xjl6wyrlc

December 7, 2023

0.1 Tasks Accomplished

1. Summarized the data. Checked how much data is present? What attributes/features are continuous valued? Which attributes are categorical?
2. Displayed the statistical values for each of the attributes, along with visualizations of the distributions for each attribute. Explained noticeable traits for key attributes. Investigated if there are any attributes that might require special treatment? If so, what special treatment might they require?
3. 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.
4. Selected 20% of the data for testing. Described how I did that and verified that the test portion of the data is representative of the entire dataset.
5. Trained a Linear Regression model using the training data with four-fold cross-validation using appropriate evaluation metric. Performed this with a closed-form solution (using the Normal Equation) and with SGD. Performed Ridge, Lasso and Elastic Net regularization – tried a few values of penalty term and described its impact. Explored the impact of other hyperparameters, like batch size and learning rate (no need for grid search). Describe your findings. For SGD, displayed the training and validation loss as a function of training iteration.
6. Repeated the previous step with polynomial regression. Explore if the model overfits/underfits the data using validation loss.
7. Made predictions of the labels on the test data, using the trained model with chosen hyperparameters. Summarized performance using the appropriate evaluation metric. Discussed the results. Included thoughts about what further can be explored to increase performance.

0.2 Acknowledgement

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