# Machine Learning 2015-2016

## Final Exam ***With Answers***

22 March 2015, 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. The exam grade must at least be 5.5 to pass the course. The exam consists of 40 multiple choice questions on 6 pages.
**Good luck!**

## Questions

### Decision Trees

With the recent excitement of AI beating the world champion at Go, you decide to develop a model to classify games at which AI will beat the world champion within five years. Table 1 contains the training data you have collected.

| $x_1$ (Team or Individual) | $x_2$ (Mental or Physical) | $x_3$ (Skill or Chance) | $y$ (Win or Lose) |
|---|---|---|---|
| T | M | S | W |
| I | M | S | W |
| T | P | S | W |
| I | M | C | W |
| T | P | S | L |
| I | M | C | L |
| T | P | C | L |
| T | P | C | L |
| T | P | C | L |
| I | P | S | W |

Table 1: Games that AI can win

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

   ***A***

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

| $x$ | 0.10 | 0.25 | 0.33 | 0.50 | 0.66 | 0.75 | 1.00 |
|---|---|---|---|---|---|---|---|
| $log_2(x)$ | -3.32 | -2.00 | -1.60 | -1.00 | -0.60 | -0.42 | 0.00 |

2. What is the entropy $H(y|x_1 = T)$? **A** -0.39 **B** -0.53 **C** 0.33 **D** 0.92

***D***

3. What is the information gain for attribute $x_2$? **A** 0.08 **B** 0.12 **C** 0.14 **D** 0.19

***B***

4. Which attribute would ID3 select as the root for the decision tree?

**A** $x_1$ **B** $x_2$ **C** $x_3$

***C***

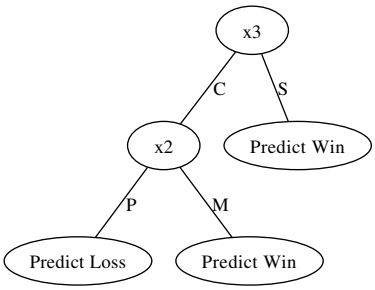



Figure 1: A decision tree

5. Suppose that you develop the tree in fig. 1 (not necessarily using ID3). What is the training set error for this tree on the data in table 1? **A** 0% **B** 10% **C** 20% **D** 30%

***C***

6. True (**A**) or False (**B**): Because decision trees learn to classify using only discrete-valued variables, they cannot overfit.

***False***

## Bayesian Learning

7. From the data in table 1, what is the prior probability of AI winning a game?

**A** 0.25 **B** 0.45 **C** 0.50 **D** 0.66

***C***

8. What would a naive Bayes density estimator trained on table 1 for the $Win$ class predict for a case $(x_1 = I, x_3 = C)$? **A** 0.20 **B** 0.48 **C** 0.50 **D** 0.60

***Oops. It's*** $\frac{3}{5} \cdot \frac{1}{5} = 0.12$***. I'll ignore this question.***

9. What would a naive Bayes classifier trained on table 1 predict for a case $(x_1 = I, x_2 = P, x_3 = C)$?
   **A** $Win$
   **B** $Lose$
   **C** Cannot compute because there are no $(x_1 = I, x_2 = P, x_3 = C)$ examples
   **D** Undecided (i.e., 50/50 chance)

   ***B***

10. What would a Joint Density Bayes classifier trained on table 1 predict for a case $(x_1 = I, x_2 = P, x_3 = C)$?
    **A** $Win$
    **B** $Lose$

**C** Cannot compute because there are no $(x_1 = I, x_2 = P, x_3 = C)$ examples
**D** Undecided (i.e., 50/50 chance)

***C***

11. Suppose that we know (from another source than the data in table 1) that the frequency of $Win$ is actually four times that of $Lose$. How would the naive Bayes Classifier decide now for a case $(x_1 = I, x_2 = P, x_3 = C)$?
**A** $Win$
**B** $Lose$
**C** Cannot compute because there are no $(x_1 = I, x_2 = P, x_3 = C)$ examples
**D** Undecided (i.e., 50/50 chance)

***A***

## Instance-based Learning

In this question, we'll consider the data in figure 2.

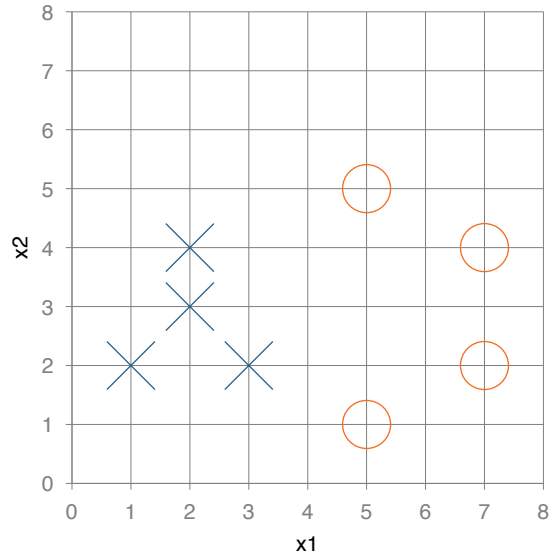



Figure 2: Classification data

12. What is the predicted class of 3-nearest neighbour for a new data-point $(x_1 = 4, x_2 = 4)$ using Euclidean distance? **A** Cross (×) **B** Circle (◯)

    ***A***

13. Now, we use $\frac{1}{d}$ as a kernel function. To simplify our calculations, we'll use Manhattan Distance[2] as the distance metric. What is the classification for the data-point $(x_1 = 4, x_2 = 4)$ now? **A** Cross (×) **B** Circle (◯)

    ***A***

14. How can we do nearest neighbour modelling with binary input vectors?
    **A** You can't
    **B** Any variant will do if we define an appropriate distance measure
    **C** Any variant with a kernel function will do
    **D** Any variant that uses averaging as the local model will do

    ***B***

15. When making a local model, we base our prediction on the examples in the neighbourhood by taking the average value, fitting a line or fitting a higher-degree polynomial. How can we choose between these options?
    **A** Select the model with the lowest variance
    **B** Select the model with the lowest bias
    **C** Use cross-validation to select the model
    **D** Don't choose any of these, but use a kernel instead

    ***C***

[2]Manhattan Distance defines the distance between two points in a grid based on a strictly horizontal and/or vertical path (that is, along the grid lines). The Manhattan distance is the simple sum of the horizontal and vertical components. Thus, the distance between two points $(1, 1)$ and $(2, 3)$ is: $(2 - 1) + (3 - 1) = 3$.

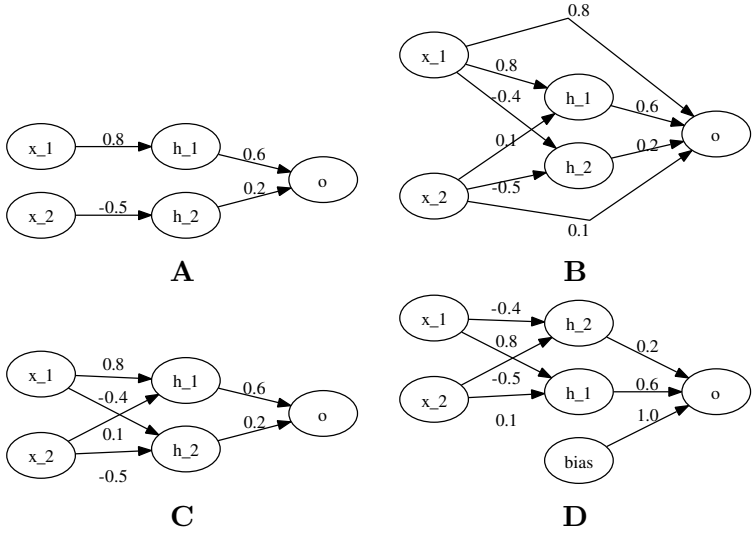
## Neural Networks and Gradient Descent

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

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

Table 2: Neural Network Weights

16. Which of the following graphs depicts the network described here?




***C***

17. What is the output of hidden node $h_1$ if the input is $(x_1 = 0.1, x_2 = 0.2)$?
**A** 0.10 **B** 0.30 **C** 0.53 **D** 0.67

***C***

18. What is the output of the output node $o$ if the input is $(x_1 = 0.1, x_2 = 0.2)$?
**A** 0.17 **B** 0.41 **C** 0.54 **D** 0.60

***D***

19. For a neural network, which one of the following structural assumptions most affects the trade-off between underfitting and overfitting?
**A** The number of hidden nodes
**B** The learning rate $\eta$
**C** The initial choice of weights
**D** The use of a constant-term unit input

***A***

20. True (**A**) or False (**B**): The main reason to add momentum in gradient descent is to avoid overfitting.

***False***

21. True (**A**) or False (**B**): If the learning rate $\eta$ in gradient descent is set too low, convergence will take longer than necessary, as reaching the correct value takes more time.

    ***True***

## Support Vector Machines (SVMs)

22. Which of the following makes most sense as a filtering algorithm to find training instances that are very unlikely to be support vectors?
    **A** Remove instances of each that are close to instances of other classes
    **B** Remove any instances that are surrounded by a large number of instances all of the same class
    **C** Remove instances with extreme attribute values
    **D** Remove instances with average attribute values

    ***B***

23. Consider the data in figure 2. How many instances in that data are support vectors for a linear maximum margin classifier?

    **A** 2 **B** 3 **C** 4 **D** 8

    ***B***

24. For the same data, what is the maximum margin linear model?
    **A** $x_1 + x_2 = 0$
    **B** $x_1 + x_2 = 4$
    **C** $2 \cdot x_1 + x_2 = 2$
    **D** $x_1 = 4$

    ***D***

25. True (**A**) or False (**B**): The use of kernels allows SVMs to model non-linear class boundaries.

    ***True***

## Evaluating Hypotheses

26. For a binary classification problem, a classifier is evaluated on a dataset of 1500 cases. There are 900 positive and 600 negative examples in the dataset. The classifier correctly identifies 621 positive cases and 502 negative cases. What is the accuracy of this classifier?

    **A** 0.41 **B** 0.67 **C** 0.75 **D** 0.86

    ***C***

27. What is the precision of the classifier from question 26?

    **A** 0.16 **B** 0.31 **C** 0.75 **D** 0.86

    ***D***

28. What is the recall of the classifier from question 26?

    **A** 0.16 **B** 0.56 **C** 0.69 **D** 0.83

    ***C***

29. Four 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?

    **A** Model 1 **B** Model 2 **C** Model 3 **D** Model 4

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| **fold 1** | 0.86 | 0.91 | 0.89 | 0.95 |
| **fold 2** | 0.94 | 0.85 | 0.86 | 0.81 |
| **fold 3** | 0.84 | 0.83 | 0.88 | 0.85 |
| **fold 4** | 0.82 | 0.87 | 0.89 | 0.83 |

Table 3: Cross-fold validation results

***C (or D if you thought the table listed errors, not accuracy. Both are counted as correct.***

30. Someone has developed a classifier that correctly classifies 750 out of 1000 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** $0.25 \pm 0.01$ **B** $0.75 \pm 0.01$ **C** $0.25 \pm 0.03$ **D** $0.75 \pm 0.03$

    ***D***

31. True (**A**) or False (**B**): When there is not enough data to split of a test set, you can still create an unbiased estimate of a model's performance by calculating a confidence interval around the training set performance.

    ***False***

## Bias, Variance and Ensemble Models

32. High bias is typically associated with:

    **A** underfitting **B** overfitting **C** Classification **D** Regression

    ***A***

33. Assume you have a small dataset from which a model has to be generated. What kind of model would you prefer?
    **A** low bias, low variance
    **B** high bias, low variance
    **C** low bias, low variance
    **D** high bias, high variance

    ***B***

34. Boosting 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

    ***C***

35. When using decision trees in the Random Forest approach, you
    **A** Add weighting to focus on difficult cases
    **B** Combine them with linear models for the leave nodes in the tree
    **C** Prune the trees to prevent overfitting
    **D** Do not prune the trees to obtain more unstable models

    ***D***

36. What is the difference between voting and stacking when training a linear perceptron as the combiner function?
    **A** None
    **B** The perceptron would add different weights to the base learners' votes
    **C** The variance would increase
    **D** The bias would increase

    ***B***

## Expectation Maximisation (EM) and Clustering

37. The EM procedure to configure Gaussian Mixture Models can be characterised as:
    **A** Determining the optimal number of clusters by gradient descent
    **B** Sequentially positioning the cluster centres by gradient descent
    **C** Alternately calculating the expectation for the current configuration and updating the configuration to reflect the expectation
    **D** Adding clusters until the within-cluster distance is minimal

    ***C***

38. EM uses a mixture of Gaussians for clustering; what purpose do the Gaussians serve?
    **A** Gaussians provide the confidence interval for the clustering accuracy
    **B** Each Gaussian models one cluster
    **C** The Gaussians approximate a binomial distribution
    **D** The Gaussians are kernels to weight distances

    ***B***

39. True (**A**) or False (**B**): Silhouettes provide a measure that we can use to choose an appropriate value of $k$ when applying k-means.

    ***true***

40. True (**A**) or False (**B**): Hierarchical clustering methods do not require a distance metric because they consider pairs of clusters, not individual cases.

    ***False***