Johannes Brinkrolf Bielefeld University

Introduction to Machine Learning (WS 2018/19)
5. Assignment

Released: Thursday, 15.11.2018.
Due: Please solve the exercises in groups of three and present your solutions in the GZI V2-222 (student computer pool) on **Monday, 26.11.2018**, at the assigned time.

- Exercises that require a documentation are marked with **[DOC]**.
- The python toolbox needed for solving the practical exercises can be found at `http://scikit-learn.org/stable/`. It comes with an excellent online description and demos. You can also look at the given demo code.
- On the computer in the GZI you just need to enable *mltools* in the RCINFO Package Selector. You will have a new version of sklearn (0.20.0), updated Spyder 3.3.1 develoment environment, and our toolbox launch Python 3.5 automatically when you start *python* or *ipython* from command line.
- If you have any questions, please ask your tutor or write an email to intromachlearn@techfak.uni-bielefeld.de.

---

1 Do I need all Features

(*5 Points*)

The iris data set is four dimensional[1]; so far we have only used the first two dimensions for their classification in order to arrive at a direct visualization. Compare the result of logistic regression from the sklearn-toolbox for the first two dimensions of the data to a logistic regression for the last two dimensions and to a logistic regression where all dimensions of the data are used. Split the data into a training (2/3 of the data) and test set (1/3 of the data). Repeat for five different train-test-splits and visualize both two dimensional classifiers[2] for at least one of the splits.
[DOC] Document the data used for training (which features, split into train / test set, ...), and the respective classification error.

2 One vs. Rest

(*7 Points*)

In this exercise you will implement the training of a multi-class problem with the One-vs.-Rest concept. Use the `sheet5_task2_skeleton.py` for this exercise, where the data `data5_1.npz` are already loaded. The data set consists of three classes and the labels are 0, 1 and 2. First train three classifiers P_j, $j \in \{0, 1, 2\}$ where class j is set to class 1 and the other two are set to 0 to train a model. You have to change your labels for each training separately.
Then write a function to predict the label of a (new) sample $\vec{x}$. Calculate the probability $P_j(c = 1 \mid \vec{x})$ and choose the class j with the biggest P_j for the winning class.
[DOC] Document the test errors and plot the data with the different decision boundary for each classifier with the function `plot_2d_decisionboundary(model,X, y)` from *utils.py*.

3 Complexity of classifiers

(*3 Points*)

For this exercise use the data stored in `data5_2.npz`. Split the data into training (1/2 of the data) and test set (1/2 of the data). Record the required time[3] for training a k-NN classifier ($k = 5$) and for training a logistic regression.
[DOC] Record the required time for the prediction of the test set as well for both classifiers and explain the results.

[1] The dataset is already loaded in `sheet5_task1_skeleton.py`.
[2] `plot_2d_decisionboundary(model,X, y)`
[3] see `import time` and `time.time()`