University of Pittsburgh
CS2750 Machine Learning Handout 5
Professor Milos Hauskrecht February 17, 2022

Problem assignment 5

Due: Thursday, February 24 2022

In this assignment we continue our investigation of the "Pima" classification dataset by developing and testing two new classification models covered in the class: Support Vector Machines (SVMs) and Multilayer perceptrons (MLP). As in the previous assignment, you are given the complete Pima dataset (*pima.csv*), as well as, *pima_train.csv* and *pima_test.csv* to be used for model training and testing purposes. You have also received the code learning and testing the vanilla logistic regression model from last assignment in file *hw5_LogReg.py*. We will use this model as a baseline model and we will seek to improve its predictive performance with the new models.

Problem 1. Support vector machines

Part a. Write and submit a program *hw5_SVM.py* that will train and evaluate the basic linear support vector machine model. You can build your code using the *hw5_LogReg.py* as a starting point. To implement it please use the scikit-earn svm implementation, in particular svm.SVC classifier model with the linear kernel. Run your code and compare the performance of the linear SVM to the baseline Logistic regression model. Include the results in your report. Also report any differences notices in the performance between the two models.

Part b. One of the advantages of the SVM model is that it can be used to train a non-linear classifier, where the shape of the decision boundary is determined by the kernel function. Please explore and try to use different kernels offered by the svm.SVC and svm.NuSVC classes in scikit-learn library such as the second degree polynomial, or the RBF kernel. Please note optimizations for some of the kernels may take a long time when used with svm.SVC class. In that case, you can try svm.NuSVC that speeds up the learning by restricting the number of support vectors considered. You do not have to submit the code to this part. Just report the models you have tried, results you have been able to achieve and analyze them by comparing their performance the Logistic Regression model and linear SVM.

Problem 2. Multilayer perceptron

Our next step is to develop and study multi-layer perceptron (MLP) models on the pima dataset.

Part a. Write and submit a program *hw5_MLP.py* that will train and evaluate an MLP model with one hidden layer that will consist of 4 units with logistic activation functions. Please use neural networks implemented in the scikit-learn libraries in particular MLPClassifer to write the code. Once again it is recommended that you build your code by starting from *hw5_LogReg.py* and by changing the model. Please run your code and compare the performance of the MLP model to the Logistic regression model and SVM model/models you have implemented and tried in Problem 1. Include the results and comparisons in your report.

Part b. The neural networks module implemented in scikit-learn gives you enough of flexibility to change the structure of the neural network, select different activation functions modeling non-linearities, select different optimization procedures (e.g. sgd, lbfgs, adam), choose the limit on the number of iterations etc. Please experiment with the different architectures using the logistic activations and report any observations. After that try to explore the different parameter settings to observe their effect on the resulting model. Report you findings in the report.