DD2437 Artificial Neural Networks and Deep Architectures (annda18)

Exam 2018-03-15 at 14:00-19:00

Use a separate sheet for each question. Brief and concrete answers are preferred by all means. Do NOT give several mutually conflicting answers to a question (ingen helgardering). If you do, the alternative with lowest score will be chosen. You are allowed to use a standard calculator (no mobile phones allowed) and a standard English-other language dictionary may be used. Please keep in mind that you are obliged to abide by *the code of honor* at EECS/KTH (https://www.kth.se/en/eecs/utbildning/hederskodex/inledning-1.17237).

The exam is divided into two parts. Successful completion of Part I is a necessary but not a sufficient condition to pass the exam with E. In other words, E can be secured by collecting 60% of all the points available in the exam (30 out of 50p) provided that most answers in Part I (80%, i.e. 16 out of 20p) are correct. Part II offers the scope for higher grades.

Good luck!

Pawel and Erik

Part I (max 20p)

Question 1 (4p)

Briefly (in maximum one or two sentences) explain the following terms (*a*-*h*) or the differences between the following concepts (*i*-*p*) focusing on what you find the key characteristics to be:

a) generalization

b) artificial neural network architecture

c) learning algorithm

d) bias-variance dilemma

e) backprop

f) dead units in competitive learning

g) autoassociative memory

h) greedy layer-wise pretraining

i) supervised vs unsupervised learning

j) recurrent vs feed-forward networks

k) Hopfield net vs Boltzmann machine

l) RBF network vs multi-layer perceptron (MLP)

m) vanishing vs exploding gradients

n) shallow vs deep learning

o) sequential (on-line) vs batch learning

p) discriminative vs generative classifier

Question 2 (4p)

For a single-layer perceptron, what are the convergence criteria of the perceptron learning rule and delta rule? List key differences between the two learning algorithms and discuss briefly the quality of the output after the learning has terminated.

Then, perform computations on the following initial weight matrix **W** (the last column contains the biases; the threshold is 0):

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 2 & -1 \\ -2 & 3 & 0 & 1 \end{bmatrix}$$

subject to training with the sequence of three patterns, p1, p2 and p3, having their desirable target outputs, t1 t2 and t3, respectively:

$p1 = [\ 1, 0, -1]^T$, $t1 = [1\ \ 1]^T$,
$p2 = [-1, 2, 0]^T$, $t2 = [0, 1]^T$,
$p3 = [\ 1, 1, 1]^T$, $t3 = [0, 1]^T$.

Show your calculations for one epoch of pattern-by-pattern training using a traditional perceptron learning rule with a step length of 0.1, explain the symbols used and sketch a network diagram including numerical values in the weight matrix obtained as a result of training (just place all the values from **W** next to the network connections in your diagram).

Finally, define the classification accuracy measure and calculate it at the end of your simplistic one-epoch training with the three patterns. Are they all classified correctly?

Question 3 (2p)

Explain briefly why overfitting is an undesirable effect and how it usually manifests itself. What techniques (name at least four) can you recommend that should be employed to minimise the risk and potential implications of overfitting in multi-layer perceptrons? What are key advantages of using Bayesian regularisation?

Question 4 (3p)

Very briefly outline the principles (no need for any math formulae) of training self-organising maps with focus on the following concepts: input vs output (grid) space, topographic mapping, neighbourhood shrinking. What are Kohonen maps useful for?

Question 5 (4p)

Assume a Hopfield network with bipolar {1, -1} nodes and the following weight matrix, **W**:

$$\mathbf{W} = \begin{bmatrix} 0 & 2 & -1 & -1 & -1 \\ 2 & 0 & -1 & -1 & -1 \\ -1 & -1 & 0 & 2 & 2 \\ -1 & -1 & 2 & 0 & 2 \\ -1 & -1 & 2 & 2 & 0 \end{bmatrix}$$

Is the following pattern, x = [-1, 1, -1, 1, -1] a stored pattern (fixed-point attractor)? Show your calculations assuming that the node output is 1 if the net input is greater or equal to 0. What is an alternative candidate for a stored pattern in this network?

What are assumptions for the weight matrix in Hopfield networks?

What learning rule would you use to arrive at the weight matrix of the Hopfield network storing a set of desirable patterns? What are characteristics of input patterns that promote large storage capacity in Hopfield networks?

Question 6 (3p)

What are key computational factors that motivate the use of deep rather than shallow neural network architectures?

At the same time, why is the use/development of deep neural networks more challenging when compared to the shallow networks? How can we overcome these difficulties/challenges in deep learning?

Name at least two popular deep neural network types (with respect to learning algorithms, architectures) and list their basic characteristics.

Part II (max 30p)

Question 7 (4p)

For each statement listed below, please declare whether it is always or never correct. If it is wrong, provide alternative wording to make it correct (more nuanced statement that is correct). Please keep your comments and statements as short as possible.

~~a) increase the number of nodes~~
~~b) decrease the number of weights~~
~~c) increase the number of learning epochs~~
~~d) decrease the number of training examples~~

Question 8 (5p)

You have been charged with a task of recognising current frame of mind/mood of 15 patients suffering from mild dementia based on a set of six measurements such as body temperature, blood pressure and a pulse rate among others (assuming that they are somewhat independent). The measurements are collected three times per day (in the morning, afternoon and evening) from each patient to produce a daily average, which is then matched against a real-valued score characterising each patient's frame of mind/mood (patients rate only once their own overall daily condition on the continuous scale from 1 to 5, where 1 corresponds to a bad and 5 to excellent frame of mind). Please bear in mind that the data collected from patients suffer from nonstationarity effects, e.g. different measurement devices can be employed on different occasions for individual patients etc.

Using data collected across multiple days from a population of patients, you are supposed to examine the effectiveness of an artificial neural network based approach to the problem of patients' frame of mind detection/recognition. In this context there are number of decisions to take and issues to consider. Please comment briefly on the following aspects and address the emerging questions (if necessary introduce extra assumptions relevant to your answer):

a) Given a dataset of three months of average daily recordings for a single patient with associated scores describing her/his daily frame of mind, propose how you would split and normalise the data to train a neural network for the given patient and test the quality of predictions (their generalizability). Suggest also a neural network model (architecture and a learning algorithm) that you would apply. (1p)

b) What are key hyperparameters that would determine your network and how would you approach the problem of model selection (choosing optimal hyperparameters)? (1p)

c) How would you design a study to examine whether one universal neural network model could be developed to make reliable predictions for many patients as opposed to neural networks individually adjusted to each patient? If it turned out that we should have separate networks tuned to individual characteristics of each patient (the universal network did not work too well), could you still suggest how one could try to exploit the data collected from other subjects? Analogously, would you be able to exploit in the training process a subset of recordings even though they were not annotated with scores describing the corresponding patient's frame of mind or would you just discard such incomplete (missing annotations/labels) data? (1p)

d) If you were offered an extra set of hundreds of measurable characteristics from each patient (on top of the original six measurement types), what would be your immediate concern related to your task or what challenge would you have to face before training a suitable neural network for frame of mind prediction? (1p)

e) In what circumstances would you consider unreasonable to work with daily measurement averages rather than all three sets of recordings collected each day (in the morning, afternoon and evening)? What type of neural network would you recommend for development (with what input) if it turned out that the aforementioned daily averaging was a bad idea and the measurements were to be collected at least twice more often (i.e. six times per day)? (1p)

Question 9 (5p)

Why does averaging the outputs from multiple models typically gives better results than using just one model. Let's assume that the outputs from 10 neural network models is to be averaged. Apart from producing decent performance by individual network models, what additional property should their collection have to be considered as a good candidate for output averaging?

Generally, what ensemble learning approaches (meta-algorithms) could you propose?

Finally, briefly discuss whether a committee of networks besides improving the regression or classification performance also produces a more robust model with greater generalisation capabilities and interpretability.

Question 10 (6p)

In each case below, what type of problem/processing is this? Propose a neural network approach to address it. For your solution, give algorithm name, network topology, describe input and output (what kind of data, how many nodes etc), how is training done, etc.

a) As machines get more and more complex, mechanisms for self-diagnostics get more important. In an automated factory, a number of parameters (sensor values, data from micro-controller memories, etc) are constantly logged. Over time, the company has saved this data (now about 122000 entries) as well as a label describing the data. This label is either a collective label ”normal operation” or, in case it was found out that the machine was not working properly, the type of error was saved. Your task is to use this data to construct an automated self-diagnostics system.

b) You are the organizer of a large trade show (exhibition) where manufacturers of technologies with positive impact on the environment will be presented. It is a large and quite diverse set of technologies and application areas obviously. You think it is a good idea if technologies that are somehow similar or related are located close in the hall so that visitors do not have to walk around in random (the hall is just too large to do a complete round-tour). You ask each exhibitor to tag their product according to a list of properties you think cover aspects of interest of the visitors. This tagging means giving a yes/no to whether each aspect/application/property applies to the product or not. You collect this information from all exhibitors. How do you decide the location of the exhibitors?

Question 11 (4p)

Reservoir computing (echo state network, liquid state machine) offers the functionality of processing temporal/sequential data. One of the central components of such networks is reservoir. Explain briefly what it is and describe its key properties that promote good performance in sequence classification tasks. What is meant behind the statement that training echo state networks is fast? What kind of training is it?

What are the fundamental differences in solving the problem of sequence mapping or sequence discrimination between echo state networks and typical (vanilla) recurrent neural networks (RNN) or long short-term memory (LSTM) networks? Also, what functional deficit of standard RNN units do LSTMs address? What features of LSTM units are of special importance in this regard?

Question 12 (6p)

Please explain in one or two sentences what you understand by a generic term of data representation and specifically by the concept of *learning representations*. In what way do deep neural networks facilitate learning of data representations in comparison with shallow feed-forward networks? Why should we care about this aspect of the network's functionality?

Apart from convolutional neural networks (CNN), there are other powerful deep learning architectures for extracting suitable data representations such as deep belief nets (DBN) and stacked autoencoders. Please describe briefly key algorithms involved in building these architectures (excluding CNNs) and thereby creating new data representation (provide names and conceptual explanation of how they work). In the RBM context, what is the meaning or probabilistic interpretation of the hidden layer? For autoencoders, how can we learn high-dimensional representations (in the higher-dimensional space than that of the input) without compromising autoencoder's capability to find a non-trivial data mapping and why isn't it really a problem in *undercomplete* autoencoders?

What do you understand by the concept of *distributed representations*? What would be an alternative coding convention(s) usually contrasted with distributed representations? Why would the distributed coding scheme be considered particularly useful for pattern recognition and prominent in the world of deep learning? Finally, explain please in one sentence what sparse distributed representations are.