DD2437 Online Re-exam Answers, 22VT (2022/06/09)

Question 1 (6p)

Based on the assumption that the data forms “multi-dimensional clouds” with the described discontinuities along each data dimension, characteristic for some form of data clustering, I would suggest the use of radial basis functions (RBFs) to model these clusters. Consequently I would choose a simple feedforward RBF network with the hidden layer composed of many RBF units, each feeding to the output layer. (2p)

The number of output units should correspond to the number of classes, i.e. four: no basic, advanced, premium service and no service. Since it is a classification problem I suggest sigmoidal activation function for the output units. The key hyperparameters are the number of RBF units (potentially their width if it is considered fixed) and the learning rate. The learning algorithm consists of the data dependent RBF unit initialisation based on either some clustering algorithm (competitive learning) or expectation maximisation (EM) algorithm. Although the initialisation is focused on identifying RBF centres/positions in the multi-dimensional input space (corresponding to some cluster centres), there is an important decision to make as to whether the width of RBF (transfer/activation function) is learnt during the initialisation too (e.g., EM algorithm can do that) or it is initialised to a fixed value. Similarly, after the initialisation during actual weight tuning with a gradient descent algorithm (with the primary focus on the hidden-to-output weights), it should be decided whether RBF widths and positions are also simultaneously adjusted/learnt or kept fixed (tuning only the weights from the hidden to output layer). (2p)

Given that only 10000 data samples are available and the dimensionality of the input space is in the order of tens of features, it might be too optimistic to assume that all free parameters can be reliably tuned. To properly validate this approach it is suggested that the 10-fold cross-validation is performed (10 iterations of 80% training – 10% validation – 10% testing split) on 80-90% of all available data (8000-9000 samples) with the cross-entropy measure as the loss function. The cross-validation is convenient for hyperparameter (model) selection. The final evaluation of the generalisation performance for the selected model should be performed on the remaining 10-20% of data (1000-2000 samples) following the training on all 8000-9000 samples (previously used for the cross-validation). (1p)

If the assumptions about the data distribution were not strictly met, the RBF network should still be able to perform well after the free parameter tuning (training with gradient descent) due to its approximating capabilities. (1p)

Question 2 (6p)

Generative neural networks are aimed at learning of a full probabilistic model of the data over the joint input and output spaces, P(x,y). This is typically achieved by introducing latent variables that facilitate building the full model as they help in capturing key underlying characteristics ("causal factors") of the data. Discriminative models only learn conditional probability to infer the output and therefore they are most often trained with a supervised learning approach whereas generative models also involve unsupervised learning (without labels). (1p)

The key advantage of generative models is that besides making predictions (like discriminative models) they can also be used to generate new data (sampling based on the latent and output space representations). Generative models also help in understanding how the data was generated to categorise an object (what are the factors that make data represent one object and not some other ones). The scope of applications is immense from data augmentation/generation/denoising, image inpainting, image-to-image translation etc. (2p)

The key capability to model the underlying data distribution (one of the most defining characteristics) is achieved in different ways for DBNs, VAEs and GANs.

-> The probabilistic model in DBN is learned with a contrastive divergence algorithm (approximate max likelihood) where the activity of visible and hidden nodes is stochastic governed by Boltzmann's probability distribution. (1p)

-> In VAEs the distribution of the data is learned via the latent space where latent variables are considered as random (and are subject to sampling) with the underlying parametric distribution, e.g. Gaussian. The parameters of these latent variables' distributions are learnt using backprop implementation of gradient descent (with the loss function corresponding to the sum of the reconstruction error, as in typical autoencoders, and a variational term imposing constraints on the parameters of the latent variables' probability distributions) (1p)

-> In GANs, it is a generator network that learns implicitly the data distribution. In essence, it amounts to sampling from a complex distribution that is implemented by sampling from a simple noise distribution that is transformed to match the data distribution. In that way, it is a nonparametric (unlike in VAEs) probability estimation via the generator, led by the loss function with the contribution of the classification error incurred by the discriminator network. The learning of the generator and discriminator is a joint process coupled with the loss function. (1p)

Question 3 (12p)

A) The identification of the vigilant vs pre-sleep phases.

- The simplest approach would be to classify each sample (spectral image corresponding to 2-s snapshot of 32-channel data) independently with a convolutional neural network (CNN). The number of inputs corresponds to the frame size with RGB coding: 16*32*3 and the number of sigmoidal output units is 3 (as the number of classes). The key hyperparameters are concerned with the model size/architecture and learning, e.g. learning rate. (1p)

- The gradient descent learning (backprop algorithm in particular) is recommended with cross-entropy loss function. The cross-validation is recommended for validation purposes and the data could be split based on human subjects (112 subjects split into 8 folds, 14 subjects each). (1p)
- The cross-validation suggested above helps evaluate the generalisation across human subjects (the most challenging one due to the heterogeneity of human subjects). Other relevant forms of generalisation are across sessions. It could be examined separately for each subject by splitting the session data to implement a cross-validation scheme. The amount of data per subject may not be sufficient though so a compromise would be to mix subjects and sessions to make a global cross-validation across sessions (stratifying to have roughly homogenous subject distribution across folds). (1p)
- The performance could be measured by the area under the ROC (AUC-ROC) or a combination of sensitivity and specificity (e.g. F1-score) to account for the uneven distribution of classes. (1p)
- The key assumption is the independence of data segments treated as separate samples here. It is a conservative assumption but this approach does not suffer from an obvious violation of this assumption. At the same time, this classification approach, given the temporal dependence between the samples, is suboptimal. Also, data processing and low-level representation assumes signal stationarity within those 2s windows. (1p)
- The challenges are concerned with inadequate number of sleep episodes per subject, considerable class imbalance, heterogeneity across subjects and variability across sessions among others. (1p)

B) To follow up the process of falling asleep the transitions between the data segments have to be explored.

- To handle temporal context a combination of a CNN approach with a recurrent network could be implemented. The inputs would the same frames as used in the A) approach, also the outputs are matched to address the same classification problem in the end. The learning approach would have to incorporate the integration of gradients over time so it should rely on backprop through time.

  However, to exploit the approach proposed in A with the aim of studying the transitions between frames (rather than building a new network that explicitly takes into account temporal correlations), it is much simpler, faster and cheaper in terms of compute resources to set up a self-organising map (SOM) using the representations extracted with the CNN network in A. The number of inputs then corresponds to the size of the selected hidden CNN representations and the output nodes could be organised into a classical 2D SOM grid (potentially 3D if more complex manifolds are needed to capture the state transition dynamics). The training builds on competitive learning with neighbour based updates in the grid (*here a bit more detailed explanation would be appreciated*). (2p)

To study the transitions, one has to identify best matching units (BMUs; the grid units that become most active for a given sample – "the winning" units) in the grid that correspond to the specific classes/states with the most important status of those early sleep segments (that directly follow the pre-sleep segments). Once the key BMUs are established the data should be applied in the recall mode to observe if the trajectory towards the sleep related BMUs follows a smooth trajectory of activated units in the grid (smooth corresponds here to the activation patterns flowing through neighbouring units rather than abruptly switching between distant units in the SOM grid before arriving at the sleep related BMUs). It is also interesting to study whether there are separate sleep related BMUs in different parts of the SOM grid, potentially with different patterns of neighbouring BMU activations converging on the target sleep BMU. (1p)

<u>(In summary, *2p fo SOM-like approach and 1p for RNN-CNN based approach + 1p extra should be given for describing how the network (SOM or RNN-CNN) can be used to study the transition patterns, i.e. a differential hypothesis stated in the question*)</u>*.*

- The most relevant generalisation to inform the approach here is concerned with the performance transfer between human subjects (besides the generalisation across sleep data segments but this is in fact what we want to study here and we did not explicitly quantify in A) – if there is lack of consistency between subjects (and potentially sessions) then no repetitive patterns of transitions can be observed. (1p)
- Extending the SOM to perform classification (brain state identification) implies the need to associate SOM units in the output grid with specific class labels. This should be done on a subset of data (training data) and validated/tested on hold-out sets (a similar approach could be adopted as proposed in A). The simplest way to assign these labels is to associate each unit with the most dominant label among those that represent input samples driving this particular unit to be the BMU, the winning node in the grid (for each sample there should be one BMU and it is likely that samples representing different classes may converge/lead to the same BMU – hence the need to define the most dominant label for each relevant unit in the grid). As mentioned, it is important to make sure the network produces one clear winner - BMU per input sample. This can be done by sufficiently narrowing the neighbourhood in the grid space towards the end of the training process. (2p)

Question 4 (6p)

Averaging the outputs of weak learners reduces the error variance if the error produced by individual learners are statistically independent. (1p)

This implies that the expected value of the cross-term / interaction between errors generated by learners (covariance) is 0. Consequently, the variance of the average of independent variables is proportional to the mean variance divided by the number of independent variables. (2p)

Boosting is an example of an ensemble learning method. (1p)

The corresponding pseudocode is as follows: (2p)

```
1) At the very beginning all the samples have the same weights 1/N.
2) Repeat L times (# learners in the sequence):
    a. train the best weak model with the current samples weights
    b. compute the value of the update coefficient that is a scalar evaluation metric of the weak learner that indicates how much this weak learner should be taken into account into the ensemble model
    c. update the strong learner by adding the new weak learner multiplied by its update coefficient
    d. compute new samples weights that inform which samples we should focus on in the next iteration (weights of samples wrongly predicted by the aggregated model increase and weights of the correctly predicted samples decrease)
3) Aggregate the L models built sequentially into a linear combination weighted by coefficients corresponding to the performance of each learner.
```

Question 5 (8p)

This is a classification problem. I use an MLP motivated by the small computational resources. (1p)

Input is data from the 64 sensors. (1p)

Output is a layer of 7 nodes, each of which giving a binary 0/1 output and capable of together coding for 128 different classes. We have 60+55+1 (not known) = 116 compounds/classes. (1.5p)

I try different number of nodes in the hidden layer during evaluation to learn which number gives the best generalization performance estimated by means of cross-validation. (1.5p)

First I normalize the data separately for each sensor to [0-1]. Training is done using sample data as input and known substance as target using Backprop. I split data following N-fold CV. (1p)

To improve generalization, I test different number of hidden nodes (see above). I also add noise to the input values. (1p)

Challenges includes a possible class imbalance (as often is the case). (1p)

Question 6 (8p)

a) For still images it is recommended to use a convolutional neural network (CNN) whereas for the video clips one could employ a synergistic combination of a CNN with a recurrent network, e.g. long short-term memory (LSTM) or a CNN wit a 3D convolutional kernel to account for time (one could also consider a two-stream network where spatial and temporal dimensions are processed by two independent CNNs). (2p)

b) Video-clips are composed of a set of images so by definition more data is needed for a video-clip approach. Additionally, due to a higher effective dimensionality of the video-clip samples when compared to still-image samples, the underlying network for video-clips has the higher number of degrees of freedom, which additionally contributes to the larger demand for samples. Furthermore, to help generalisation the collection of videos should be procedurally constrained so that similar type of videos are taken. There are also constraints (spatial – similar object(s) should feature) for still images but it is easier as there are no extra constraint for repeatable temporal structure. (1p)

c) As for the shared concerns between the two data-driven diagnostic approaches, the great challenge is generalisation considering the heterogeneity of subjects/patients (clinical/biological diversity). In addition, the data are typically collected by different examiners, potentially in different labs using different equipment etc. – it all contributes to the heterogeneity that is challenging for the generalisation. (2p)

d) Ultimately, the two approaches can only be compared using diagnostic specific criteria. One option is the diagnostic accuracy but it is prone to imbalanced classes so a better approach is ROC curve (area under the curve) or F1-score that accounts for both specificity and sensitivity (alternatively both of these measures can be reported). These measures can be reported either on a subject basis (where each subject is treated as a sample unit for evaluation) or on a sample basis (an image vs a video-clip). It is recommended that for fairness the subject-based evaluation is used. These evaluation approaches are built on top of the diagnostic networks, which themselves may be trained with different loss functions depending on how a video-lip is processed and classified (ultimately, however, the loss could be a cross-entropy function for gradient descent based learning in both networks). (2p)

e) Other considerations may account for different criteria for the quality assessment of images vs video-clips. It may also be worth examining if the video-clips could be composed of lower-resolution images as the temporal correlations could compensate for the potential loss in spatial information (naturally the question of temporal resolution should also be raised) (1p)