UNIVERSITY COLLEGE LONDON

END ASSESSMENT

FOR INTERNAL STUDENTS

MODULE CODE :	**SPCE0038**
MODULE NAME :	**Machine Learning with Big Data**
LEVEL: :	**Postgraduate**
SUBMISSION DATE:	**12 May 2020**
TIME :	**12:00 BST**
TIME ALLOWED :	**1 week**

SPCE0038: Machine Learning with Big-Data

Alternative Assessment 2020

Guidelines:

- Answer all of the FIVE questions provided.
- Each question has equal marks (30 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) Draw a diagram of the basic logistic unit that is used as the core building block of artificial neural networks. For simplicity you can ignore the inclusion of a bias term. Describe the components of your diagram in words.

[3 marks]

(b) Specify the equations that define the output a of the logistic unit given the inputs x_j and weights θ_j. Again, you may ignore a bias term for simplicity.

[4 marks]

(c) Using your logistic unit as a base building block, draw a diagram of a fully connected, feed-forward artificial neural network with three layers (one input, one hidden and one output layer), three input units, three hidden units, and one output node. Again, you may ignore a bias term for simplicity.

[4 marks]

(d) Specify the equations defining the full artificial neural network of part (c), extending your equations given for a single logistic unit that you specified above in part (b). Again, you may ignore a bias term for simplicity.

[6 marks]

(e) What typical cost functions are used to train neural networks for regression and classification problems? Specify the corresponding cost function equations for targets $y_j^{(i)}$ and predictions $p_j^{(i)}$, where i denotes training instance and j the output node.

[6 marks]

(f) Explain what is meant for a network to be deep?

[1 marks]

(g) Why do deep networks provide a powerful representation framework? Include a discussion of the universal approximation theorem.

[6 marks]

Question 2

Gradient descent algorithms take a step η in the direction of decreasing gradient, where the update of parameter θ is given by a form similar to

$$\theta \leftarrow \theta - \eta \nabla_\theta C(\theta),$$

where C denotes the cost function and $\nabla_\theta C$ the gradient of the cost function with respect to θ. The variable η is often called the learning rate. Gradient descent based algorithms are often used to train deep learning models.

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

[4 marks]

(b) Although stochastic gradient descent is often very effective, why are alternative optimisation algorithms typically considered for training?

[2 mark]

(c) Describe the momentum optimisation algorithm, including the update equations.

[3 marks]

(d) Describe the Nesterov variant of the momentum algorithm, including the update equations.

[3 marks]

(e) Explain the concept behind the AdaGrad algorithm and how this can help with training (no need to include update equations).

[4 marks]

(f) Explain the concept behind the RMSProp algorithm and how this can help with training (no need to include update equations).

[4 marks]

(g) Adam is the standard go-to algorithm for training deep networks. Explain the components of the algorithms considered so far that are included in the Adam algorithm.

[3 marks]

(h) Deep networks have very large numbers of parameters and so can be prone to overfitting. Explain the dropout regularisation technique to avoid overfitting. Support your explanation with a diagram.

[7 marks]

Question 3

(a) Describe the knowledge based approach to artificial intelligence.

[4 marks]

(b) Describe the machine learning approach to artificial intelligence.

[2 marks]

(c) Describe the traditional machine learning approach of feature engineering?

[4 marks]

(d) Briefly describe supervised, unsupervised and reinforcement learning.

[3 marks]

(e) For supervised learning, briefly describe the difference between regression and classification problems.

[2 marks]

(f) Consider logistic regression for K classes, where the predicted probabilities for each class k are given by

$$\hat{p}_k = \frac{\exp\left(s_k(x)\right)}{\sum_{k'=1}^{K} \exp\left(s_{k'}(x)\right)}, \quad \text{with} \quad s_k(x) = \left(\theta^{(k)}\right)^{\mathrm{T}} x,$$

for input x and parameters $\theta^{(k)}$ (recall each $\theta^{(k)}$ includes n features).

Consider the generalised cost function for logistic regression given by the cross entropy

$$C(\Theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} y_k^{(i)} \log\left(\hat{p}_k^{(i)}\right),$$

where i denotes training instance and m the total number of training instances. The target value of instance i for class k is denoted $y_k^{(i)}$.

Show that the derivative of the cost function is given by

$$\frac{\partial C}{\partial \theta^{(k)}} = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{p}_k^{(i)} - y_k^{(i)}\right) x^{(i)}.$$

Hint: For the term $\frac{\partial \hat{p}_k}{\partial s_{k'}}$ it may be convenient to consider the cases $k = k'$ and $k \neq k'$ separately and then combine. Note also that $\sum_{k=1}^{K} y_k = 1$.

[15 marks]

Question 4

(a) Explain the computational model of TensorFlow in terms of computational graph construction and execution.

[3 marks]

(b) Explain the difference between TensorFlow `Variable` and `Constant` types.

[3 marks]

(c) Explain what a TensorFlow `Placeholder` variable is and why it may be useful.

[4 marks]

(d) Explain autodiff and its advantages.

[4 marks]

(e) Consider the following TensorFlow code to set up a computational graph and execute it. Assume `scaled_housing_data_plus_bias` is an $m \times (n+1)$ feature matrix and `housing_data_target` is an $m \times 1$ target vector, where m denotes the number of training instances and n the number of features ($n+1$ is the number of features when including a bias).

(i) Set up computational graph:

```python
import tensorflow as tf
reset_graph()

n_epochs = 1000
learning_rate = 0.01

X = tf.constant(scaled_housing_data_plus_bias, dtype=tf.float32,
                name="X")
y = tf.constant(housing_data_target, dtype=tf.float32, name="y")

theta = tf.Variable(tf.random_uniform([n + 1, 1], -1.0, 1.0),
                    name="theta")
y_pred = tf.matmul(X, theta, name="predictions")
error = y_pred - y
mse = tf.reduce_mean(tf.square(error), name="mse")

optimizer = tf.train.GradientDescentOptimizer(learning_rate)
training_op = optimizer.minimize(mse)
```

(ii) Execute:

```python
init = tf.global_variables_initializer()

with tf.Session() as sess:
    sess.run(init)

    for epoch in range(n_epochs):
        if epoch % 100 == 0:
            print("Epoch", epoch, "MSE=", mse.eval())
        sess.run(training_op)

    best_theta = theta.eval()
```

What machine learning problem does this TensorFlow code solve? What optimisation algorithm is used?

[4 marks]

(f) Write code to solve the problem given in part (e) using mini-batch gradient descent. You may find it helpful to base your answer on the code given in part (e) and then revise it where necessary. Assume you have available a function `fetch_batch` to fetch each mini-batch, with signature specified below:

```
def fetch_batch(epoch, batch_index, batch_size):
    ...
    return X_batch, y_batch
```

[12 marks]

Question 5

(a) Describe what Principal Component Analysis (PCA) is.

[3 marks]

(b) Define the explained variance ratio.

[2 marks]

(c) Explain what Kernel PCA is.

[5 marks]

(d) Define the process of Local Linear Embedding (LLE).

[4 marks]

(e) In the first step of LLE, for a set of training instances x_i, with k nearest neighbours LLE will first reconstruct the x_i as a linear function of these neighbours. Write down an equation that would describe this process, and any normalisation that is applied.

[8 marks]

(f) The second step of LLE is to map the training instances into a d-dimensional space while preserving local relationships as much as possible. If z_i is the d-space equivalent of x_i then describe the condition that must be met.

[8 marks]