CMPSC 445

M4 Assignment - Lab Report

Dhruv Shrivastava

---

3. Summarize the Dataset

3.1 Dimensions of the Dataset

```
Dimensions -> (150, 5)
```

3.2 Peak at the Data

	sepal-length	sepal-width	petal-length	petal-width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa

3.3 Statistical Summary

	sepal-length	sepal-width	petal-length	petal-width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

4. Data Visualization

Box-and-Whisker Plots

The plots are box plots of each of the features of the Iris dataset, indicating their statistical distribution and variation amongst the instances of the Iris flower. The petal dimensions (length and width) appear to have a larger variation than the sepal dimensions. Some outliers have been indicated in the sepal width plot. From the plots, the ranges and the median values of each of the features in the dataset can be interpreted. The Sepal length appears to be relatively symmetric; the Sepal width appears to have several outliers on both ends of the distribution. For the petal length, the distribution appears slightly skewed towards the lower values, and lastly, for the petal width, the distribution appears to have a clustered set of data points.

Histograms

The histograms show varied distributions across the four iris flower measurements. Sepal length shows a roughly normal distribution with a slight right skew, while sepal width is approximately normal with a slight left skew. In contrast, both petal length and petal width display distinct bimodal distributions, strongly suggesting the presence of at least two separate groups/classes within the dataset or having a bimodal representation. The two peaks in the petal measurements likely correspond to the separation between Iris setosa (with smaller petals) and the other two flower types (with larger petals). The sepal measurements, although are less clearly separated, still hint at differences in flower type through their slight skewness.

5. Evaluate Algorithms

5.2. Build Models

After running the 3 models on the training data of the Iris dataset, with default hyperparameters and no cross-validation, the accuracy scores reflect K-Nearest Neighbors (KNN) classifier to be the most accurate with an accuracy percentage of 100%, indicating that all its predictions were correct when evaluating with the test data. Referring to the documentation of scikit-learn, the default value for the hyperparameter k is 5, amongst other parameters. Considering that the training data is of 120 instances, a 100% accuracy score is possibly due to cases of overfitting rather than better performance. The model is maybe memorizing the data rather than learning the inherent pattern. This indicates that it would not perform as well, on unseen or unfamiliar data.

6. Make Predictions

To train the algorithm on the entire training dataset and make predictions on the test data set, the Support Vector Machine (SVM) classifier was chosen. This is because it is a more sophisticated model which consistently performs better than the other 2 models with the cross-validation technique, within the context of the Iris dataset. Looking at the box plots for the algorithmic comparison, even though the Inter-Quartile Ranges (IQR) for all 3 models are very similar, the SVM has a higher median value compared to the other two. So, for the Iris dataset and trailed model choices with their default hyperparameter configurations, the best selected model would be SVM.

SVM Model Accuracy Score (in %) was 96.67%.

7. Model Tuning

- K-Nearest Neighbors
    - After experimenting on the entire training set and with cross-validation, the most optimal k value considering the possibilities for over-fitting, would likely be K = 7.
- Support Vector Machines (SVM)
    - A logarithmic scale was attempted for the tuning exploration to understand the scale and permeability of the SVM on the Iris dataset.
    - Interestingly, when the various permutations with the regularization parameters and the kernel types of the nonlinear SVM classifications were trained on the entire training set, they resulted in a varied set of values. But both nonlinear and linear support vector classifiers (SVCs) resulted in the same (and highest) accuracy across all permutations of the regularization parameters when training with cross validation. This could be due to the size and relative simplicity of the dataset.
    - This led me to infer that both linear and nonlinear SVM training methods are optimal choices when done with cross validation, for the context of the Iris dataset.
    - But, if I had to say the best choices of hyperparameters for SVM on the Iris dataset, I would say it with a nonlinear SVM with C = 10 with the kernel type polynomial.