SPCE0038: Machine Learning with Big-Data

Exam 2019

Guidelines:

- Answer **FOUR** questions of the FIVE questions provided.
- Exam duration is **TWO** hours.
- Each question has equal marks (20 marks per question).
- Markers place importance on clarity and a portion of the marks are awarded for clear descriptions, answers, drawings, and diagrams, and attention to precision in quantitative answers.

Question 1

(a) Describe three reasons why machine learning improved in effectiveness so dramatically over recent years.

[3 marks]

(b) Briefly describe batch gradient descent and stochastic gradient descent at a conceptual level.

[2 marks]

(c) Describe the properties of batch and stochastic gradient descent.

[3 marks]

(d) Write pseudo code (i.e. outline the algorithmic steps) defining a stochastic gradient descent algorithm to estimate parameters `theta`, given training feature matrix `X_train` and target vector `y_train`. Assume the functions `compute_random_index`, `compute_gradient`, and `compute_learning_rate` are available to you (i.e. you do not need to explicitly define them), although you should specify their inputs and outputs when you make use of them. Assume that the training set is already randomised.

[5 marks]

(e) Fill in the four missing entries of the confusion matrix below for a binary classifier to show which entries correspond to true-positives (TP), false-positives (FP), true-negatives (TN), and false-negatives (FN). Copy the confusion matrix and complete it in your answer book.

[2 marks]

		Predicted Negative	Positive
Actual	Negative	-	-
	Positive	-	-

(f) Define the true positive rate and the false positive rate in words and mathematically using the entries in your confusion matrix above.

[2 marks]

(g) Explain how a ROC curve is constructed and how the threshold defining the the ROC curve varies along the curve. Illustrate your explanation with a diagram.

[3 marks]

Question 2

(a) For the logistic unit shown below, specify the equations that define the output a of the logistic unit given the inputs x_j and parameters θ_j.

[2 marks]

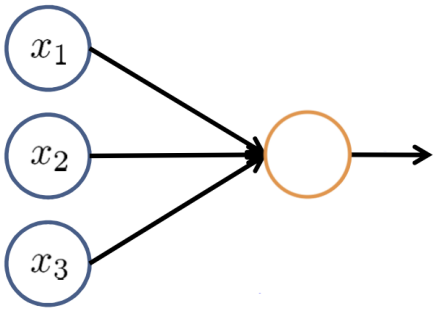

Fig. 1: Logistic unit.

(b) Specify the equations defining the logistic unit when a bias term is included.

[2 marks]

(c) Explain how the softmax function may be used to adapt a neural network to perform multi-class classification.

[2 marks]

(d) Define the softmax function mapping inputs a_j to outputs p_j, for $j = 1, \ldots, n$.

[1 mark]

(e) Show the outputs of the softmax function p_j satisfy the following properties:

(i) $\sum_{j=1}^{n} p_j = 1$;
(ii) $0 \leq p_j \leq 1$ for all j.

[3 marks]

(f) The gradient of the cost function is often used when training neural networks. Describe two ways gradients may be computed efficiently in practice for typical neural network cost functions.

[2 marks]

(g) Describe the vanishing gradient problem when training neural networks and its cause.

[2 marks]

(h) Give mathematical expressions for the sigmoid and ReLU activation functions and plot them. Which activation function is generally better and why?

[3 marks]

(i) Describe a shortcoming of the ReLU activation function.

[1 marks]

(j) Describe an improved version of the ReLU activation function to mitigate this shortcoming. Give the equation defining the improved activation function and illustrate your description with a diagram.

[2 marks]

Question 3

Consider a classification problem with features $x_j^{(i)}$ and targets $y^{(i)}$, where there are n_{features} features and n_{objects} objects (assume i and j are indexed from 1).

(a) Construct the feature matrix $\mathbf{X}$ and target vector $\boldsymbol{y}$ from features $x_j^{(i)}$ and targets $y^{(i)}$.

[1 mark]

(b) Give the size of the feature matrix $\mathbf{X}$ and target vector $\boldsymbol{y}$.

[1 mark]

(c) Explain the process of n-fold cross-validation and why it is useful. Use a diagram to illustrate your answer.

[4 marks]

(d) What are the four key steps in training and applying a classification model in Scikit-Learn (assuming data are already set up appropriately).

[4 marks]

Consider the underlying (true) model

$$y = f(\boldsymbol{x}) + \epsilon,$$

where f is the true model, to be approximated by h, $\boldsymbol{x}$ is an object feature vector, and ϵ is noise, with zero mean and variance σ^2.

(e) Explain the three contributions to the mean square error.

[3 marks]

(f) Show that

$$\mathrm{E}\left[(y - h(\boldsymbol{x}))^2\right] = \mathrm{Bias}^2\left[h(\boldsymbol{x})\right] + \mathrm{Var}\left[h(\boldsymbol{x})\right] + \sigma^2$$

[3 marks]

(g) Explain the bias-variance trade-off and how it relates to model complexity. Illustrate your explanation with a diagram.

[4 marks]

Question 4

(a) State which type of neural network would be appropriate for the following problems, stating your reasons.

(i) An image classification problem.

(ii) A language translation problem.

(iii) Sentiment score analysis.

[6 marks]

(b) Describe what a pooling layer in a neural network is, and state some reasons why such a layer may be included.

[4 marks]

(c) For a convolutional neural network describe what stride parameter defines.

[2 marks]

(d) Consider the following piece of Python code using Tensor Flow:

```
1 import tensorflow as tf
2 reset_graph()
3 n_inputs = 3
4 missing_line
5 n_outputs = n_inputs
6 learning_rate = 0.01
7 X = tf.placeholder(tf.float32, shape=[None, n_inputs])
8 hidden = tf.layers.dense(X, missing_variable)
9 outputs = tf.layers.dense(hidden, n_outputs)
10 reconstruction_loss = tf.reduce_mean(tf.square(outputs - X))
11 optimizer = tf.train.AdamOptimizer(learning_rate)
12 training_op = optimizer.minimize(reconstruction_loss)
```

(i) State what this code is doing in general terms.

(ii) Identify the missing variable on line 8.

(iii) Provide a value for the missing variable on line 4, justifying its value. If `n_inputs = 4` how would this change your answer?

(iv) Describe in words what the learning rate on line 6 is. If this was set too high how might the results change.

[8 marks]

Question 5

This question focuses on data formats, normal forms, SQL, markup languages and semantic models.

(a) (i) What do we mean by the **serialization** of an object or data structure?

[1 mark]

(ii) Provide at least two reasons why it might be beneficial to serialize.

[1 mark]

(b) The Anonymous gallery possesses a physical archive where it stores information on all the purchases made by its customers over time. These purchases are recorded in files, an example of which is given in Fig. 2. The gallery also wishes to maintain records of data on customers, artists and objects. There might be several objects authored by a single artist and objects may be bought and sold several times over. In other words, the gallery may sell something, buy it back at a later date and sell it to another customer.

Customer purchase history file — Customer ID: 387

Customer

John Brown
8 King Street
SW1Y 6QT London
Phone: 000111222333
Email: john.brown999@gmail.com

Purchases

Artist	Title	Sales date	Hammer price
03 – Andy Warhol	*Campbell's Soup I*	01/01/2000	1.999.999$
12 – Umberto Boccioni	*Dynamism of a Cyclist*	13/04/2007	12.345.777$
99 – Jan van Goyen	*River Scene*	12/12/2010	64.956$

Fig. 2: Example of a customer's file in the Anonymous gallery archives.

(i) Provide a database schema **not in normal form (UNF)**. Use the following syntax: table_name [column_1, (sub_column_1, sub_column_2), ...] grouping attributes which end-up in the same column using parentheses.

[2 marks]

(ii) Provide a database schema in **first normal form (1NF)**.

[2 marks]

(iii) Provide a database schema in **second normal form (2NF)**.

[2 marks]

(iv) Provide a database schema in **third normal form (3NF)**.

[4 marks]

(c) Write a query in **SQL** that lists all the works of art bought by John Brown (roughly as Fig. 2). You might need *some* of the following SQL syntax: SELECT (AS), FROM, WHERE, JOIN (ON), GROUP BY, ORDER BY.

[3 marks]

(d) Clarify the differences between URI (Uniform Resource Indicator), URL (Locator), URN (Name) and namespaces.

[2 marks]

(e) Use **RDFS** (Resource Description Framework Schema) to specify the schema of artists, their works and relations among them, as in Fig. 2. You might need to use some of the following components: RDFS:Class, RDF:Property, XSD:string, XSD:integer, RDFS:domain, RDFS:range. XSD stands for XMLSchema. When defining your namespaces, feel free to ignore providing the correct URL. A partial solution is fine, provided you show an understanding of: namespaces, classes and properties, domains and ranges, data types, RDFS syntax.

[3 marks]