ve5dh5ayu

December 7, 2023

0.1 Tasks Accomplished

1. Displayed the statistical values for each of the attributes, along with visualizations of the distributions for each attribute. Checked if there are any attributes that might require special treatment? If so, what special treatment might they require?
2. Analyzed and discussed the relationships between the data attributes, and between the data attributes and label. This involved computing the Pearson Correlation Coefficient (PCC) and generating scatter plots.
3. Selected 20% of the data for testing and 20% for validation and use the remaining 60% of the data for training. Described how I did that and verified that my test and validation portions of the data are representative of the entire dataset.
4. Trained different classifiers and tweaked the hyperparameters to improve performance. Reported training, validation and testing performance (classification accuracy, precision, recall and F1 score) and discussed the impact of the hyperparameters:
   A. Multinomial Logistic Regression (softmax regression); hyperparameters I explored: C, solver, max number of iterations.
   B. Support vector mach ines (using kernels); hyperparameters I explored: C, kernel, degree of polynomial kernel, gamma.
5. Implemented Random Forest classifier (also analyzed feature importance); hyperparameters I explored: the number of trees, max depth, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node.
6. Combined the classifiers into an ensemble and tried to outperform each individual classifier on the validation set. Once I found a good one, I tried it on the test set. Described and discussed my findings.

0.2 Acknowledgement

I would like to extend my gratitude towards Prof. Zoran Tiganj for his guidance and support throughout the project.

0.3 References

[1] Geron, A. (2019). Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

[2] Machine Learning Notebooks(8,9,10) on GitHub by A. Geron. https://github.com/ageron/handson-ml2 , https://github.com/ageron/handson-ml3

[2] API Reference. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/classes.html

[3] API Reference. (n.d.). Pandas. https://pandas.pydata.org/docs/user_guide/index.html

[4] API Reference. (n.d.). NumPy. https://numpy.org/doc/stable/user/index.html#user