University of Maryland Global Campus

Exercise 1a Suppoprt Vector Machines

Johanson Bombaes

DATA 440 6980 Advanced Machine Learning

Dr. Chalachew Jemberie

Text

AI-generated content may be incorrect.

**Dataset Details:**

After loading the data, I ran print(mice.details[“id”]) to confirm the OpenML ID and locate the dataset details online. MiceProtein contains expression levels of 77 proteins measured in the brains of 72 mice, including 38 control and 34 mice with Down Syndrome (trisomy). Each mouse was measured 15 times, resulting in 1080 samples. target\_variable was set to [“class”]; which, according to the documentation on the dataset, is divided into the following 8:

\* c-CS-s: control mice, stimulated to learn, injected with saline (9 mice)

\* c-CS-m: control mice, stimulated to learn, injected with memantine (10 mice)

\* c-SC-s: control mice, not stimulated to learn, injected with saline (9 mice)

\* c-SC-m: control mice, not stimulated to learn, injected with memantine (10 mice)

\* t-CS-s: trisomy mice, stimulated to learn, injected with saline (7 mice)

\* t-CS-m: trisomy mice, stimulated to learn, injected with memantine (9 mice)

\* t-SC-s: trisomy mice, not stimulated to learn, injected with saline (9 mice)

\* t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (9 mice)

Explain what precision,  recall, f1-score and support are and compare these statistics between the train and test cases.Text

AI-generated content may be incorrect.

* Imports the function to load datasets from OpenMl and sets the dataset as a dataframe, mice.frame
* Creates an array of the column names from the df

Text

AI-generated content may be incorrect.

* Imports necessary libraries/modules and functions

Text

AI-generated content may be incorrect.

* Defines the target variable
* Creates an array of feature columns by dropping the target
* Normalizes predictors by dividing the values with the max to allow scaling from 0-1
* Generates the summary statistics for each column
* Removes rows with missing values

Text

AI-generated content may be incorrect.

* Extracts the predictor (X) and target (y)
* Turns strings/objects into numerical data

Text

AI-generated content may be incorrect.

* Splits the data, 70% for training and 30% for testing
* Imports SVM classifier with one-vs-one decision strategy initialized, then fitting the model on the training data
* Uses the trained model to predict cases for both training and testing sets

Text

AI-generated content may be incorrect.

* Prints the confusion matrix and detailed classification report for training predictions as well as for test predictions to evaluate how the model did.

**ovo Parameter:**

The ovo parameter specifies one-vs-one, meaning that a binary classifier is trained for each pair of classes

* Red vs Blue
* Red vs Green
* Red vs Yellow
* Blue vs Green
* Blue vs Yellow
* Green vs Yellow

ovr stands for one-vs-rest which is when a binary classifier is trained for a class against all others

* Red vs [Blue, Green, Yellow]
* Blue vs [Red, Green, Yellow]
* Green vs [Red, Blue, Yellow]
* Yellow vs [Red, Blue, Green]

When I ran the model using ovr, the accuracy was slightly lower at 94% compared to using ovo (97%). The ovo approach resulted in fewer misclassifications being able to better differentiate between the 8 classes specially because it tends to perform better when the classes are balanced and closely related much like the MiceProtein dataset being used.

OVO:

A picture containing calendar

AI-generated content may be incorrect.

OVR:  
A picture containing calendar

AI-generated content may be incorrect.

**Confusion Matrix:**

The confusion matrix shows how well the model predicted each class by comparing actual vs predicted values. The rows are the actual values while the columns are the predilected values. Looking at the ovr testing confusion matrix above as an example, the diagonal values (10, 27, 20, … 29) indicate correct predictions, while the ones off of the diagonal are misclassifications. Compared to the training set:

A picture containing arrow

AI-generated content may be incorrect.

The training set is nearly perfect while the test has a few more miscalculations. In training cases, precision, recall, and F1-scores are all near perfect (1.0) showing the model fits the training data very well. Support values range from 35-62. In testing cases, the precision, recall, and F1-score drop in some classes. For example in class 0, precision dropped from 0.95 to 0.71. It still performed well overall and the small drop in overall accuracy is expected in real applications.

Precision is the correct positive predictions out of all predicted positives.

Recall is how many actual positives were correctly predicted.

F1-score balances both precision and recall.

Support is the number of true samples in each class.