Final programming assignments Basics of Artificial Intelligence

Mgr inż. Jolanta Podolszańska

1 Ogólne informacje

During the programming assignment you will design a program in Python based on the *data.csv* dataset, which you will find in the same section as the assignments. Please read the questions carefully. Using **Chat GPT** to generate solutions is **forbidden**. You can use the prepared notes for the colloquium I discussed in class. Ultimately, you can use StackOverflow in case of syntax error, not looking for a solution to tasks.

The maximum number of points to be scored is 12 points. Credit for the programming assignment is from 6 points.

2 Zadania

Please perform the following tasks.

1. Load the file mushrooms.csv into a Pandas DataFrame. Display the first few rows (df.head()), Print column names (df.columns), Check the number of samples and columns (df.shape) (2 points)

2. Identify the column that serves as the label (often class), containing 'e' (edible) and 'p' (poisonous). Map the values 'e' to 0 (edible) and 'p' to 1 (poisonous). That way, you have a binary label (0/1) (2 points)

3. Exclude (or separate) the class column from the DataFrame, so that X contains only the mushroom attributes (e.g., cap-shape, odor, gill-size, etc.). Apply One-Hot Encoding (e.g., pd.get_dummies) to the remaining columns (since they are all categorical). Perform dimensionality reduction to 2 components using PCA(n_components=2). This will give you a (n_samples, 2) matrix, which you can easily plot in 2D (3 points)

4. Split the data into three parts: Training (70%), Validation (15%), Test (15%). Use train_test_split twice to achieve this split. Scale the 2D PCA-transformed data using StandardScaler. Perceptrons often benefit from scaling, as it can help the algorithm converge more effectively. Print the

sizes of the train/val/test sets to ensure they sum up to the total number of samples (3 points).

5. Train a perceptron (e.g., Perceptron(max_iter=5000, eta0=0.01, random_state=42)) on the training set. Measure accuracy on both the validation set and the test set using accuracy_score. **Note** whether the model achieves high accuracy (the Mushrooms dataset is often separable).

6. Divide the points into class 0 (no disease) and class 1 (disease), give them different colors (e.g. red / blue). Calculate and draw the decision line of the perceptron:

$$y = \frac{b + w_1 x}{w_2}$$

where w1, w2 = perceptron.coef_[0] and b = perceptron.intercept_[0]. Use the library *Matplotlib* for this. (2 points)