# DD2437 Online Re-exam Answers, 21HT (2021/12/17)

## Part I - Quiz **(23p)** – *please see the corresponding set of questions where correct answers are marked*

## Part II **(23p)**

### **Question II.1** (3p)

Since the experts do not agree on the labelling and the number of music styles the recordings should be categorised into, an explorative analysis without any imposed labels should be performed. For such an unsupervised learning scenario a self-organising map (SOM) appears as a suitable approach. It allows for topographic mapping of music recording representations fed into input nodes to the activations of selective output nodes organised into a grid, i.e. similar recordings should cause activation of output nodes, best matching units, close to each other in the grid (local neighbourhood). The representation of a music recording as input could be either a set of low-level features, e.g. time-frequency map, Fourier transform, or time series/sequence representation, potentially pre-processed by a recurrent network.

The granularity of mapping (how distinct these activated nodes in the output grid are depending on the similarity of the inputs) is determined in the process of learning by a parameter controlling the gradual decay of the size of the neighbourhood in the grid space as well as the learning rate. The learning algorithm this way organises the output grid into groups of units (a set of local neighbourhoods) that could be associated with categories found for input music recordings. The number of such clusters and their distinctiveness (how selective the best matching units are) correspond to the aforementioned "granularity" of mapping established during learning. The association of clusters with music styles is made by experts based on the patterns of output activations.

After the learning phase experts could use the SOM to study clusters of music styles etc. by feeding the network with music recordings as input data representations and observing patterns of best matching unit activations in the output grid space. This way they could study if the music pieces produced by the same composers cause activations of the same (or nearby) output units, previously associated by experts with a given music style.

## **Question II.2** (4p)

*A coherent and a well formulated answer providing the following details is expected:*

Ad.1)

- It is a classification problem.
- I use an MLP trained using the generalized delta rule.
- Input is the data from sensors, registers etc.
- Output is the type/class given for the data (Normal operation, one of a set of common error codes).
- Use saved class labels as targets to train the net.
- Each data channel is separately normalized in [0, 1].

Ad.2)

- This is an optimization problem similar to the traveling salesperson problem.
- I use a Bolzmann machine.
- I set up a cost for visiting each estate (medium weight, proportional to distance to the estate) and a cost for missing an estate (large cost).
- I run the system, with a large temperature at start and low at end.
- For each temperature, I iterate over all nodes many times (to reach "thermal equilibrium").
- Alt. I use a SOM like in the lab.
- I use a cyclic connectivity to enforce that a closed loop is created.
- Use as many nodes as estates, or more to make more likely all estates are visited. If more are used, remove nodes that do not represent an estate.
- Start with a large neighbourhood and shrink during iterations.
- Initialize nodes randomly or according to the center of gravity of the data..

## **Question II.3** (3p)

The proposed network for modelling such memory phenomena is a Hopfield network. It could be trained with a Hebbian type learning rule.

Memory recall in such auto-associative memory networks is conducted by providing the input corresponding to the memory pattern itself or its noisy/incomplete version (content addressable memory) and performing an iterative update (synchronous or asynchronous) until the network converges.

Once the network approaches the storage capacity, new memory patterns will be encoded/memorised at the cost of those already stored patterns. At some point however, a classical

Hopfield network trained with conventional Hebbian learning will suffer catastrophic forgetting (its performance will drop rather abruptly). (There are heuristic modifications of the Hebbian-like learning rule and strategy that render the associative memory networks more robust.)

The number of unique memory patterns that the network is reliably able to store is determined by the size of the network, learning rule as well as the level of orthogonality of patterns (this is rather a data property).

## **Question II.4** (5p)

The problem is a classification problem.

I use an MLP. Motivation of MLP is that the temporal nature of the data is not that complicated, there are no long-range correlations to be found as the person will be in one of the 3 states for the entire time-window of one data sample.

Topology is a typical N layer MLP network, (number of hidden layers determined by grid search).

Input patterns come from the set of sensors (number of nodes=8*number of sensors).

I use one or several layers of hidden nodes (with number of nodes determined by grid search), and

an output layer of 3 nodes corresponding to the classes ”normal”, ”fallen”, ”confused”.

The target labels (training data) comes from the labels generated by the experts.

I use backpropagation of errors and early stopping.

Optimize and estimate generalization. I use 5-fold CV. I run a grid search including number of hidden layers, number of hidden nodes, learning rate to find the hyperparameters that give the best generalization.

Key challenges and potential difficulties: Unbalanced data, class “Normal” will probably be 99% of the data.

## **Question II.5** (5p)

The problem to address is a time series prediction with focus on short-term horizon. Potential neural network approaches to test are among others feed-forward multi-layer perceptrons (MLPs) with time-lagged representations and recurrent architectures with long short-term memory (LSTM) networks as suitable candidates.

Since the aim is to predict 24 steps ahead, a simplified approach would be to produce one prediction at a time (the 24th step ahead) and a somewhat more complex problem (requiring more training and likely more tunable parameters) would be to produce 24 predictions at a time (the 1st, 2nd, ... 24th step). So, the basic configuration is either a sequence-to-one or sequence-to-sequence mapping. Importantly, the number of inputs largely depends on the historical context to be taken into account for each prediction. As we are aware of strong daily periodicity one option would be to offer a full 24 hour input either at full hourly resolution or subsampled. However, there is also a weekly periodicity trend, so another option could be to incorporate input samples from previous days (potentially at a lower temporal resolution).

Some key network hyperparameters for both MLPs and LSTMs are the number of hidden nodes and layers, learning rate (and momentum), activation functions, dropout (especially for LSTMs). It would be relevant to search for best network configurations resorting to, for example, a grid search.

There are different aspects of the fairness of the proposed comparison between selected networks: the same problem specification in terms of input-output mapping, comparable complexity of the models (quantified by, for example, the number of tunable/trainable parameters – on the one hand we want to optimise our network candidates and, on the other hand, we should account for the complexity of the resulting models in the comparative analysis), the same performance measures.

Prediction accuracy could be measured with the mean squared error. An estimate of the generalisation performance should be made with cross-validation (important for model selection and network comparisons). To this end, 3x365 weeks of data could be split into, say 6 half-year consecutive blocks, to run 6-fold cross-validation. Apart from the prediction performance one could also use computational load of the training process as a comparative criterion. To provide reliable results a statistical analysis should be performed accounting for stochastic effects of weight initialisation and data splits among others. To this end, multiple repetitions as well as cross-validation (to study generalisation performance) should be made. The mean results along with the variability (second moment like variance) should be presented, and statistical null hypothesis testing should be employed for the final network comparison.

## **Question II.6** (3p)

We refer to deep belief networks (DBNs) as hybrid generative and discriminative models as the same model can be explicitly used for both classification and sample generation.

Generative capabilities are owed to a probabilistic representation of the joint (input+output) data distribution with the support of latent variables (hidden units) (0.5p). Discrimination is facilitated

by modelling the projections to the class label layer (thus estimating class conditional probabilities P(output | input)).

*Here a description of pre-training a DBN for a simple binary discrimination task comes.*