Importing and Partitioning the Data

1.1 Get and describe the data

	V1	V2	V3	V4
count	5456.000000	5456.000000	5456.000000	5456.000000
mean	1.290307	0.681269	1.881819	1.273689
std	0.626700	0.342594	0.562533	0.821750
min	0.495000	0.340000	0.836000	0.367000
25%	0.826000	0.495000	1.472000	0.788000
50%	1.089500	0.612000	1.753000	1.066500
75%	1.519500	0.753000	2.139000	1.400500
max	5.000000	5.000000	5.000000	5.000000

1.2 Describe the targets

```
1    2205
2    2097
4     826
3     328
dtype: int64
```

1.3 Make an 80/20 Training/Test set split

X_train statistics:

	V1	V2	V3	V4
count	4364.000000	4364.000000	4364.000000	4364.000000
mean	1.287900	0.682126	1.882211	1.272436
std	0.617155	0.343029	0.564946	0.825196
min	0.495000	0.340000	0.836000	0.367000
25%	0.827000	0.494000	1.472000	0.785000
50%	1.090500	0.615000	1.754000	1.065000
75%	1.522000	0.754000	2.136250	1.397000
max	5.000000	5.000000	5.000000	5.000000

X_test statistics:

	V1	V2	V3	V4
count	1092.000000	1092.000000	1092.000000	1092.000000
mean	1.299925	0.677845	1.880249	1.278697
std	0.663680	0.340987	0.553037	0.808190
min	0.496000	0.340000	0.871000	0.367000
25%	0.824750	0.497000	1.475750	0.797750
50%	1.084500	0.599500	1.751500	1.072500
75%	1.498750	0.748000	2.143250	1.410250
max	5.000000	4.408000	3.252000	5.000000

Y_train statistics:

```
1.0    1751
2.0    1670
4.0     675
3.0     268
dtype: int64
```

Y_test statistics:

```
1.0    454
2.0    427
4.0    151
3.0     60
dtype: int64
```

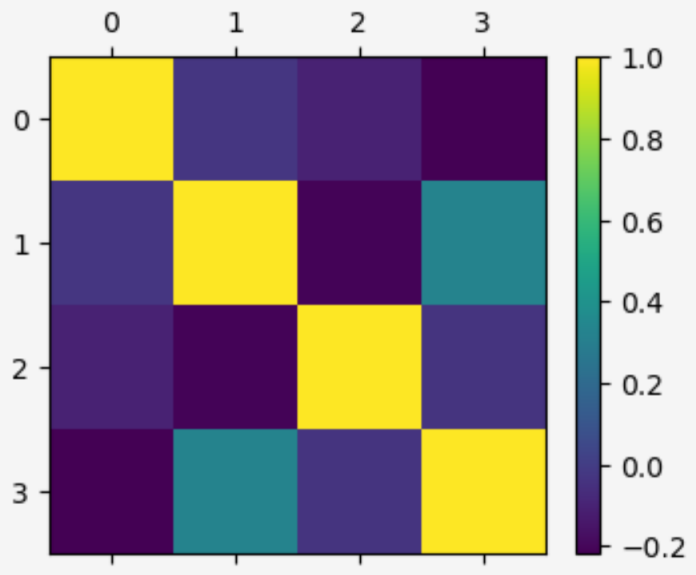
1.4 Binarize the target sets

1938 True commands in the binarized training target set, 487 in the test target set, 2425 total True commands

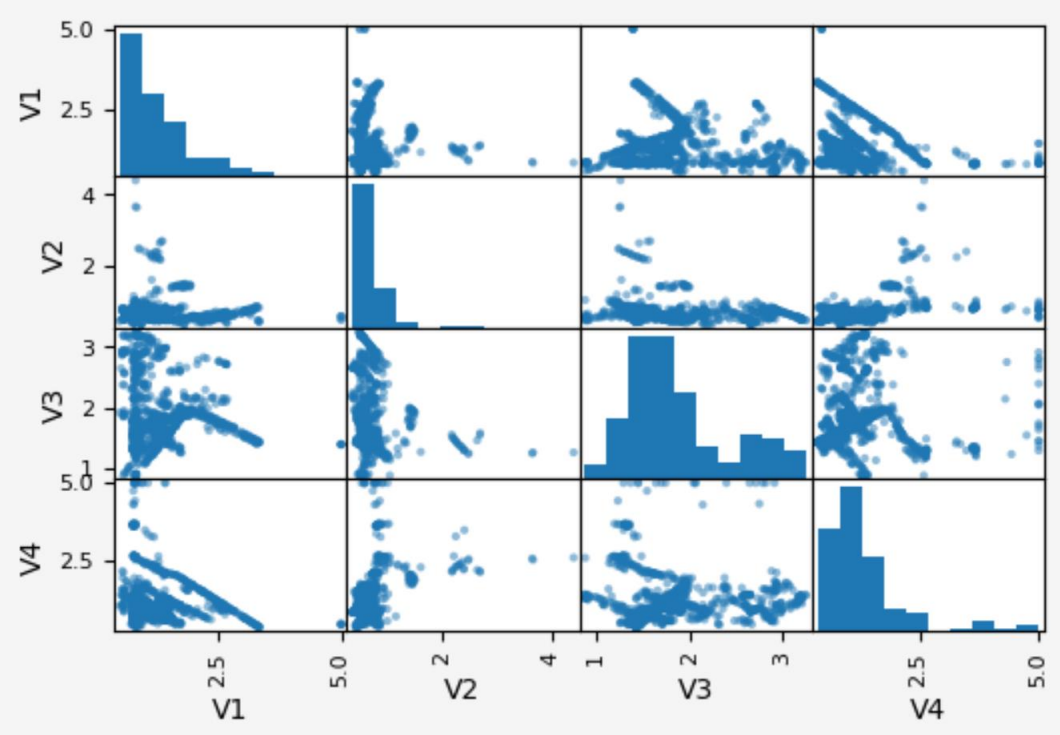
1.5 Training Data correlations

The left and the back feature measurements are the most correlated (correlation coefficient = 0.32587839)

1.6 Plotting the correlations:



1.7 Scatter plots and Histograms

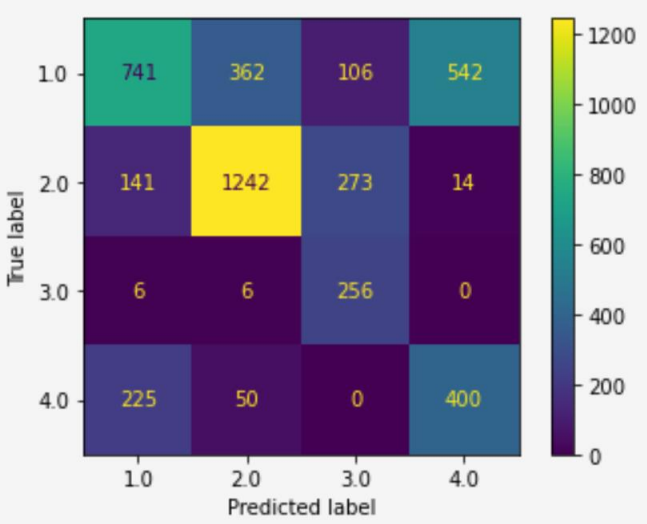


2 Centroid Classification

2.1 Baseline estimator

The accuracy of the baseline centroid estimator on the training data is 0.60472

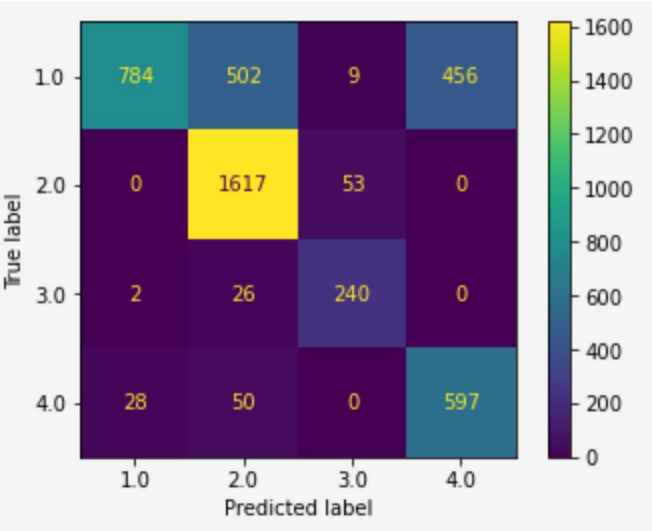
2.2 Baseline confusion



2.3 Scaled Estimator

The accuracy of the Scaled centroid classifier is 0.74197

2.4 Scaled confusion



The following two misclassifications are the most improved:

Misclassifying Slight-Right-Turn (label=2) as Move-Forward (label=1) (141 misclassifications in the baseline estimator, down to 0 in the scaled classifier)

Misclassifying Slight-Right-Turn (label=2) as Slight-Left-Turn (label=4) (14 misclassifications in the baseline estimator, down to 0 in the scaled classifier)

2.5 Scaled Estimator

The accuracy of the Scaled centroid classifier including polynomial features is 0.758249

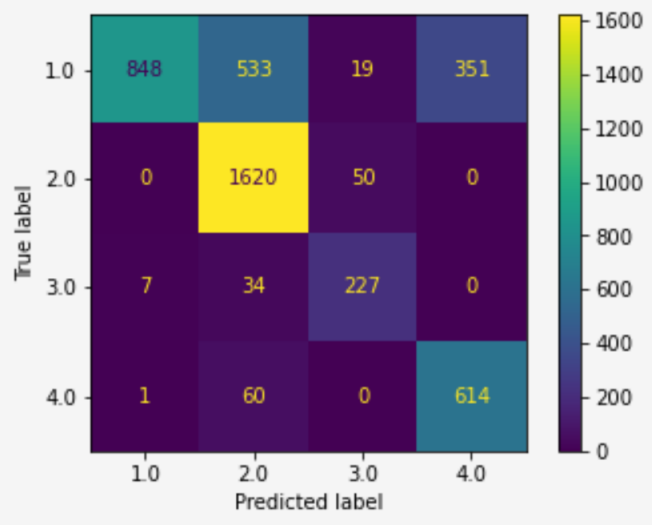
2.6 Best by grid search

The best polynomial degree is 2

2.7 Best score

The score from the best estimator is 0.7582

2.8 Best confusion



The following labels are misclassified more often compared to the baseline estimator:

Move-Forward (label=1) is misclassified as Slight-Right-Turn (label=2) (533 times in the best estimator compared to 362 times in the baseline estimator)

Sharp-Right-Turn (label=3) is misclassified as Move-Forward (label=1) (7 times in the best estimator compared to 6 times in the baseline estimator)

Sharp-Right-Turn (label=3) is misclassified as Slight-Right-Turn (label=2) (34 times in the best estimator compared to 6 times in the baseline estimator)

Slight-Left-Turn (label=4) is misclassified as Slight-Right-Turn (label=2) (60 times in the best estimator compared to 50 times in the baseline estimator)

2.9 Test score and analysis

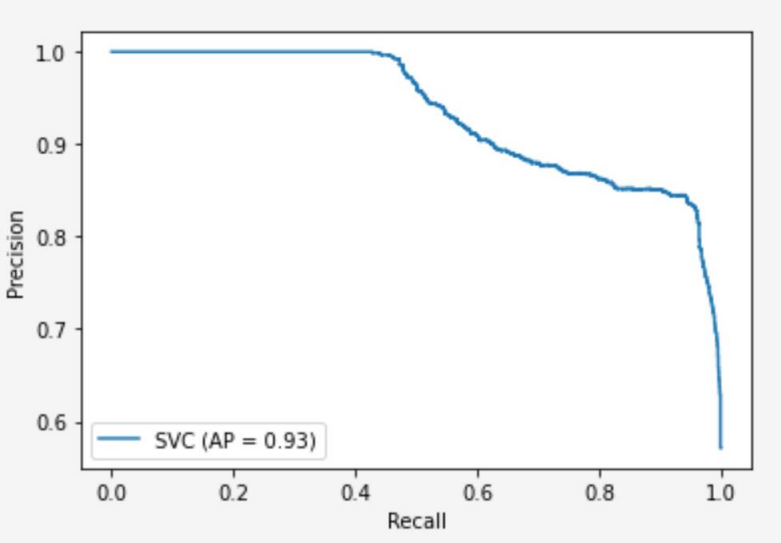
The accuracy of the best estimator on the test data is 0.7509. As the classifier drives the robot in the wrong (or somewhat wrong) direction about 25% of the times, I would not trust it to drive my robot.

3. Support Vector Classification

3.1 Baseline binary estimator

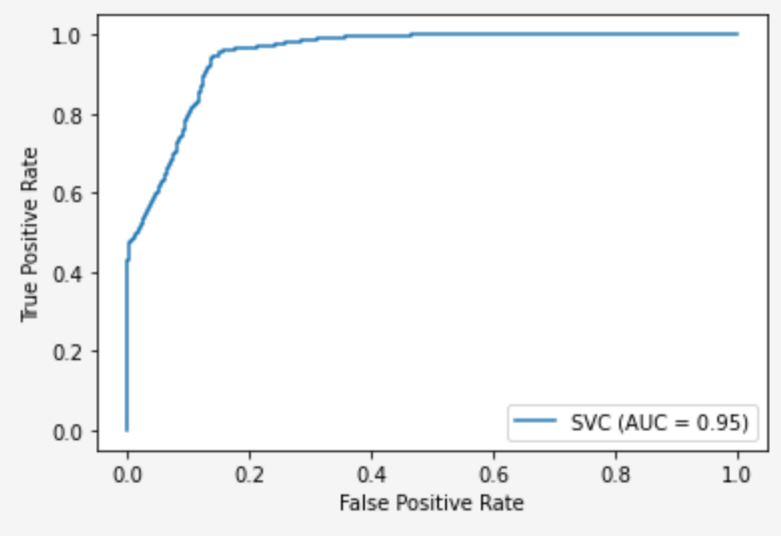
The accuracy of the baseline binary SVC is 0.89436

3.2 Baseline binary precision and recall



The recall for the classifier with precision=0.9 is 0.62538

3.3 Baseline binary ROC curve



The AUC is 0.95 for the baseline estimator

3.4 Best binary by grid search

The accuracy of the best estimator is 0.9855

3.5 Best binary precision and recall

The precision is 1.0 (0.999 when I calculated this on my own by np.mean(precision_array))

3.6 Best binary ROC curve

AUC is 1

3.7 Overfit detection

C=1 and gamma=1000 gives the largest difference between the training and validation accuracies

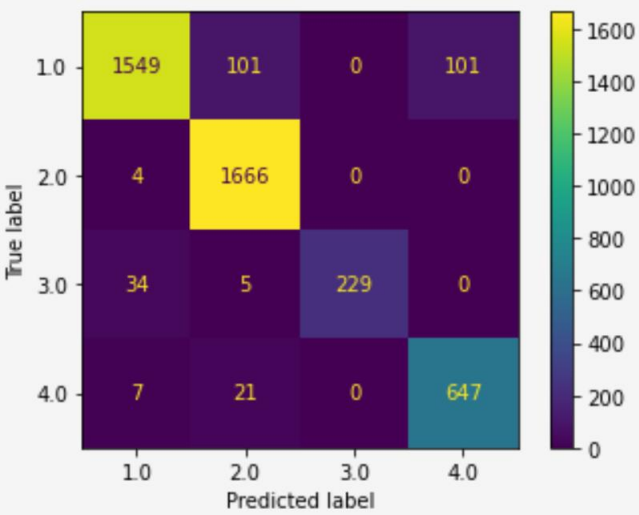
3.8 Binary test score and analysis

The test score is 0.987179

3.9 Baseline multiclass estimator

The accuracy on the entire training dataset is 0.9374

3.10 Baseline multiclass confusion

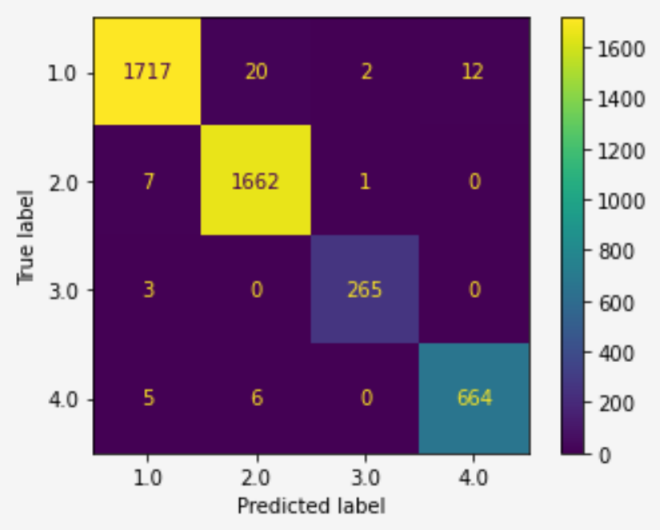


The multiclass SVC misclassifies Move-Forward (label=1) as Slight-Right-Turn (label=2) and Slight-Left-Turn (label=4) the most (101 times each)

3.11 Best multiclass by grid search

The accuracy of the best by grid search estimator is 0.987167

3.12 Best multiclass confusion



The best by grid search multiclass SVC misclassifies Move-Forward (label=1) as Slight-Right-Turn (label=2) 20 times

3.13 Overfit detection

The parameters C=1, gamma=1000 are the most overfit

3.14 Multiclass test score and analysis

The test accuracy is 0.97435

I think that the multiclass SVC performed worse than its binary counterpart because the multiclass SVC does multiple binary classifications, each of which should be at least as 'difficult' as the task given to the binary classifier. While the binary classifier only had to distinguish between right and not-right, the multiclass classifier had to go one step further and classify, say, sharp-right vs slight-right + not right. Any overlap between the input variables corresponding to the sharp right and slight right commands would only make the performance of the multiclass classifier worse.

I would still not trust either classifier to drive my robot, since they both get directions wrong on the training set itself. As a next step, I would probably think about using different input data/augmenting existing data (say with Lidar data) rather than creating a fancier classifier.