IMPERIAL COLLEGE LONDON

MSc EXAMINATION 2020

For internal students of Imperial College London

Taken by students of the Masters of Applied Computational Science and Engineering

ACSE-8 Machine Learning

Friday 29 May 2020, 13:00-15:00

The total number of marks is 100.

This exam comprises six questions. Please answer ALL questions.

Contact details in case of technical difficulties:

Ying Ashton
Tel: +44 (0)20 759 43067
Email: y.ashton@imperial.ac.uk

Gareth Collins
Tel: +44 (0)20 759 41518

James Percival
Tel: +44 (0)20 724 90698

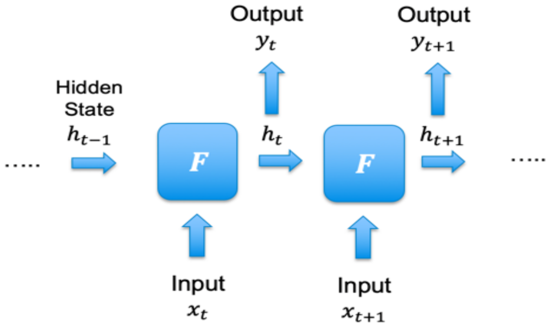
Olivier Dubrule
Tel: +33 (0)627666340

While this time-limited remote assessment was not originally designed to be open book, in the present circumstances it is being run as an open-book examination. We have worked hard to create an exam that assesses synthesis of knowledge rather than factual recall. Thus, access to the internet, notes or other sources of factual information in the time provided will not be helpful and may well limit your time to successfully synthesise the answers required. Where individual questions rely more on factual recall and may, therefore, be less discriminatory in an open book context, we may compare the performance on these questions to similar style questions in previous years and we may scale or ignore the marks associated with such questions or parts of the questions. The use of the work of another student, past or present, constitutes plagiarism. Giving your work to another student to use may also constitute an offence. Collusion is a form of plagiarism and will be treated in a similar manner. This is an individual assessment and thus should be completed solely by you. The College will investigate all instances where an examination or assessment offence is reported or suspected, using plagiarism software, vivas and other tools, and apply appropriate penalties to students. In all examinations, we will analyse exam performance against previous performance and against data from previous years and use an evidence-based approach to maintain a fair and robust examination. As with all exams, the best strategy is to read the question carefully and answer as fully as possible, taking account of the time and number of marks available.

ACSE-8 Exam

Question 1 (20 marks): Recurrent Neural Network

Suppose we want to train a recurrent neural network to predict, given the average atmospheric pressure and temperature of today (and of each of the previous days), what will be the average atmospheric pressure and the temperature of the next day. We build the following classical set-up for the network, with a dimension of the hidden state vector h_t taken as a hyperparameter d :



We have the standard recurrent network formulas, where y_t is the forecast for the next day $(t + 1)$.

$$h_t = F(h_{t-1}, x_t) = tanh(W_{hh}h_{t-1} + W_{xh}x_t) \text{ and } y_t = W_{hy}h_t$$

a. Explain what are, for our specific example, the vectors x_t and y_t, and what is the total number of training parameters. Calculate this number first in the case without bias terms, then in the case with bias terms. (10 marks)

b. If we were dealing with a supervised neural network, the training set data pairs would be $(x_i, y_i)_{i=1,\dots,m}$ with m equal to the number of training data and y_i the label associated with each vector x_i. Here we assume that we have a training set of 1000 values of x_t . What are the training set data pairs in this recurrent neural network context? (5 marks)

c. Assuming that we train the network by chunks - or batches - of 20 training examples, explain in less than 8 lines how the training process works and suggest which loss function could be used. (5 marks)

Question 2 (20 marks): Softmax 2D combined with Feed-Forward Neural Network

Suppose a feed-forward neural network has been trained to classify each pixel of an image into three classes of colour, Red, Green or Blue. Classes are one-hot encoded in a vector of dimension 3 where Red, Green, Blue respectively correspond to the first, second and third coordinate.

The network is very simple: the input layer contains the two coordinates (x_1, x_2) of the pixel in the image, the next layer is linear with three neurons , and the last layer is a Softmax layer. The 6 weights and the three biases of the linear layer have been trained already and the network is as follows:

$$\begin{pmatrix}x_1\\x_2\end{pmatrix} \longrightarrow \begin{pmatrix}1+x_1+x_2\\3x_1-x_2\\-1+x_1-x_2\end{pmatrix} \longrightarrow Softmax\begin{pmatrix}1+x_1+x_2\\3x_1-x_2\\-1+x_1-x_2\end{pmatrix}$$

Calculate the output of the network for the point $\begin{pmatrix}1\\1\end{pmatrix}$. Which colour is this point going to be classified into? (5 marks)

Suppose that we know that the three pixels of respective coordinates $\begin{pmatrix}1\\0\end{pmatrix}$, $\begin{pmatrix}0\\1\end{pmatrix}$ and $\begin{pmatrix}1\\1\end{pmatrix}$ are respectively Red, Green and Blue. What is the cross-entropy between the three values predicted by the network and the three actual pixel values? (15 marks)

Question 3 (10 marks): Variance of a Uniform Distribution

In their classic paper, X. Glorot and Y. Bengio recommend to initialize the neural network's training weights from a distribution of mean 0 and variance $\frac{1}{n}$, where n is the number of neurons in the considered layer. Obviously, this can be done by sampling a normal distribution $N\left(x; 0, \frac{1}{n}\right)$. However, if instead of a normal distribution we use a uniform distribution between $-a$ and a, which value of a should we use to obtain a mean of zero and a variance of $\frac{1}{n}$?

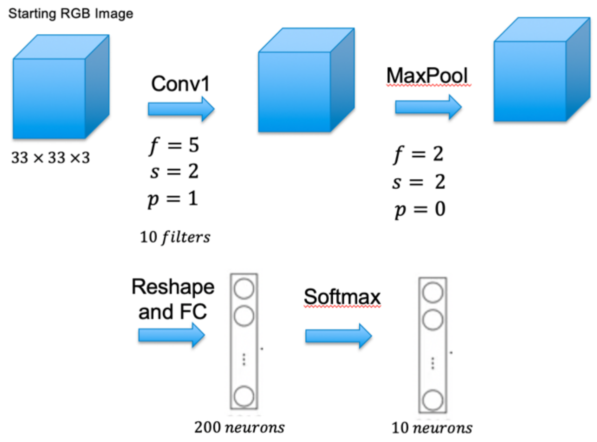
Question 4 (20 marks): Convolutional Neural Network

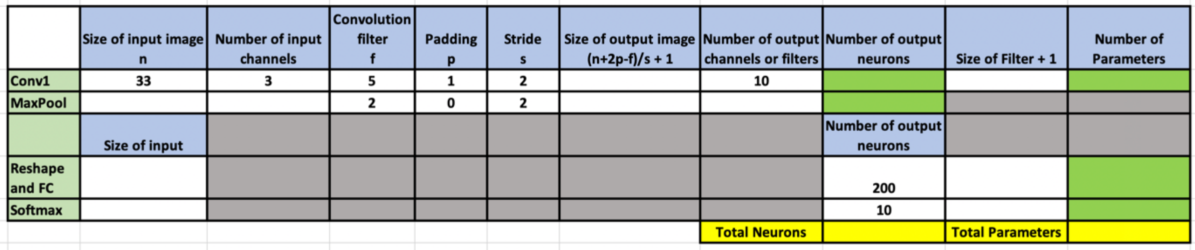
The figure below describes a simple convolutional network. We assume that each convolution operation is performed with a bias term.

a. Calculate the number of parameters associated with each of the three layers (Conv1, Reshape and FC, and Softmax). (10 marks)

b. Calculate the number of output neurons in the Conv1 and MaxPool layers. (10 marks)

The table below is given as a help for the calculation. The answers to the two above questions are associated with the green cells. There is no need to reproduce the whole table in your answer.





	Size of input image n	Number of input channels	Convolution filter f	Padding p	Stride s	Size of output image (n+2p-f)/s + 1	Number of output channels or filters	Number of output neurons	Size of Filter + 1	Number of Parameters
Conv1	33	3	5	1	2		10			
MaxPool			2	0	2					
	Size of input							Number of output neurons		
Reshape and FC								200		
Softmax								10		
							Total Neurons		Total Parameters	

Question 5 (15 marks): Maximum Likelihood and KL Divergence

We are dealing with a simplistic unsupervised learning problem. Suppose we have a Training Set of $m = 3$ IID training examples $x^{(i)}$ each of dimension 1:

$$x^{(1)} = 2 \quad x^{(2)} = 5 \quad x^{(3)} = 6$$

We wish to fit to this Set a Normal distribution $N(x;\ \mu, \sigma^2)$ of mean μ and variance σ^2.

a. What is the log-likelihood of the Training Set as a function of μ and σ^2? (10 marks)

b. Using the Training Set, show why calculating the values μ and σ^2 maximizing the log-likelihood of the Training Set (as calculated in part a.) is equivalent to calculating the values μ and σ^2 minimizing the Kullback-Leibler Divergence $KL(p_d \| N)$ between the experimental distribution p_d of the Training Set and the Normal distribution $N(x;\ \mu, \sigma^2)$. (5 marks)

Question 6 (15 marks): Unsupervised Machine Learning

We are working on a dataset containing 100,000 images of people's faces, with each image composed of 100x100 grey-scale pixels. The answer to each of the three questions below should be about 5 lines.

a. Suppose we want to reduce the dimensionality of the data, and represent each image by a latent vector of dimension 10. What is a good approach to achieve this? (5 marks)

b. Now suppose we want to make the previous approach more general, and able not only to do dimensionality reduction of the data themselves, but also generate some latent vectors and images that statistically look like the 100,000 images of the training dataset, but are not identical to them. Which approach should we use? (5 marks)

c. Which approach could we also use if we were not interested in the latent vector associated with each person's image, but just in the generation of images that statistically look like the 100,000 images? (5 marks)

END OF EXAMINATION