University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

SEMESTER II EXAMINATIONS

ACADEMIC YEAR 2017/2018

COMP 47590

Advanced Machine Learning

Dr. V. Dimitrova

Prof. P. Cunningham

Dr. B. Mac Namee *

Time Allowed: 2 Hours

Instructions for Candidates

Answer any **four** out of five questions. All questions carry equal marks. Total marks available **100**. The value of each part of each question is shown in brackets next to it.

Instructions for invigilators

This is a Closed Book/Notes exam.
Students are **not** permitted to bring materials to the Exam Hall.
Non-programmable calculators allowed.

1. (a) The **Bayes Optimal Classifier** is defined as follows:

$$t = \underset{l \in levels(t)}{\operatorname{argmax}} \sum_{\mathbb{M}_i \in \mathbb{M}} P(l|\mathbb{M}_i) P(\mathcal{D}|\mathbb{M}_i) P(\mathbb{M}_i)$$

(i) Describe what the Bayes Optimal Classifier computes. In your answer ensure to describe each term in the summation.

[4]

(ii) Explain why the **Bayes Optimal Classifier** is never actually implemented in practice.

[3]

(b) Thomas Deittrich describes three motivations for using ensemble models: **statistical**, **computational**, and **representational**. Describe each of these motivations and how they explain the performance of ensemble models.

[8]

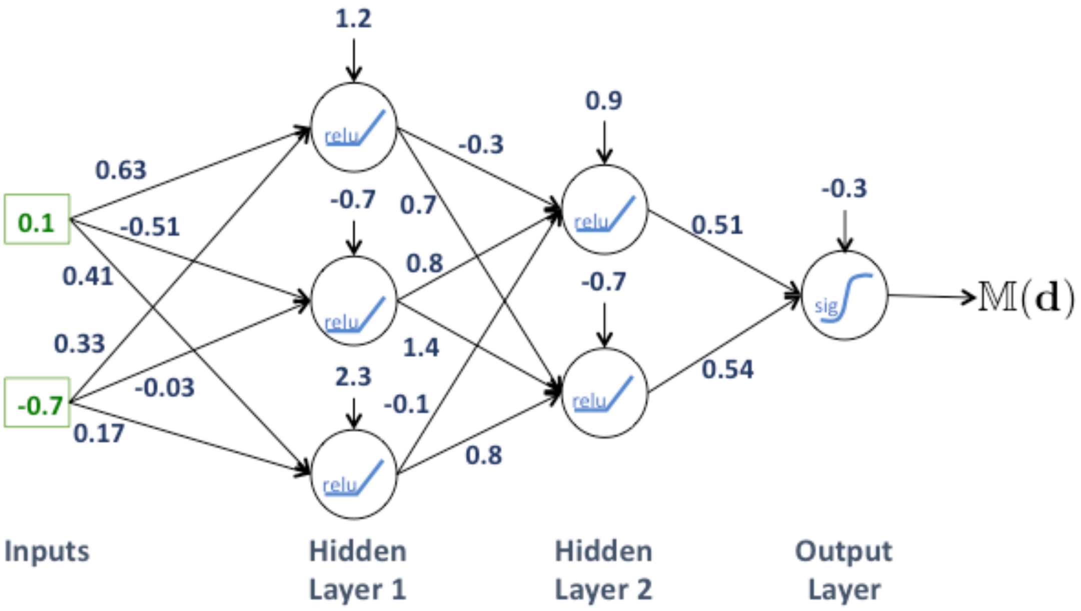
(c) Benchmark experiments have found repeatedly that ensemble models based on **bagging** are more robust to noise in the target features of a training dataset than ensemble models trained using **boosting**. Explain why this is the case. In your answer provide a short explanation of the bagging and boosting techniques.

[5]

(d) **Gradient boosting** has recently been shown to offer significant performance improvements over other boosted ensemble approaches. Explain what the gradient boosting algorithm trains its base models to predict.

[5]

2. (a) The image below shows a feed forward artificial network. The computational units in the two hidden layers use rectified linear (relu) activation functions and the output layer unit uses a sigmoid activation function. The weights and biases are shown along the links in the network.



(i) Perform a **forward propagation** through the network using an input feature vector of (0.1, -0.7).

[10]

(ii) If the target feature value for the current input vector is 1.0, calculate the **loss** associated with this training instance using **log loss**.

[3]

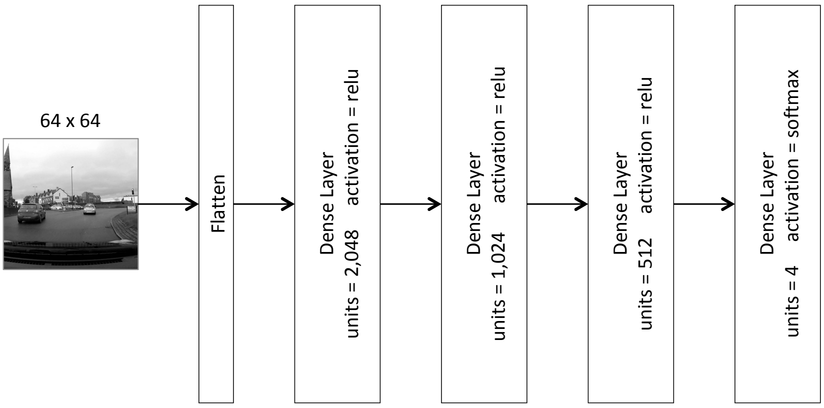
(b) In modern artificial neural networks the previously popular **sigmoid** and **tanh** activation functions have been largely replaced with the **rectified linear** (**relu**) activation function. Describe each of these activation functions (include appropriate diagrams) and explain the advantage of using relu.

[4]

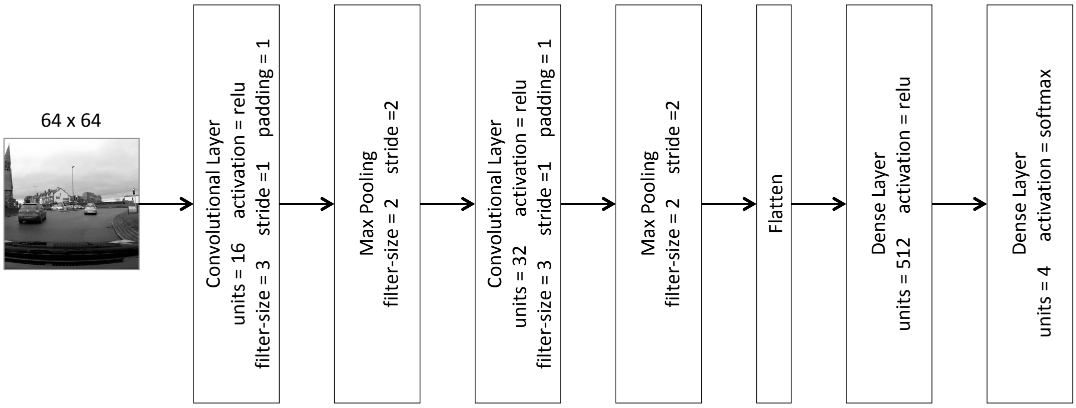
(c) There are three common variants of the gradient descent algorithm when training deep neural networks: **batch gradient descent**, **mini-batch gradient descent**, and **stochastic gradient descent**. Describe each of these approaches and discuss the advantages and disadvantages of each.

[8]

3. (a) You have been tasked with building a neural network system for controlling a self-driving car. The car has just four controls: accelerate, brake, turn left, and turn right. The only input is a 64 pixel by 64 pixel image from a dashboard camera within the car. Image (a) shows a multi-layer perceptron neural network architecture designed for this problem. Image (b) shows a convolutional neural network architecture designed for this problem. Both architectures are composed of four layers.


64 x 64
Flatten
Dense Layer
units = 2,048 activation = relu
Dense Layer
units = 1,024 activation = relu
Dense Layer
units = 512 activation = relu
Dense Layer
units = 4 activation = softmax

(a) A multi-layer perceptron network for self-driving car control


64 x 64
Convolutional Layer
units = 16 activation = relu
filter-size = 3 stride =1 padding = 1
Max Pooling
filter-size = 2 stride =2
Convolutional Layer
units = 32 activation = relu
filter-size = 3 stride =1 padding = 1
Max Pooling
filter-size = 2 stride =2
Flatten
Dense Layer
units = 512 activation = relu
Dense Layer
units = 4 activation = softmax

(b) A convolutional neural network for self-driving car control

Calculate the number of parameters (weights and biases) that need to be learned in each network architecture.

[8]

(b) The success of **convolutional neural networks** (**CNNs**) is often attributed to two characteristics: **sparse connections** and **shared weights**.

(i) Describe these two characteristics and the benefits they offer to CNNs.

[6]

(ii) The use of shared weights can make convolutional neural network models for image recognition **translation invariant**. Explain what this means.

[5]

(c) If we have a **recurrent neural network** (**RNN**), we can view it as a different type of network by "*unrolling it through time*". Explain what this means.

[6]

4. (a) Why is the **reward** only an indirect measure of the agent's performance in **reinforcement learning**? (Provide at least two reasons.)

[6]

(b) Explain the difference between the (**State-Action-Reward-State-Action**) **SARSA** and **Q-learning** algorithms for reinforcement learning.

[8]

(c) You are tasked with building an automated player of an endless runner video game (e.g. Temple Run, FlappyBird) with the following properties:

- no a priori knowledge of the world
- the character is constantly moving forward at a fixed speed
- the character is aware only of incoming elements of the game (enemies or obstacles) that are within a certain distance
- possible actions are jumping (to avoid low enemies or obstacles) or ducking (to avoid high enemies or obstacles)
- touching an enemy or obstacle leads to a restart of the game
- the goal is to maximise the distance covered by the character in a single run

(i) Which **reinforcement learning** method you would use to build this system? Give reasons for your answer.

[8]

(ii) Describe what constitutes a **state**, an **action** and whatever **other parameters** are relevant to the chosen method for this scenario (there is no need to describe the actual algorithm in detail).

[3]

5. (a) Inductive machine learning is often referred to as an **ill-posed problem**. Explain why this is the case and discuss the implications that follow from it. In your answer be sure to refer to examples of specific inductive machine learning algorithms.

[8]

(b) The table below shows the results of a benchmark experiment to compare the performance of a number of variants of a new learning algorithm, *YALA*, against each other and two baseline methods (random forests and multi-layer perceptrons). The performance of these algorithms has been measured across five different classification datasets using *10-fold cross validation*. Performance is measured in all cases using *macro-averaged f1-score*.

	YALA-1	YALA-2	YALA-3	Random Forest	MLP
abalone	0.462	0.437	0.436	0.451	0.448
arcene	0.804	0.799	0.749	0.742	0.802
dorothea	0.697	0.541	0.692	0.676	0.659
ecoli	0.449	0.437	0.436	0.447	0.451
iris	0.944	0.943	0.911	0.851	0.951

(i) Explain why using a *macro-averaged f1-score* is more appropriate than a *micro-average f1 score* for this experiment.

[4]

(ii) Convert the results table provided to a ranks table, including a row for average ranks.

[3]

(iii) To further investigate the differences between the different algorithms statistical significance testing based on the ranks table is recommended. Describe an appropriate set of tests to perform. (You do not actually need to perform any tests).

[4]

(c) The new EU General Data Protection Regulation (GDPR) comes into force this year on May 25th. Prof. Pedro Domingez, author of The Master Algorithm, recently stated that:

> "*Starting May 25, the European Union will require algorithms to explain their output,* ***making deep learning illegal****.*"

Discuss this claim.

[6]