DD2437 Artificial Neural Networks and Deep Architectures (annda18)

Re-exam 2018-06-05 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 necessary but not sufficient to pass the exam with E. In other words, E can be secured by collecting more than 50% of all the points available in the exam (above 25 out of 50p) provided that at least 14p (i.e., 70%: 14 out of 20p) are obtained in Part I. 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-d*) or the differences between the following concepts (*e-h*) focusing on what you find the key characteristics to be:

a) activation function

b) learning algorithm

c) bias-variance dilemma

d) overfitting

e) supervised vs unsupervised learning

f) recurrent vs feed-forward networks

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

h) sequential (on-line) vs batch learning

Question 2 (3p)

Explain briefly the concept of generalisation. What are the major factors that determine the neural network's generalisation capability? What is the founding principle of Bayesian framework for regularisation in classical multi-layer perceptrons? What are key practical advantages of using Bayesian regularisation?

Question 3 (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 for the step activation function is 0),

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & -3 & -1 & 0.5 \end{bmatrix},$$

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

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

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 this one-epoch 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 4 (2p)

Very briefly explain the concept of topology-preserving mapping in Kohonen networks. How is this property promoted in the learning algorithm?

Question 5 (4p)

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

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

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 26p)

Question 7 (6p)

You have been requested to help psychologists to understand their empirical data with the use of artificial neural networks. In particular, the researchers collected information about a large set of odours. Namely, for each odour there was a single linguistic descriptor linked to how it is perceived by humans, e.g. citrus, coffee, woody, fishy etc., and 17 characteristics describing chemical/molecular properties. The researchers' aim was to study the relationship between chemical and perceptual properties of odours. To start with, they have approached you to help them investigate whether these 17 chemical features explain some perceptual similarities between odours. For example, they have asked you if it is possible to predict how new odours are perceived, or what other known odours they are similar to in terms of perceived smell based only on these known 17 chemical characteristics.

a) What kind of neural network would you propose to address these challenging questions? Please describe the network architecture, type of computations and learning process. Would it be a supervised or unsupervised learning approach? How would you visualise data and your findings for the researchers to facilitate their intuitive understanding?

b) How would you extend your neural network approach if instead of 17 there were additional 100 molecular features added to the dataset for each odour? For your network to function properly would you critically need also the higher number of samples (odours) in the dataset?

c) Now imagine that the goal of the researcher's project has been changed to just predicting whether an odour is pleasant, unpleasant or neutral based on the chemical characteristics. Your dataset is then enhanced by these extra labels about the odour valence obtained from a panel of experts. Would you change your neural network approach to test how reliable are such predictions? Would your approach be sensitive to the number of chemical descriptors (e.g., if you had extra 100 molecular features, as before)? If yes, in what way?

Question 8 (4p)

Please characterise briefly main advantages and disadvantages of ensemble learning. Provide also a simple example of how you would build and train an ensemble of neural networks of your own choice (consider how you would use available data). What are desirable properties of a collection of weak learners in the context of ensemble learning? What effect does the averaging of outputs produced by multiple models have on the bias-variance properties of the resulting estimate (based on the mean of outputs) of the ensemble performance when compared to that of individual learners?

Question 9 (4p)

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

a) In the attempt to facilitate some form of "auto-complete" function for digitised hand-written text, a lot of hand-written texts have been collected. All the words in the texts have been segmented into strings of images of individual letters and these images were annotated as letter symbols. The intended neural network based system is supposed to be developed and trained to offer online suggestions for the multiple subsequent letters (or maybe even words) as the user's is entering hand-written text letter by letter. It is sufficient if these multi-step-ahead predictions are made on the symbolic level, i.e. letter symbols are predicted, not their handwritten versions (though the input is still handwritten text). What would it take to generate handwritten text (not a string of letter symbols, as originally asked) as a prediction?

b) In the care of elderly, there is a huge need for autonomous systems that can monitor the senior citizen and send an alarm in the case the person has fallen, appears confused or otherwise behaves different from normal. In a study, a set of sensors (ultrasound movement detectors, microphones, etc) were installed in the home of a set of elderly. Additionally, to construct the system, video cameras were also installed (however in the final commercial system, no cameras were needed). The video was later monitored by human experts and situations were labelled as ”fall” (typically when there was very little large-scale movement), ”confusion” (typically erratic movement from room to room) or ”normal”. The time-segments corresponding to these labelled examples were then extracted from the set of sensors and the temporal nature of the data was converted into a spatial pattern using 8 dimensions per sensor. This means that that for each separate time-segment there would be the data from the set of sensors as well as the human classification into one of the three classes. In total over all test persons and the entire time data was collected, 8913 time-segments were produced. The system was then to be connected to an alarm service so that ”fall” would invoke an immediate visit, ”confusion” would result in a visit during the day, and ”normal” would not change the visit schedule.

Question 10 (6p)

Please design an empirical study to resolve a question as to which neural network approaches are suitable for forecasting the consumption of electrical energy in one of Stockholm's largests residential areas. Please bear in mind that the focus is on short-term prediction, max 24 hours ahead. The dataset for your study contains samples of measured energy load on an hourly basis from the last three years with large seasonality trends removed (data normalised across four seasons of the year) and remaining strong weekly as well as daily periodicity. When describing your experimental study please keep the following questions/issues in mind:

- What could be candidate neural network types for your comparative study? What configurations would you consider worth testing (how would you go about parameter settings and what parameters)? What would you consider a fair comparison in terms of network configurations?
- How would you suggest using your data, how would you use/divide these roughly 26280 (3x365x24) samples to ensure reliable training, validation and testing?
- What would you feed to the networks, what would be inputs and outputs? You could support your short description with suitable illustrations.
- What would be the criterion/criteria for your comparison (what would you measure)? How would you ensure that your findings account for stochastic nature of training neural networks so that generalizable conclusion could be drawn (rather than being an unreliable product of random effects)? How would you demonstrate your study outcomes?

Question 11 (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? What is the fundamental conceptual difference in processing data samples/handling data to identify data representations suitable for a given pattern recognition task between shallow and deep neural network approaches?

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.