Aprendizagem 2023
Homework I — Group 003
(ist1107028, ist1107137)

Part I: Pen and paper

For questions in this group, show your numerical results with 5 decimals or scientific notation. Hint: we highly recommend the use of `numpy` (e.g., `linalg.pinv` for inverse) or other programmatic facilities to support the calculus involved in both questions (1) and (2).

Below is a training dataset D composed by two input variables and two output variables, one of which is numerical ($\mathbf{y_{num}}$) and the other categorical ($\mathbf{y_{class}}$) . Consider a polynomial basis function $\phi(\mathbf{y_1}, \mathbf{y_2}) = \mathbf{y_1} \times \mathbf{y_2}$ that transforms the original space into a new one-dimensional space.

D	y_1	y_2	y_{num}	y_{class}
x_1	1	1	1.25	B
x_2	1	3	7.0	A
x_3	3	2	2.7	C
x_4	3	3	3.2	A
x_5	2	4	5.5	B

1. **Learn a regression model on the transformed feature space using the OLS closed form solution to predict the continuous output variable $\mathbf{y_{num}}$.**

2. **Repeat the previous exercise, but this time learn a Ridge regression with penalty factor $\lambda = \mathbf{1}$. Compare the learnt coefficients with the ones from the previous exercise and discuss how regularization affects them.**

3. **Given three new test observations and their corresponding output $\mathbf{x_6} = (\mathbf{2}, \mathbf{2}, \mathbf{0.7}), \mathbf{x_7} = (\mathbf{1}, \mathbf{2}, \mathbf{1.1})$, and $\mathbf{x_8} = (\mathbf{5}, \mathbf{1}, \mathbf{2.2})$, compare the train and test RMSE of the two models obtained in (1) and (2). Explain if the results go according to what is expected.**

4. **Consider an MLP to predict the output $\mathbf{y_{class}}$ characterized by the weights**

$$W^{[1]} = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0.1 \end{bmatrix}, \quad b^{[1]} = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix}, \quad W^{[2]} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad b^{[2]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

the output activation function

$$\text{softmax}(z_c^{[\text{out}]}) = \frac{e^{z_c^{[\text{out}]}}}{\sum_{l=1}^{|C|} e^{z_c^{[\text{out}]}}}$$

, no activations on the hidden layer(s) and the cross-entropy loss:

$$\text{CE} = -\sum_{i=1}^{N}\sum_{l=1}^{|C|} t_l^{(\text{i})}\log(Z_l^{[\text{out}](\text{i})})$$

Consider also that the output layer of the MLP gives the predictions for the classes A, B and C in this order. Perform one stochastic gradient descent update to all the weights and biases with learning rate $\eta = 0.1$ using the training observation x_1.

Part II: Programming

Consider the parkinsons.csv dataset (available at the course's webpage), where the goal is to predict a patient's score on the Unified Parkinson's Disease Rating Scale based on various biomedical measurements. To answer question (5), average the performance of the models over 10 separate runs. In each run, use a different 80 − 20 train test split by setting a `random_state = i`, with i = 1...10.

5. **Train a Linear Regression model, an MLP Regressor with 2 hidden layers of 10 neurons each and no activation functions, and another MLP Regressor with 2 hidden layers of 10 neurons each using ReLU activation functions. (Use `random_state=0` on the MLPs, regardless of the run). Plot a boxplot of the test MAE of each model.**

6. **Compare a Linear Regression with a MLP with no activations, and explain the impact and the importance of using activation functions in a MLP. Support your reasoning with the results from the boxplots.**

7. **Using a 80 − 20 train-test split with `random_state=0`, use a Grid Search to tune the hyperparameters of an MLP regressor with two hidden layers (size 10 each). The parameters to search over are: (i) L2 penalty, with the values $\{0.0001, 0.001, 0.01\}$; (ii) learning rate, with the values $\{0.001, 0.01, 0.1\}$; and (iii) batch size, with the values $\{32, 64, 128\}$. Plot the test MAE for each combination of hyperparameters, report the best combination, and discuss the trade-offs between the combinations.**