Stellenbosch University

Computer Science Division
RW 441/741 (Machine Learning)
Exam: October 2024

Totaal / *Total*: [320]
Tyd / *Time*: 3-day take-away (28-30 October)

Internal Examiners: Prof. A.P. Engelbrecht
Moderator: Dr. M Ngxande
External Examiner: Prof. L. Leenen (University of Western Cape)

Instructions

1. Answer all the questions of this paper.
2. You are allowed to consult any literature, provided that you include proper citations to such literature.
3. You are not allowed to make use of any Artificial Intelligence tools.
4. You are not allowed to discuss the questions with anyone.
5. You have to submit your own work.
6. Submit your **typed** answers on or before 30 October, 23:59, via SunLearn, not via email. All submissions have to be in pdf format. Please do not submit in any other format, and make sure that your pdf document does not contain any errors.
7. Give answers in your own words. Do not copy from any other text. Marks will be given for insight and interpretation, not just facts. **Give motivations for all your answers**.
8. Name your file `????????RWx41exam.pdf`, where the question marks are replaced with your SU number. Replace `x` with 4 if you are registered for RW441 and with 7 if you are registered for RW741. Also provide at least your student number in the header of the first page of your pdf document.
9. You may send me email for clarifications on questions.

Section A: Introduction to Machine Learning

Total: [49]

1. Explain what you understand about Ockham's razor, and discuss how his principle applies to predictive models. (5)

2. Provide a clear and concise definition of the bias-variance dilemma. In your definition, define the meaning of the two terms `bias` and `variance`. (6)

3. Underfitting and overfitting are two important concepts in supervised learning.

   (a) Provide a clear definition of underfitting (2)

   (b) Provide a clear definition of overfitting (2)

   (c) Discuss underfitting and overfitting with reference to the bias-variance dilemma (2)

4. Criticize the following statement: In all situations where a numeric feature has missing values, the best approach to imputation is to replace the missing value with the mean over the available values of that feature. (5)

5. Discuss why feature selection is important for machine learning algorithms. (5)

6. Data used to construct predictive models can be either stationary or non-stationary.

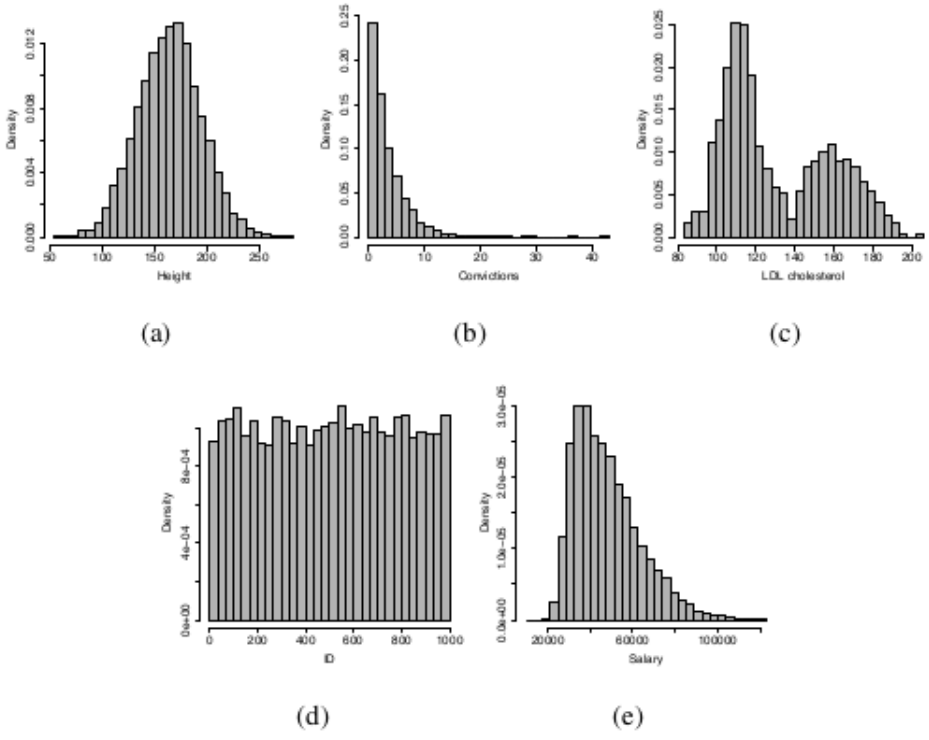
   (a) What do we mean by data being non-stationary compared to stationary data? (2)

   (b) Give an example of a real-world problem that exhibits non-stationary data. (1)

   (c) What are the main implications of non-stationary data for the construction and training of predictive models? (3)

7. When binning is used to convert a numeric feature into a categorical feature, what are the consequences of too few bins versus too many bins? (5)

8. Consider the following distributions for some arbitrary features, and answer the questions that follow.



   (a) Which of these distributions may indicate a feature with outliers? Motivate your answer. (3)

   (b) Which one of these distributions indicate a feature that should be considered for removal? Motivate your answer. (2)

   (c) For which of the features will mean imputation not be sensible? (5)

Section B: Information-based Learning

Totaal / *Total* [50]

1. Consider a data set that has 35% missing values for one of the categorical descriptive features. The data set has 40000 instances. Which of the following approaches should be followed with reference to this descriptive feature if a C4.5 classification tree is induced on this data set? Simply write down the letter(s) of the correct approach(es): (2)

   (a) Remove the descriptive feature

   (b) Impute the missing values using the mean or median over the available values

   (c) Impute the missing values using the mode over all the available values

   (d) Apply complete case analysis, and remove all of the instances with missing values for this feature

   (e) Nothing needs to be done with these missing values

2. The figure below shows eight numbers.

   (a) What is the entropy of the letters in this set? Show all your calculations. (5)

   (b) What would be the reduction in entropy (i.e. information gain) if these letters are split into two sets, one containing the even numbers and the other containing the odd numbers? Show all your calculations. (5)

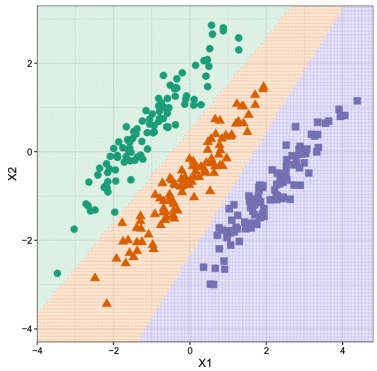
3	2	4	6	5	8	7	6

3. Consider the data set given below, and assume that information gain is used to decide on splits. Which descriptive feature will be split upon in the root of the classification tree? Show all of your calculations. Note that the last column represents the target feature. (10)

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

4. Classification trees induced using information gain have four inductive biases. Name these biases, discuss the consequences of each, and how these biases can be addressed. (12)

5. For the classification problem illustrated below, which of a standard classification tree induction algorithm (such as C4.5 or ID3) or an oblique tree induction algorithm will result in the smallest tree structure? Motivate your answer. (3)



6. Are regression trees robust to outliers with respect to the target feature? Motivate your answer. In your motivation consider the two splitting conditions:

(a) Split while the number of instances in the resulting subsets is larger than a user specified threshold.

(b) Split until the mean squared error for the resulting subsets have reached a specified error threshold.

(6)

7. Are model trees robust to target features? Only state aspects that have not been stated for regression trees above in question 6. Motivate your answer. (3)

8. Consider a classification tree induction algorithm that stops splitting of nodes when the number of instances in the data subsets are equal to or less than a specified thershold. Discuss advantages and disadvantages of this approach to stopping classification tree induction. (4)

Section C: Similarity-based Learning

Totaal / *Total* [29]

1. Consider a data set that has 5% missing values for one of the categorical descriptive features. The data set has 40000 instances. Which of the following approaches should be followed to deal with missing values if a k-nearest neighbor algorithm is used? Assume that a categorical distance measure is used. Write down the letter(s) of the correct approach(es): (4)

   (a) Apply complete case analysis and remove all instances with a missing value for this feature

   (b) Simply ignore the feature when the similarity measure is applied to find nearest neighbors

   (c) Use mean imputation to replace the missing value

   (d) Remove the feature

2. A k-nearest neighbour algorithm is used to develop a classifier. The data set contains noise. Which of the following strategies will help to reduce sensitivity to noise? Simply write down the letter(s) of the correct approach(es). Note that incorrect answers will be penalized. (4)

   (a) Reduce the value of k

   (b) Use a larger value for k

   (c) Find all Tomek links and remove both instances that form the Tomek link

   (d) Nothing can be done to reduce sensitivity to noise

   (e) Use SMOTE to oversample the minority class

3. Consider a regression problem, where the data set has the following characteristics:

   - There are 185 instances
   - There is one categorical and five numeric descriptive features
   - The categorical feature has three possible values. One of these values occurs for 70% of the instances, and the other two values occur respectively for 13% and 17% instances
   - Four of the numeric features have values in the range $[0, 1]$, and the fifth numeric feature has values in the range $[100, 100000]$
   - For 1% of the instances there are outliers for the target feature
   - One of the numeric descriptive features has a few outliers
   - One of the numeric descriptive features has 3% missing values

   A k-nearest neighbour algorithm is used and Euclidean distance is used as the similarity measure. Which of the following statements are correct? Simply write down the letter(s) of the correct statement(s). Be careful; marks will be subtracted for incorrect answers. (6)

   (a) The value for k has to be large

   (b) One-hot encoding has to be applied to the categorical feature

   (c) The numerical features have to be scaled to the same range, e.g. $[0, 10]$

   (d) The outliers in the numeric descriptive feature have to be removed

   (e) The missing values have to be imputed

(f) Under-sampling or over-sampling has to be applied to the categorical feature to balance the distribution of possible values

(g) The predicted value is calculated using the average over the target values of the neighbors

4. What is the inductive bias of the k-nearest neighbour algorithm? (2)

5. Is it necessary to normalize input features when a k-nearest neighbour algorithm is used? Motivate your answer. (3)

6. Explain how k-nearest neighbours can be used to impute missing values. (3)

7. Can the k-nearest neighbour algorithm be applied to problems with categorical descriptive features? Motivate your answer. (2)

8. Discuss the consequences of different values for k when k-nearest neigbours is applied to regression problems. (5)

Section D: Error-based Learning

Totaal / *Total* [49]

Note that any reference to neural networks in this section is to shallow networks, and not deep networks.

1. Consider a data set that has 5% missing values for one of the numerical descriptive features. The data set has 3200 instances, and there are also outliers for this feature. These outliers have not been removed. Which of the following approaches should be followed with reference to this descriptive feature if a neural network is trained on this data set? Just provide the letter(s) of the correct answer(s). (2)

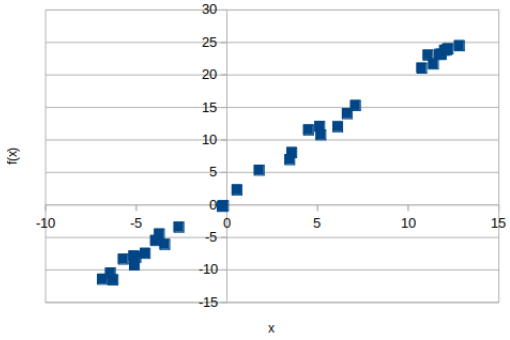
   (a) Remove the feature from the data set

   (b) Impute using the mean over all of the available values

   (c) Impute using the median over all available values

   (d) Do nothing

2. Consider the plot below. Which of the following approaches is best suited to develop a predictive model for this regression problem? Just provide the letter(s) of the correct answer(s). (2)



   (a) Logistic regression

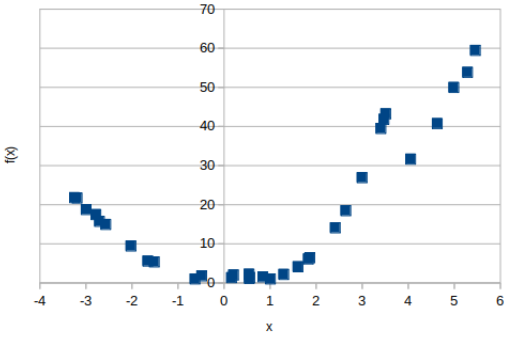
   (b) A neural network with one input unit, two hidden units and one output unit

   (c) A regression tree

   (d) Linear regression

   (e) Model tree

3. Consider the plot below. Which of the following approaches is best suited to develop a predictive model for this regression problem? Just provide the letter(s) of the correct answer(s). (2)



   (a) Any recurrent neural network
   (b) Polynomial regression using a second order polynomial kernel
   (c) Polynomial regression using a third order polynomial kernel
   (d) A neural network with one input unit, three hidden units and one output unit
   (e) Logistic regression
   (f) A regression tree
   (g) Linear regression

4. Consider a data set for a classification problem. The problem has three decision boundaries, there are five descriptive features and two class values. A neural network has been trained to separate the two classes, and contains enough input units, one output unit, and a hidden layer with five hidden units. The data set also contains noise, and has 1500 instances. After training the neural network, the prediction accuracy on the training set is 95% and the prediction accuracy on the test set is 67%. Which of the following statements are true for this predictive model? Incorrect answers will be penalized. Write down only the letter(s) of your answer(s). (8)

   (a) The model overfits the training set
   (b) Accuracy on the test set can be improved by reducing the number of hidden units
   (c) Accuracy on the test set can be improved by increasing the number of hidden units
   (d) Accuracy will be improved if the learning rate is increased
   (e) Accuracy can be improved by using an active learning approach
   (f) Accuracy on the test set will improve if the network is trained for longer
   (g) Accuracy on the test set may improve if regularization is applied

5. Suppose that the sigmoid activation function is used in the hidden and output layers of the neural network, and that gradient descent is used to adjust the weights.

   (a) Is normalization/scaling of the target feature necessary? Motivate your answer. (3)
   (b) Will it be prudent to normalize/scale the input features? Motivate your answer. (4)
   (c) Why is it not a good strategy to initialize weights and biases to large absolute values? (3)

6. Assume again that gradient descent is used, explain why a momentum term is necessary for stochastic learning. (4)

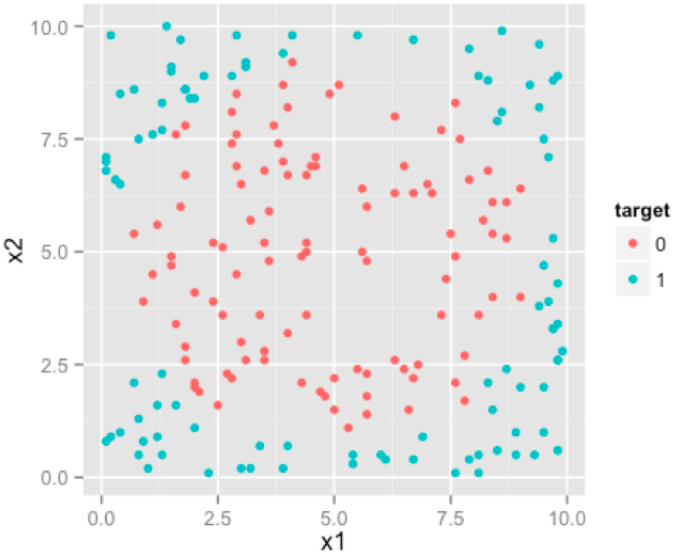
7. Is the following statement true or false: A single artificial neuron can only separate linearly separable classes. Motivate your answer. (4)

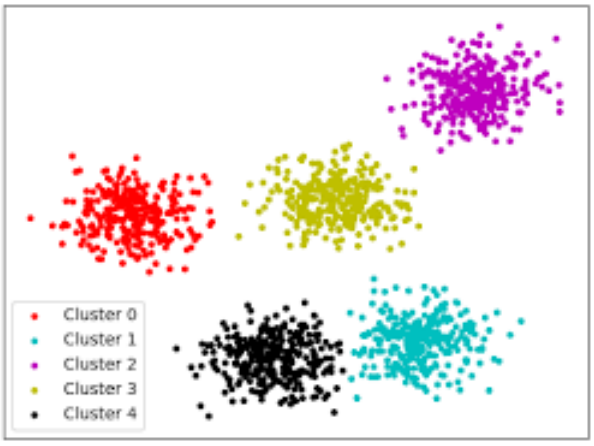
8. Compared to gradient descent, what are the main advantages of using scaled conjugate gradient to train a neural network? (2)

9. What are the main advantages in using population-based meta-heuristics to train neural networks? (3)

10. What is the rationale behind active learning in neural networks? (5)

11. Consider the classification problem depicted in the figure below. Can logistic regression be used to separate the two classes? Motivate your answer. (4)





12. Consider the classification problem depicted in the figure below. Explain in detail how logistic regression can be used to separate the classes. (3)

13. Are neural networks that make use of bounded activation functions robust to scale differences in the values of descriptive features used as input to the neural network? (3)

14. Consider classification problems, and explain why adaptive steepness of activation functions in the hidden units are of benefit. (3)

15. Assume that a neural network is used to approximate a highly non-linear mapping using one hidden layer. To achieve the same generalization performance, which of a linear activation function or sigmoid activation function will result in less hidden units? Motivate your answer. (3)

Section E: Unsupervised Learning

Totaal / *Total* [51]

1. Which of the following statements are false with reference to self-organizing feature maps? Incorrect answers will be penalized. Write down only the letter(s) of your answer(s) (6)

   (a) Training minimizes the mean-squared error

   (b) Self-organizing feature maps use a competitive training approach

   (c) They are robust to skew class distributions

(d) The best approach to weight initialization is to initialize the weights by sampling weight values from a uniform distribution in the range [-0.1,0.1] to ensure that initial weights have small values

(e) Training is a computationally expensive process

(f) Batch training will reduce the computational cost

(g) Can be used for regression problems

2. Explain how a self-organizing map (SOM) can be used to profile customers of a medical aid company. (5)

3. Training of a SOM is computationally expensive. Discuss three approaches to reduce the computational cost of training a SOM. (9)

4. Can a SOM be used as a classifier? Motivate your answer. (4)

5. Explain how a SOM can be used for data imputation. (4)

6. Unsupervised learning algorithms do not use target output values. So, what does unsupervised learning learn? (2)

7. You are using a self-organizing feature map to develop a recommender system for Netflix. Only descriptive features characterising movies are used. Assume that the self-organizing feature map has been trained on a large number of movies.

   (a) After watching one movie, how can the trained map be used to recommend the next movie? (2)

   (b) After having watched a large number of movies, how can the trained map be used to build a profile of the types of movies watched? (3)

   (c) Growing self-organized maps are an approach to find optimal map sizes. Assume that a square map has to be maintained, answer the following questions:

      i. Which node should be selected to be split via addition of an additional row and column next to that node? Justify your answer. (2)

      ii. What is the best choice for the neighbouring neuron inbetween which the new row or column should be inserted? Justify your answer. (2)

      iii. When the new row or column is added and weights initialized for the nodes in the row or column, should interpolation between the two existing neurons bias towards that neuron with the lowest quantization error or the highest quantization error. Justify your answer. (2)

Section F: Kernel-based Learning

Totaal / *Total* [30]

1. What is the inductive bias of linear support vector machines (SVMs)? (2)

2. Discuss the rationale of the maximum margin approach of support vector machine training. (3)

3. What is the consequence of noise for linear SVMs? (2)

4. How are linear SVMs adapted to cope with noise? (2)

5. What are the advantages and disadvantages of this approach to cope with noise? (4)

6. Are SVMs scalable to large data sets? Motivate your answer. (4)

7. How do kernels help to apply SVMs to linearly non-separable problems? (3)

8. Discuss the similarities between radial basis function neural networks and Gaussian mixture models. (4)

9. Is support vector regression robust to target outliers? Motivate your answer. (3)

10. Is support vector regression robust to noise? Motivate your answer. (3)

Section G: Ensemble Learning

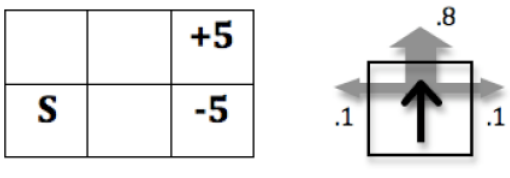
Totaal / *Total* [35]

1. Select all statements that support advantages of random forests over individual decision trees. Incorrect answers will be penalized. Write down the letter(s) of your answer(s). (4)
   (a) A random forest is easier to train in terms of computational resources
   (b) A random forest is an ensemble learning approach
   (c) With reference to the bias-variance dilemma, random forests result in larger bias
   (d) Random forests overfit less
2. What is the main rationale behind ensemble learning? (2)
3. Discuss why heterogeneous mixtures of experts are expected to perform better than homogeneous mixtures of experts. (4)
4. Why is weighted voting used in AdaBoost, instead of allowing each expert in the boosting pipeline to have the same weight? (4)
5. Consider an ensemble of k-NN algorithms. Why does it make sense to include k-NN experts with different values for k? (5)
6. How do random forests ensure that the different decision trees in the forest exhibit different behavior? (4)
7. Are regression random forests sensitive to outliers? Motivate your answer. (4)
8. Is adaptive boosting robust to skew class distributions? Motivate your answer. (3)
9. Why does bagging ensembles result in reduced variance? (2)
10. Why should boosting ensembles make use of weak learners? (3)

Section H: Reinforcement Learning

Totaal / *Total* [27]

1. What is Q-learning? (2)
2. Consider the following Markov decision process, with the gridworld and transition function as illustrated below. The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1), marked with the letter S. There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (North, South, West, or East) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.



(a) Gridworld MDP. (b) Transition function.

   (a) Draw the optimal policy for this grid. (5)
   (b) Suppose the agent knows the transition probabilities. Give the first two rounds of value iteration updates for each state, with a discount of 0.9. (Assume V_0 is 0 everywhere and compute V_i for times $i = 1, 2$). (8)

(c) Suppose the agent does not know the transition probabilities. What does it need (or must it have available) in order to learn the optimal policy? (3)

(d) The agent starts with the policy that always chooses to go right, and executes the following three trials:

1) (1,1)-(1,2)-(1,3),
2) (1,1)-(1,2)-(2,2)-(2,3), and
3) (1,1)-(2,1)-(2,2)-(2,3).

What are the monte carlo (direct utility) estimates for states (1,1) and (2,2), given these traces? (4)

(e) Using a learning rate of 0.1 and assuming initial values of 0, what updates does the TD-learning agent make after trials 1 and 2, above? First give the TD-learning update equation, and then provide the updates after the two trials. (5)