Machine Learning 2014-2015—Final Exam

With Answers

27 March 2015, 15:15 – 18:15

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	20
5. Instance-based Learning	15
6. Classification and Performance Measures	15
Total	**100**

Questions

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

   (a) (**True** or **False**): As long as a decision tree only uses binary splits, it cannot overfit.
   False

   (b) (**True** or **False**): Clustering techniques like k-means require some labeled data to initialise the cluster centres.
   False

   (c) (**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.
   True. Stochastic gradient descent updates the weights after every pattern presentation. Batch learning only accumulates the weight changes until the whole batch of patterns has been presented once.

   (d) (**True** or **False**): Support Vector Machines can only model linear decision boundaries.
   False, the kernel trick allows non-linear decision boundaries.

   (e) (**True** or **False**): Cross Validation is a method that can help select proper settings for machine learning algorithms.
   True

2. **Decision Trees**

You're taking a date to the movies. To be sure that you both enjoy the night out, you choose to select a movie using a decision tree that you develop on data about movies you both watched. Table 1 contains the training data.

Genre	Oscar nominee	Lead	Liked by both?
Romantic Comedy	Yes	Male	No
Horror	Yes	Male	No
Romantic Comedy	No	Male	No
Arthouse	No	Male	No
Arthouse	No	Male	Yes
Romantic Comedy	No	Female	Yes
Horror	No	Male	Yes
Horror	Yes	Female	Yes

Table 1: Movies

(a) *(5 points)* What are the entropies $H(Liked|Genre = Arthouse)$ and $H(Liked|Lead = Female)$? [1]

1 and 0

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

Genre

(c) *(5 points)* What will the algorithm decide for the $Lead = Female$ branch?

No further split: both examples were liked, so predict Yes

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

The tree would misclassify 2 cases (One Horror and one Arthouse, both with male lead), which is 25% of the cases. Your answer might be different depending on your answer to question 2b.

3. **Bayesian Learning**

(a) *(5 points)* From the data in table 1, what is the prior probability of both you and your date liking a movie? Show your calculations.

50%

(b) *(5 points)* What would a naive Bayes classifier trained on that data predict for a romantic comedy with a male lead? Show your calculations.

$y_{predict} = maxarg(P(E = v|data)$**.**

For naive Bayes,

$P(Liked = Yes|Genre = Romantic\ Comedy, Lead = Male) = P(Liked = Yes) \cdot P(Genre = Romantic\ Comedy|Liked = Yes) \cdot P(Lead = Male|Liked = Yes) = \frac{4}{8} \cdot \frac{1}{4} \cdot \frac{2}{4} = 0.0625$

$P(Liked = No|Genre = Romantic\ Comedy, Lead = Male) = P(Liked = No) \cdot P(Genre = Romantic\ Comedy|Liked = No) \cdot P(ead = Male|Liked = No) = \frac{4}{8} \cdot \frac{2}{4} \cdot \frac{4}{4} = 0.25$

$0.25 > 0.0625 \rightarrow Liked = No$

[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.

(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)?
No difference because of equal priors

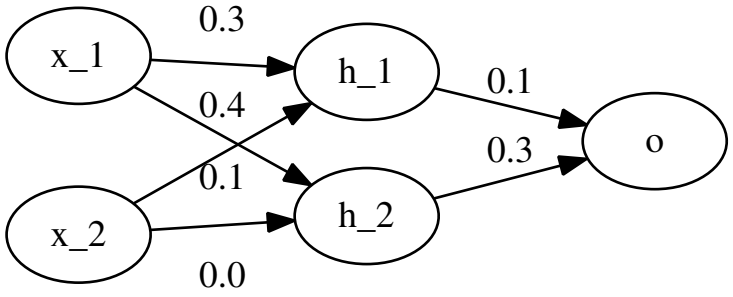
4. **Neural Networks**

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
$x_1 \to h_1$	0.3
$x_1 \to h_2$	0.4
$x_2 \to h_1$	0.1
$x_2 \to h_2$	0.0
$h_1 \to o$	0.1
$h_2 \to o$	0.3

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 $(0.5, 0.5)$? Show your calculations.

0.5548

(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 $h_1 \to o$ using backpropagation. Show your calculations.

$h_1 \to o = 0.0993$

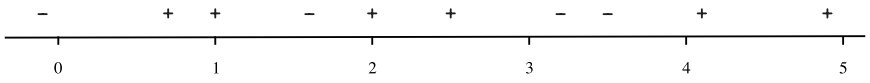
(d) *(5 points)* For a neural network, which one of the following structural assumptions most affects the trade-off between underfitting (i.e. low performance) and overfitting? Explain your reasoning.

i. The learning rate
ii. The initial choice of weights
iii. The number of hidden nodes
iv. The use of a constant-term unit input

4(d)iii

5. **Instance-based Learning**

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



The precise x values are as follows: $\{-0.1; 0.7; 1.0; 1.6; 2.0; 2.5; 3.2; 3.5; 4.1; 4.9\}$

(a) *(2.5 points)* What is the predicted class of 1-nearest neighbour for a new data-point $x = 0.0$?

-

(b) *(2.5 points)* What is the predicted class of 3-nearest neighbour for a new data-point $x = 0.0$?

+

(c) *(5 points)* What is the leave-one-out cross-validation [2] error of 1-nearest neighbour on this dataset? Give your answer as the number of misclassifications.

4

(d) *(5 points)* Now, we will use a simple uniform kernel function that weights all data points that are at a distance of max. 1 equally, i.e.:

$$k(d) = \begin{cases} 1, & \text{if d} \leq 1. \\ 0, & \text{otherwise.} \end{cases}$$

What is the predicted class of nearest neighbour with this kernel for a new data-point $x = 2.1$?

The kernel leads to including the neighbours at $\{1.6; 2.0; 2.5\}$***: two '+' and one '-', so: +***

6. **Classification and Performance Measures**

(a) *(5 points)* For a binary classification problem, a classifier is evaluated on a dataset of 1000 cases. There are 700 positive and 300 negative examples in the dataset. It correctly identifies 586 positive cases and 234 negative cases. Draw the confusion matrix that follows from these numbers.

	Predicted		
	pos	neg	total
actual pos	586	114	700
actual neg	66	234	300
total	652	348	1000

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

0.899, .82 and .837

	Model 1	**Model 2**	**Model 3**
fold 1	0.87	0.89	0.90
fold 2	0.78	0.85	0.86
fold 3	0.79	0.84	0.76
fold 4	0.82	0.9	0.89

Table 3: Cross-fold validation results

(c) *(5 points)* Three classifiers are compared using cross-fold validation: the results on each of the four hold-out samples is reported in table 3. Which model should be selected? Show your calculations.

The average over the folds for the three models is 0.815, 0.87 and 0.8525, respectively. So, Model 2 has the best average and should be selected.

[2]Remember: Leave-one-out cross-validation = cross-validation using each separate data-point as the hold-out sample once, i.e., first use $\langle -0.1, + \rangle$ as the test set, then $\langle 0.7, + \rangle$, and so on.