CSC461 – Assignment3 – Machine Learning

26-11-2023

Muhammad Umar Asif

FA21-BSE-075

Question 1

Provide responses to the following questions about the dataset.

1. How many instances does the dataset contain?

The dataset contains 110 instances.

2. How many input attributes does the dataset contain?

The dataset contains 7 input attributes and one output attribute is "gender".

3. How many possible values does the output attribute have?

The output attribute have 2 value.

4. How many input attributes are categorical?

There are 4 categorical input attributes. 'beard', 'hair_length', 'scarf', 'eye_color'

5. What is the class ratio (male vs female) in the dataset?

Ratio of male vs female:

male 0.56

female 0.44

Question 2

Logistic Regression

1. How many instances are incorrectly classified?

There is 1 false positive and 2 false negative. So model incorrectly classified 3 instances.

2. Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.

Logistic Regression Model gives accuracy of 95 train/test split ratios of 80/20. Pression, recall and f1-score are also very good for this split.

3. Name 2 attributes that you believe are the most “powerful” in the prediction task. Explain why?

Feature importance of **beard** and **shoe_size** are higher than other attributes.

1. beard: The presence or absence of a beard might be a strong indicator of gender. Typically, beards are more associated with males, and this attribute might have a significant impact on predictions.

2. shoe_size: The shoe_size is often associated with gender stereotypes. Small shoe_size is traditionally associated with females, while large shoe_size is associated with males. This attribute could carry valuable information for gender prediction.

4. Try to exclude these 2 attribute(s) from the dataset. Rerun the experiment (using 80/20 train/test split), did you find any change in the results? Explain.

After removing important attributes the accuracy of model decrease from 95% to 86% and precision, recall and f1-score also decreases.

Support Vector Machines

1. How many instances are incorrectly classified?

There are 1 false positves and 2 false negatves. Model incorrectly classified 3 instances.

2. Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.

Accuracy of model accuracy decreases 92 to 82 with train/test split ratio of 80/20.

3. Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?

Feature importance of beard and hair_length are higher than other attributes.

1. beard: The presence or absence of a beard might be a strong indicator of gender. Typically, beards are more associated with males, and this attribute might have a significant impact on predictions.

2. hair_length: The length of hair is often associated with gender stereotypes. Longer hair is traditionally associated with females, while shorter hair is associated with males. This attribute could carry valuable information for gender prediction.

4. Try to exclude these 2 attribute(s) from the dataset. Rerun the experiment (using 80/20 train/test split), did you find any change in the results? Explain.

After excluding hair_lenght and beard from dataset, the model's accuracy and other classification attributes like precision, recall and f1-score also decreased.

Multilayer Perceptron classification

1. How many instances are incorrectly classified?

There are 0 false positves and 2 false negatves. Model incorrectly classified 2 instances.

2. Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.

As we can see that accuracy of model increases with with train/test split ratio of 80/20. Now model incorrectly classified only one instance.

3. Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?

Feature importance of scarf and weight are higher than other attributes.

4. Try to exclude these 2 attribute(s) from the dataset. Rerun the experiment (using 80/20 train/test split), did you find any change in the results? Explain.

After excluding scarf and weight from dataset,there is not much change in the model's accuracy and other classification attributes like precision, recall and f1-score.

Question 3

Apply Random Forest classification algorithm (using Python) on the gender prediction dataset with Monte Carlo cross-validation and Leave P-Out cross-validation. Report F1 scores for both cross-validation strategies.

Monte Carlo cross-validation

Model = RandomForestClassifier(n_estimators=100)

cv = ShuffleSplit(n_splits=5, test_size=0.33, random_state=42)

scoring = f1_scorer

F1 score = 0.98

Leave P-Out cross-validation

model = RandomForestClassifier(n_estimators=100)

lpo = StratifiedKFold(n_splits=5)

F1 score = 0.96

Question 4

Run the ML experiment (using Python) by training the model using Gaussian Naïve Bayes classification algorithm and all the instances from the gender prediction dataset. Evaluate the trained model using the newly added 10 test instances. Report accuracy, precision, and recall scores.

Newly added 10 instances

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
64	165	no	bald	40	no	brown	male
70	115	yes	medium	42	no	black	male
72	168	no	medium	42	no	black	male
65	121	no	medium	39	yes	blue	female
66	150	yes	short	41	no	green	male
63	110	no	long	37	no	black	female
70	181	yes	medium	45	no	blue	male
65	119	no	medium	37	yes	green	female
68	154	yes	short	43	no	brown	male
64	128	no	long	40	yes	brown	female

accuracy = 100%

precision = 1

recall = 1

f1-score = 1