Elements of ML and DS
WS 2023
Exercise Sheet 5

Exercise 5: Neural Networks & Hyperparameter Optimization

due **on** 2024-01-27

Important information regarding the exercises:

- Use the Moodle system to submit your solution. You will also find your corrections there.
- Upload your pen & paper solutions as PDF. If your solutions are handwritten, ensure that your handwriting is legible and that the pictures are not blurred and are taken under appropriate lighting conditions. All non-readable submissions will be discarded immediately.
- Submit coding exercises in a separate Python file.

Problem 1 (Neural Networks, 13 Points)

a) Consider a multi-layer perceptron (MLP) with $N > 1$ layers, each equipped with bias. The input vector size is denoted as D, and the output vector size is denoted as K. The size of the hidden layers is denoted as H. The logistic sigmoid function σ is employed as the activation function. Determine the total number of learnable parameters in the network. (**3 Points**)

b) Provide the closed-form expression for the forward pass of the network mentioned in part a) for the case of $N = 3$. Here, $\mathbf{x} \in \mathbb{R}^{D\times 1}$ denotes the input, and $\mathbf{y} \in \mathbb{R}^{K\times 1}$ represents the output. The weight matrices are denoted as $\mathbf{W}_i$, and the bias vectors as $\mathbf{b}_i$, where i ranges from 1 to N. Additionally, specify the shapes of every $\mathbf{W}_i$ and $\mathbf{b}_i$. (**3 Points**)

c) Suppose we use a linear activation function in an MLP. Demonstrate that this modification results in the network collapsing into a single linear layer. In other words, we can represent the MLP's function equivalently as a single linear layer defined by a solitary weight matrix $\mathbf{W}$ and bias vector $\mathbf{b}$. Explain why this outcome is undesirable. (**4 Points**)

d) Describe the process of adjusting the weights of a neural network to the training data. (**3 Points**)

Problem 2 (Hyperparameter Optimization, 7 Points)

a) Random Search and Grid Search are two common baselines for hyperparameter optimization. Explain in one or two sentences how grid search works. (**1 Point**)

b) In the situation when on a same problem Grid Search and Random Search obtain a similar budget, Random Search can in many cases come with a better result. Explain why this is, and what are the situations that this happens. (**1.5 Points**)

c) Bayesian optimization is a commonly used method for hyperparameter optimization. Important concepts of Bayesian optimization are the initialization phase, the surrogate function, and the acquisition function. Explain how Bayesian optimization works in words with at most 6 sentences. Explain in particular the initialization phase, the surrogate function, and the acquisition function. (**2.5 Points**)

d) Consider the following pseudo-code for Bayesian optimization:

```
for $t$ = $0, 1, \dots$ do
    Find $x_t$ by optimizing the the acquisition function over the Gaussian Process:
      $x_t = \text{argmax}_x u(x|D_{1:t-1})$
    Sample the objective function $y_t = f(x_t) + \epsilon_t$
end
```

where x is the input, u is the acquisition function, $D_{1:t-1}$ is the data, y is the target, f is a black-box function and ϵ is some Gaussian noise.
What crucial step are we missing here? (**1.5 Points**)

e) Please answer the following multiple-choice question by giving the letters of the correct answer. Multiple answers can be correct. Partial points will be deducted for incorrect answers, but the question will count at least as 0 points.

Hyperparameters can be numerical or categorical. Which of the following are examples of categorical hyperparameters? (**0.5 Points**)

A: Random State of a Randomized Classifier.
B: Gamma of a Support Vector Machine.
C: Complexity of a Support Vector Machine.
D: Split Criterion of a Decision Tree.