Rapport

Lab work 2

le 04 décembre 2024,

version 1

Mohamed Toujani,

Fonction

[mohamed.toujani@ecole.ensicaen.fr](mailto:mohamed.toujani@ecole.ensicaen.fr)

Tuteur école : TSAFACK PIUGIE Armand Florent



# Question A

* Categorical variables: For attributes like hair, feathers, eggs, etc. (binary values: "yes" or "no").
* Target variable: type with 7 classes (e.g., mammal, bird, fish, etc.).

A screenshot of a computer

Description automatically generated

# Question B

Using hair on the x-axis and feathers on the y-axis gives a clear view of the data, as it separates mammals (which have hair) from birds (which have feathers) quite well. It’s a simple and effective way to visualize the differences between these groups.

A screenshot of a computer

Description automatically generated

# Question C

The widget shows how features like milk, feathers, and fins separate classes (e.g., mammals, birds, fish). Optimization improves class grouping for clearer patterns.

A screenshot of a computer

Description automatically generated

# Question D

The PCA projection helps separate some classes like mammals and birds. However, classes like amphibians and reptiles or fish and invertebrates are harder to distinguish due to overlapping features.

# Question E

* Normalization: Standardized to mean = 0, variance = 1.
* PCA: Reduced to 10 components for dimensionality reduction.

A screenshot of a computer

Description automatically generated

# Question F

The dataset is divided into 10 equal parts. Each part is used once as the test set while the remaining 9 parts are used for training. This process repeats 10 times, and the results are averaged for a reliable evaluation of the model's performance.

A screenshot of a computer

Description automatically generated

# Question G

The Neural Network achieves the highest classification accuracy (CA = 97.0%), making it the best-performing algorithm in this workflow.

# Question H

Increasing the number of neighbors, changing the metric (e.g., Manhattan or Chebyshev), or adjusting the weight (e.g., distance-based) may improve classification accuracy. Experiment with these settings and observe real-time updates.

A screenshot of a computer

Description automatically generated

# Question I

The classification accuracy typically starts decreasing when the tree depth exceeds a certain point, such as 10 to 15 levels, due to overfitting. Exact results depend on the dataset.

A screenshot of a computer

Description automatically generated

# Question J

The formula is:



Yes, the RBF kernel provides excellent results for this dataset, as indicated by its high classification accuracy (SVM: 93.0%).

A screenshot of a computer

Description automatically generated

# Question K

A decrease in classification accuracy is typically observed when the number of neurons in the hidden layer exceeds 150–200, as it may lead to overfitting.

A screenshot of a computer

Description automatically generated

# Question L

Accuracy is calculated as:

Accuracy = correct predictions / total predictions

For kNN, correct predictions = 3 + 20 + 13 + 8 + 7 + 41 = 92

Total predictions = 100

Accuracy = 92%

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Question M

Misclassified data points are near overlaps between classes, like mammals and invertebrates, showing where KNN struggles with similar features.

A screenshot of a computer

Description automatically generated

# Conclusion

Through this lab, I explored various data visualization and machine learning techniques using Orange. Key takeaways include understanding how features like hair or feathers help separate classes, optimizing algorithms like kNN and neural networks for better accuracy, and recognizing challenges like overlapping classes. This practical work highlighted the strengths and limitations of different methods in analyzing and classifying data effectively.