**): Clustering methods such as k-means try to minimise the mean squared error over the training data.

   false

2. **Decision Trees**

To improve your results when playing FIFA Football Manager, you decide to collect some data on recent player transfers and use that to train a decision tree. Table 1 contains the data.

Left- or Rightfooted	International	Position	Successful Transfer?
Right	Yes	Forward	Yes
Right	Yes	Midfield	Yes
Right	Yes	Defender	Yes
Right	Yes	Keeper	Yes
Right	No	Forward	No
Right	No	Midfield	No
Right	No	Defender	No
Left	Yes	Keeper	No
Left	No	Forward	Yes
Left	No	Midfield	Yes

Table 1: Players and Transfers

(a) *(5 points)* What are the entropies $H(Successfultransfer)$ and $H(Successfultransfer|Leftfooted)$? [1] Show your calculations.

0.971 and 0.918

(b) *(5 points)* What is the information gain for *Left- or Rightfooted*? Show your calculations.

0.0059

(c) *(5 points)* The ID3 algorithm would select *International* as the root node of the tree. What will the algorithm split on next when $International = Yes$? Show your calculations.

left- or rightfooted

(d) *(5 points)* Suppose that the decision tree is not elaborated further, i.e., there is no split other than the root (on *International*) and the one you identified in question 2c. Draw the tree and calculate its training set error.

3. **Bayesian Learning**

(a) *(5 points)* From the data in table 1, what is the prior probability of a successful transfer? Show your calculations.

60%

(b) *(5 points)* What would a naive Bayes classifier trained on that data predict for a transfer of a midfield player? Show your calculations.

TODO

(c) *(5 points)* Would there be a difference between an MLE and an MAP prediction for a Bayes classifier trained on the data in Table 1? Why (not)?

There is a difference because prior probability of successful transfer is 0.6

[1] Note: if you need a calculator and yours can't do log_2, use the following equation: $log_2(x) = 1.44 \cdot ln(x)$. If you didn't bring a calculator, you may give the answer as an expression.

4. **Neural Networks**

In this question, we consider a neural network with two input nodes x_1 and x_2, two hidden nodes h_1 and h_2 and one output node o. The nodes have sigmoid activation functions. The weights for the connections are as follows:

Connection	Weight
$x_1 \rightarrow h_1$	0.6
$x_1 \rightarrow h_2$	-0.2
$x_2 \rightarrow h_1$	0.1
$x_2 \rightarrow h_2$	-0.5
$h_1 \rightarrow o$	0.6
$h_2 \rightarrow o$	0.2

Table 2: Neural Network Weights

(a) *(5 points)* Draw the network with the edges labelled with the weights.

(b) *(5 points)* What is the output of the network if the input is $(0.6, 0.6)$? Show your calculations.

0.6086

(c) *(5 points)* Let's assume that for these inputs, the target value $t = 0.5$ and the learning rate $\eta = 0.1$. Compute the new weight for the connection $h_1 \rightarrow o$ using backpropagation. Show your calculations.

$h_1 \rightarrow o = 0.5984$

5. **Instance-based Learning**

Consider the following dataset with one real-valued input x and one real-valued output y. We are going to use k-nearest neighbour with unweighted Euclidean distance to predict y for a given x.

x	y
5	-0.96
6	-0.28
9	0.41
11	-1.00
15	0.65
16	-0.29
17	-0.96
18	-0.75
20	0.91

Table 3: kNN data set

(a) *(2.5 points)* What is the predicted outcome of 2-nearest neighbour for a new datapoint $x = 5.5$?

$\frac{(-0.96)+(-0.28)}{2} = 0.62$

(b) *(2.5 points)* What is the predicted outcome of 3-nearest neighbour for a new datapoint $x = 10$?

$\frac{(-0.28)+0.41+(-1.0)}{3} = -0.29$

(c) *(5 points)* Why does using a kernel function allow instance-based algorithms to consider all points instead of only the nearest neighbours?

(d) *(5 points)* Techniques such as decision trees are known as batch learners that require the availability of all training data to build their hypothesis. Thus the arrival of additional training data needs to be handled carefully. Does kNN suffer from this problem and why (not)? (This question moved from 4. Neural Networks, where it ended up by mistake in the original exam).

6. **Classification and Performance Measures**

(a) *(5 points)* In your own words, describe what is meant by the term *overfitting*. What can cause overfitting when training an artificial neural network on a data set? How can one avoid overfitting?

(b) *(5 points)* When selecting a machine learning technique for some problem, we can measure the performance on a test set to measure the performance of the techniques were comparing, or we can use cross-validation. Name one advantage of using cross-validation.

(c) *(5 points)* For a binary classification problem, a classifier is evaluated on a dataset of 750 cases. There are 202 positive and 548 negative examples in the dataset. The classifier correctly identifies 148 positive cases and 468 negative cases. Draw the confusion matrix that follows from these numbers.

(d) *(5 points)* Give the precision, accuracy and recall for this classifier. Show your calculations.

.649, .8213 and .733

	Predicted		
	pos	neg	total
actual pos	148	54	202
actual neg	80	468	548
total	228	522	750