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

Question 1 (6p)

An internet service provider has collected a vast amount of data about the users. In other words, for ca. 10000 users 27 demographic features alongside 12 quantifiers describing the service usage pattern are available. The service provider is interested in better understanding the user base and, more importantly, predicting if they are going to continue the service after the first 6-month period free of any charges. Since there are three versions of the available paid service: basic, advanced and premium, the provider ask you for help in developing a neural network based system for predicting the user's decision about continuing the service and in what paid version. Upon closer inspection of the data you discover that many of the features, both demographic and quantifiers of the service usage, are discontinuous and form clearly separable intervals (the distribution of each dimension), for example age is well divided into 18-22, 24-29, 32-45 and 50-80 age groups. Since it is puzzling you visually inspect the data in various subspaces – two- and three-dimensional projections of different feature combinations. As a result, you discover to your dismay that a lot of these projections reveal cluster-like grouping of the data in respective subspaces. You derive an assumption that the data form distinct clusters (multi-dimensional clouds). What neural network approach would you propose to address the challenge posed by the service provider? How would you design the network, what are the key hyperparameters and how would you optimise them? Please describe what learning algorithms you would choose and how you would proceed with the training, validation and testing the network model's generalisation capability. Finally, what would be likely consequences of deploying your model to the given prediction problem if the assumption about the data distribution did not strictly hold. Please motivate your answers.

Question 2 (6p)

Generative models have received growing attention in the deep learning community mainly due to their utilitarian aspects and applicability. Could you briefly explain what makes generative models particularly attractive and what applications they are suitable for, especially when contrasted with discriminative models? What is/are the most defining characteristic(s) of generative models and how are they implemented/manifested in deep belief nets (DBNs), generative adversarial networks (GANs) and variational autoencoders (VAEs)?

Question 3 (12p)

You are involved in a scientific project where cognitive neuroscientists collect recordings of electrical activity of the human brain in the attempt to identify activity patterns that can help predict the oncoming phase of dozing off (a short lasting, a few seconds, transition into shallow sleep). To this end, they have collected data from 10 session for each one of 112 human subjects, where each session is of varying duration depending how quickly the subject dozes off (there can be basically not more than 3-4 episodes of dozing off per session, after each such episode the subject is alerted by loud sound to return to the vigilant state). The recordings were conducted in the same lab by the same examiner using the same recording equipment, which produces 32 channels of electrical signals at the temporal resolution of 1 ms (1e-3 s). Since the raw signals have been shown to be noisy with excessive sampling rate it is assumed that the fast Fourier transform performed in non-overlapping time windows of 2s offers good representation. As a result, the data for each 2-s window is a matrix 32 (channels) by 16 (frequency subbands), and each such 32-by-16 frame is annotated by the examiner with one of three possible labels: vigilance, pre-sleep (just about to doze off), sleep. So, for each subject per session there are on average 3.5 episodes of falling asleep (pre-sleep lasting 6-10 s and the subsequent sleep lasting 15-20 s) plus 20 minutes of vigilance in total (before and after sleep episodes). Here are the two main tasks (A and B) that you have been requested to perform using a neural network approach on the data described above (please answer the questions listed for each task and provide motivation for your choices).

A) Develop and validate a system for the identification of the vigilant vs pre-sleep phases.

- What neural network architecture would you choose, what is the number of inputs and outputs and what are key hyperparameters?
- What learning algorithm would you suggest and how would you perform the learning and validation process, how would you use and split the data for this purpose considering that you should also optimise hyperparameters?
- What are key generalisation issues to validate, what generalisation types would you want the system to be good at (or at least you would want to check if this is the case)? How would you use the data to validate the generalisation performance in different scenarios (this is related to the earlier question about the training and validation process)?
- How would you quantify and measure the performance?
- Are there any assumptions that you make about the data in your approach?
- Importantly, what are key challenges, potential problems, risks (regarding the data) that you can identify already at the planning stage?

B) Given that the neural network method proposed in A provides a rigid identification of pre-sleep vs vigilant state, it does not provide much insight into the actual process of falling asleep. In particular, you would like a method that helps the scientists verify if the activity patterns prior to the pre-sleep phase or maybe those within the pre-sleep phase are similar to each other, maybe there is a trajectory of similar states that eventually evolves into the sleep pattern. For this you do not need a rigorous method but rather a neural network based tool that provides these insights or, in other words, that helps you qualitatively answer the question whether the brain activity smoothly develops (implying that one pattern activity evolves into another similar one etc.) or rather abruptly switches (between dissimilar patterns) as we approach dozing off.

- What would be your idea for a neural network solution if the simplicity over scientific rigour is preferred (a simple solution that just offers qualitative insights rather than quantitative assessment)? What are the inputs and outputs, and what learning approach do you propose?
- In what way the lesson you learnt about the generalisation in A would be valuable in studying the problem here in B?
- How would you in a simple way (again, without requiring scientific rigour) extend your approach to perform the brain state identification requested in A with the sole purpose of simplistic validation of the network proposed here in B (just to verify that it acts reasonably well and offers reasonable pre-sleep vs vigilant state identification, still expectedly worse than in A). How would you make the network produce the kind of output that is suitable for measuring the performance?

Question 4 (6p)

One way to combine the outputs of component networks is simple averaging. What effect does it 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? Please motivate your answer with arguments of statistical nature.

What other ensemble learning approaches (meta-algorithms) besides the aforementioned bagging could you propose? Please name at least one of them and provide the corresponding (simplified) pseudocode.

Question 5 (8p)

Your team has received a task to construct a system for one of the European space agency's programs using micro-satellites. The satellite will be sent to the recently discovered asteroid 2010 WC9 197 to study what it is composed of. The idea is to equip the satellite with a small chemical analysis lab so that the chemical nature of the asteroid can be analyzed. The lab has been designed to use very little computational resources as the on-board computer which is used for all computations has very small memory and low computing power. The lab uses a novel array of 64 sensors, where sensors are uniquely sensitive to different combination of chemical components (building blocks of substances) and have their specific output ranges (ranges of values produced as an output response). The objective is to use the lab to identify what is the single most dominant chemical substance in each sample. The combination of output from all 64 sensors is estimated to give a unique detection of all the 61 known substances (from previous explorations) plus a range of 55 substances we only imagine could be present. The sensor array has been used on well-characterized space samples and samples from the earth. The total of 648000 samples together with their known substances (pairs of sample and substance) have been collected.

What type of problem does the company want to address? Accordingly, propose a neural network approach to meet the objective formulated above. In particular, motivate the choice of your network type, briefly characterise its topology clearly indicating the inputs and outputs, and describe how the network should be trained – how the data should be used and what learning algorithm you recommend. Please, explain also how you would optimise and estimate the generalization capacity. Finally, identify key challenges and potential difficulties/risks concerning the problem and your neural network approach.

Question 6 (8p)

Biomedical image analysis has become an essential part of many clinical diagnostic pipelines. Typically a set of images is collected for various cohorts of patients and healthy individuals as a control group (with different imaging approaches depending on the available technology, clinical context and economical constraints), then they are subject to domain specific preprocessing, augmentation and, finally, automated analysis and interpretation for diagnostic purposes (as a support for the clinical diagnosis). In practice, static biomedical images used for training deep neural networks have to be manually processed or selected to ensure adequate image quality. To handle this issue, there are new methods supporting these manual efforts but they are often excessively conservative. Growing attention in the community is thus turned to automated deep learning based video-clip rather than static image analysis. On the one hand, there are obvious advantages of this approach from the clinical perspective. On the other hand, they are technically more challenging and susceptible to multiple risks.

From a methodological perspective, especially considering the neural network's point of view, please compare approaches to the problem of clinical diagnosis based on still images vs video-clips (sequence of images collected at high temporal resolution in the examination process). To guide your answer, please address the following points in particular:

a) what family of neural networks would you propose to support clinical diagnosis based on pre-selected still images vs video-clips?

b) what are fundamental differences between the two approaches (image vs video-clip classification) in terms of data demands, their characteristics and the resulting challenges for the respective neural networks? (obviously images are different video-clips but what are implications of working with video-clips for the data collection process and then for developing a robust classifier with good generalisation potential when compared to still images)

c) what are shared concerns and risks incurred by the deployment of neural network methods for image vs video clip based diagnosis support?

d) what evaluation criteria would you adopt to systematically compare the neural network based approaches to image vs video-clip based clinical diagnosis support?

e) any other thoughts, considerations, important differences?