Machine Learning

Final Exam ***With Answers***

28 March 2017, 08:45 – 11:30

Answers in bold italics. The exam is *closed* book, but you may bring and use:

- One cheat sheet (i.e., 2 pages) of *handwritten* notes;
- A (graphical) 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. Your exam grade must at least be 5.5 to pass the course.
This exam consists of 40 multiple choice questions on 6 pages.
Good luck!

Questions

General

1. Imagine an application of machine learning where it is important that the user understands the model. Which of the following techniques is *least* applicable?
   A Decision Tree
   B 5-Nearest Neighbour
   C Neural Network
   D Linear Discriminant
   C

2. In another application, a model must provide real-time classification for a website that serves millions of visitors per day. Which kind of model is *least* applicable?
   A Naive Bayes Classifier
   B 5-Nearest Neighbour
   C Neural Network
   D Linear Discriminant
   B

Decision Trees

After this week of gruelling exams, you plan to relax with a good book. To decide what book to buy, you have compiled your experiences of books in table 1 on which you will build a decision tree.

x_1 (Genre)	x_2 (Recommendation)	x_3 (Part of a series)	y (Enjoyed the book)
Literary Fiction	No	Yes	Yes
Literary Fiction	No	No	No
Science Fiction	No	Yes	Yes
Science Fiction	Yes	Yes	Yes
Thriller	No	Yes	Yes
Thriller	No	No	No
Science Fiction	Yes	No	No
Thriller	Yes	Yes	No

Table 1: Books you read

3. What is the entropy $H(y)$?[1] **A** 1 **B** 0.5 **C** 0.25 **D** 0

   A

4. What is the entropy $H(y|x_2 = No)$? **A** -0.53 **B** -0.05 **C** 0.92 **D** 0.97

   D

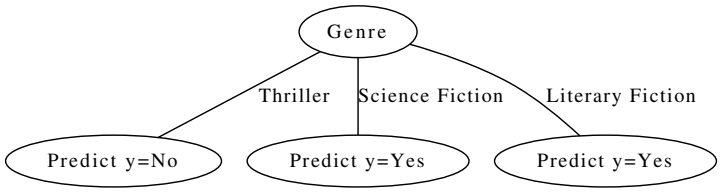
5. What is the information gain for attribute x_1? **A** 0.06 **B** 0.50 **C** 0.92 **D** 1.00

   A

6. Suppose that you decide to convert the decision stump in figure 1 to a scoring tree. Which ranking would best reflect the data in table 1?

	Literary Fiction	Science Fiction	Thriller
A	1	2	3
B	2	3	1
C	3	2	1
D	None of the above		

Figure 1: A decision stump



[1] you can look up the value for log_2 in this table:

x	0.1	0.2	0.25	0.33	0.375	0.40	0.5	0.6	0.625	0.66	0.75	0.8	1
$log_2(x)$	−3.32	−2.32	−2.00	−1.60	−1.42	−1.32	−1.00	−0.74	−0.68	−0.60	−0.42	−0.32	0.00

7. Why would you prune a decision tree developed with ID3?
   A To be able to use numeric attributes
   B To explain the model
   C To handle missing values
   D To reduce overfitting

Bayesian Learning

8. From the data in table 1, what is the prior probability of you liking a book?
   A 0.00 **B** 0.25 **C** 0.50 **D** 1.00

   C

9. What probability would a naive Bayes density estimator trained on table 1 for the $y = Yes$ class calculate for a case $(x_1 = Thriller, x_2 = Yes)$?
   A 0.00 **B** 0.06 **C** 0.50 **D** Cannot compute because there are no $(x_1 = Thriller, x_2 = Yes, y = Yes)$ examples

   B

10. What would a naive Bayes classifier trained on table 1 predict for a case $(x_1 = Thriller, x_2 = Yes)$?
    A $y = Yes$
    B $y = No$
    C Cannot compute without a value for x_3
    D Undecided (i.e., 50/50 chance)

    B

11. What does the underlying 'Naive' assumption in naive Bayes classifiers entail?
    A Any multi-class classification problem can be described as a combination of binary classification problems
    B The classes are linearly separable
    C Attribute values are independent
    D Attribute values are (approximately) normally distributed

    C

Instance-based Learning

12. Which of the following is a drawback of lazy learning?
    A It takes many iterations to train a model
    B It can only represent very simple relationships
    C It discards a lot of information
    D It is slow at query time

    D

13. What is the predicted class of 3-nearest neighbour trained on the data in table 2 for a new data-point $(x_1 = 5, x_2 = 9)$ using Euclidean distance? **A** + **B** − **C** undecided

    A

14. In k-nearest neighbour, you can scale the attributes. This allows you to:
    A Regulate the importance of attributes
    B Handle large amounts of training data
    C Use symbolic attributes
    D Reduce overfitting

x_1	x_2	y
9	9	-
9	6	-
5	6	+
3	6	+

Table 2: 3-NN training data

A

15. Increasing the value of k in k-nearest neighbour implies:
    A The use of a kernel function
    B Smoothing of noise
    C Distance weighting
    D Faster convergence

16. When using the same base learner (e.g., linear regression) in an instance-based setting instead of as a global model, the model complexity:
    A Increases because it implies many local models instead of a single one
    B Reduces because the models are based on small samples of data
    C Does not change because the base learner remains the same
    D None of the above

Neural Networks and Gradient Descent

17. 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. All nodes have sigmoid activation functions. The weights for the connections are as follows:

Connection	Weight
$x_1 \to h_1$	0.5
$x_1 \to h_2$	-0.9
$x_2 \to h_1$	0.1
$x_2 \to h_2$	-0.7
$h_1 \to o$	0.8
$h_2 \to o$	0.1

Table 3: Neural Network Weights

Oh, this is not actually a question...

18. What is the output of hidden node h_1 if the input is $(x_1 = 0.1, x_2 = 0.5)$?
    A 0.10 **B** 0.52 **C** 0.60 **D** 1.20

B

19. What is the output of the output node o if the input is $(x_1 = 0.1, x_2 = 0.5)$?
    A 0.04 **B** 0.46 **C** 0.51 **D** 0.61

D

20. Assume that for these inputs $(x_1 = 0.1, x_2 = 0.5)$, the target value $t = 1.00$ and the learning rate $\eta = 0.1$. Compute the new weight for the connection $h_2 \to o$ using back-propagation.
    A 0.0964 **B** 0.0991 **C** 0.1036 **D** 0.1387

C

21. How can neural networks handle a classification problem where the classes are not linearly separable?
    A By adding an extra output
    B By adding a 'bias' input node
    C By adding a hidden layer
    D They cannot
    C

22. When training a neural network for a regression task, you see that the error rate decreases very, very slowly. Does this imply that
    A The learning rate is set too high
    B The learning rate is set too low
    C You should add momentum
    D The problem is not linearly separable
    B

23. How can you be certain that a neural network has converged to a global optimum?
    A Gradient descent is guaranteed converge to a global optimum
    B By using cross-validation
    C If a second run with a larger learning rate achieves the same error rate.
    D You cannot
    D

Linear Discriminants and Support Vector Machines (SVMs)

24. What is the margin of a linear classifier?
    A The area under the separating hyperplane
    B The region around the separating hyperplane that does not contain any data points
    C Any data points on the separating hyperplane
    D The set of data points that are misclassified
    B

25. What is the difference between discriminant-based and generative classifiers?
    A Generative classifiers cannot use symbolic attributes, but discriminant-based classifiers can
    B Generative classifiers can be trained with unlabelled data
    C Discriminant-based can only model linear relationships, generative classifiers can model any relationship
    D Generative classifiers base their prediction on a model of the joint probability of the inputs x and the label y, discriminant-based classifiers model learn a direct map from inputs x to the class label y.
    D

26. What is the role of the kernel in the kernel trick for support vector machines?
    A Dimension reduction to handle large number of attributes
    B It redefines distance calculation and so allows for non-linear discriminant
    C It provides a weighting scheme to ignore irrelevant attributes
    D Weighted sampling of cases to handle uneven distribution of outcome classes
    B

Evaluating Hypotheses

27. For a binary classification problem, a classifier is evaluated on a dataset D of 2,000 cases. There are 900 positive and 1,100 negative examples in the dataset. The classifier correctly

identifies 684 positive cases and 958 negative cases. What is the accuracy of this classifier?

A 18% **B** 45% **C** 76% **D** 82%

D

28. What is the recall of the classifier from question 27 on that same dataset D?

A 0.17 **B** 0.34 **C** 0.76 **D** 0.83

C

29. Four classifiers are compared using cross-fold validation: the accuracy on each of the four hold-out samples is reported in table 4. Which model should be selected to achieve maximum performance? **A** Model 1 **B** Model 2 **C** Model 3 **D** Model 4

D

	Model 1	**Model 2**	**Model 3**	**Model 4**
fold 1	0.68	0.71	0.69	0.68
fold 2	0.67	0.66	0.61	0.70
fold 3	0.75	0.72	0.78	0.70
fold 4	0.70	0.70	0.70	0.74

Table 4: Cross-fold validation results

30. Someone has developed a classifier that correctly classifies 300 out of 400 test examples. Give an estimate of the true accuracy of this classifier, and a 95% confidence interval around that estimate. Note that the z-value for 95% is 1.96.

A 0.25 ± 0.01 **B** 0.75 ± 0.01 **C** 0.25 ± 0.04 **D** 0.75 ± 0.04

D

31. What is an advantage of 20-fold cross-validation over 3-fold cross validation?
    A No benefits
    B It requires less training of models
    C It reduces variance of the estimated performance
    D It achieves better training set performance

32. Which of the following gives the most accurate unbiased estimate of model performance?
    A Training set error
    B Average error on bootstrap samples from training set
    C Test set error
    D All are equally accurate

Bias, Variance and Ensemble Models

33. High variance is typically associated with:

A Underfitting **B** Overfitting **C** Classification **D** Regression

B

34. Bagging combines multiple base learners by
    A Training each model on different subsamples of the data
    B Training each model on a separate set of attributes
    C Training each model to focus on the errors of previous models
    D Training a separate model to combine simpler base models

A

35. The effect of boosting is that it...

    A Reduces variance **B** Increases variance **C** Reduces bias **D** Increases bias

    C

36. True (**A**) or False (**B**): A model with high variance gives a wider range of results over subsamples of a dataset than does a low variance model

    A

37. True (**A**) or False (**B**): Simple models typically have lower bias and higher variance

    B

38. Random forest is...
    A A bagging method
    B A boosting method
    C A combination of regression and classification for ordered but symbolic labels
    D Not an ensemble method
    A

Expectation Maximisation (EM) and Clustering

39. K-means clustering...
    A Determines the optimal value for k using cross-validation
    B Results in different clusterings depending on the initial configuration of centroids
    C Uses kernels to (re)define the distance between two cases
    D Requires labelled records

40. Single Linkage Hierarchical Clustering iteratively combines clusters...
    A That contain the fewest data points
    B That have minimal entropy
    C That are most similar
    D That have maximal within-cluster distance