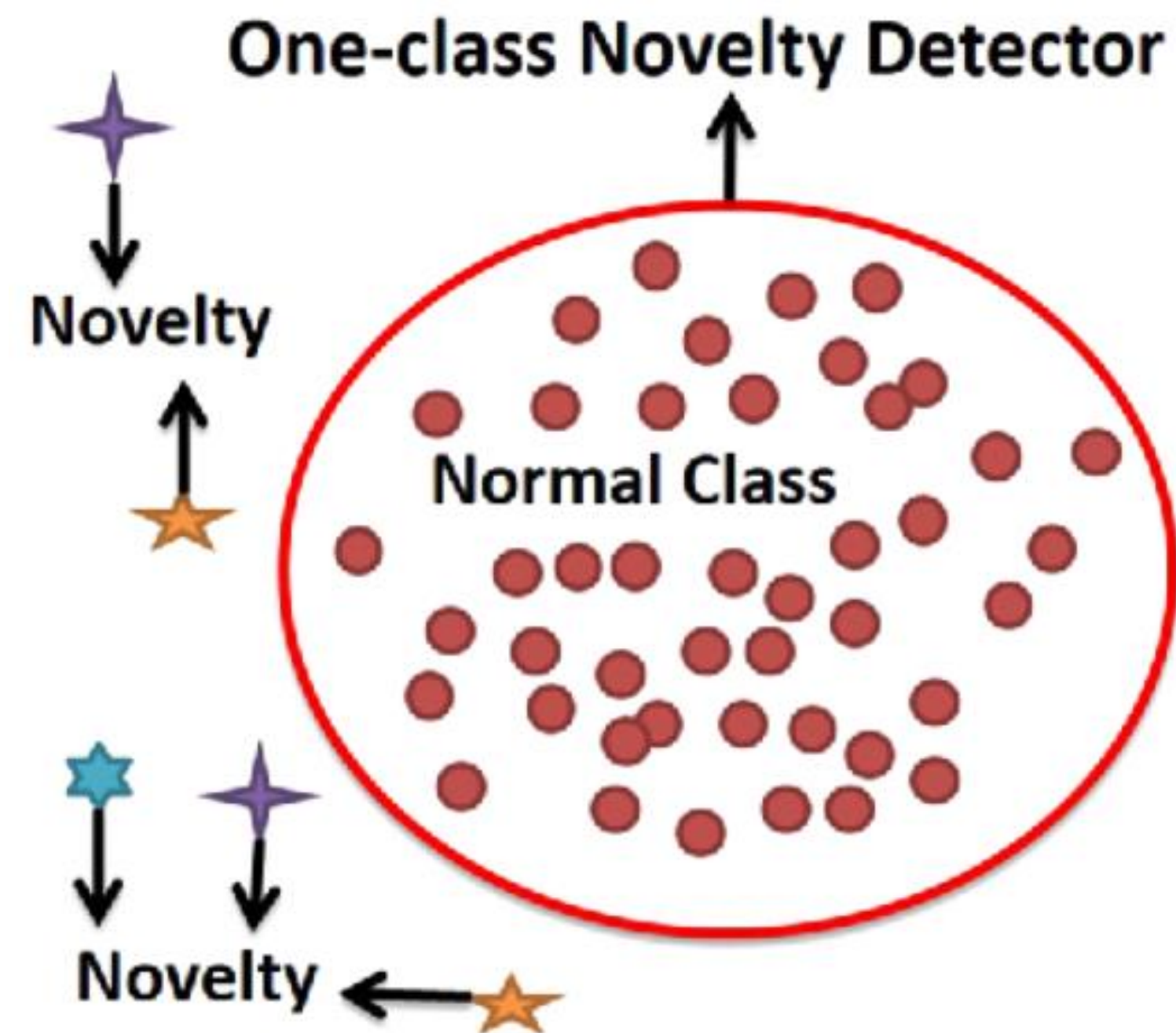
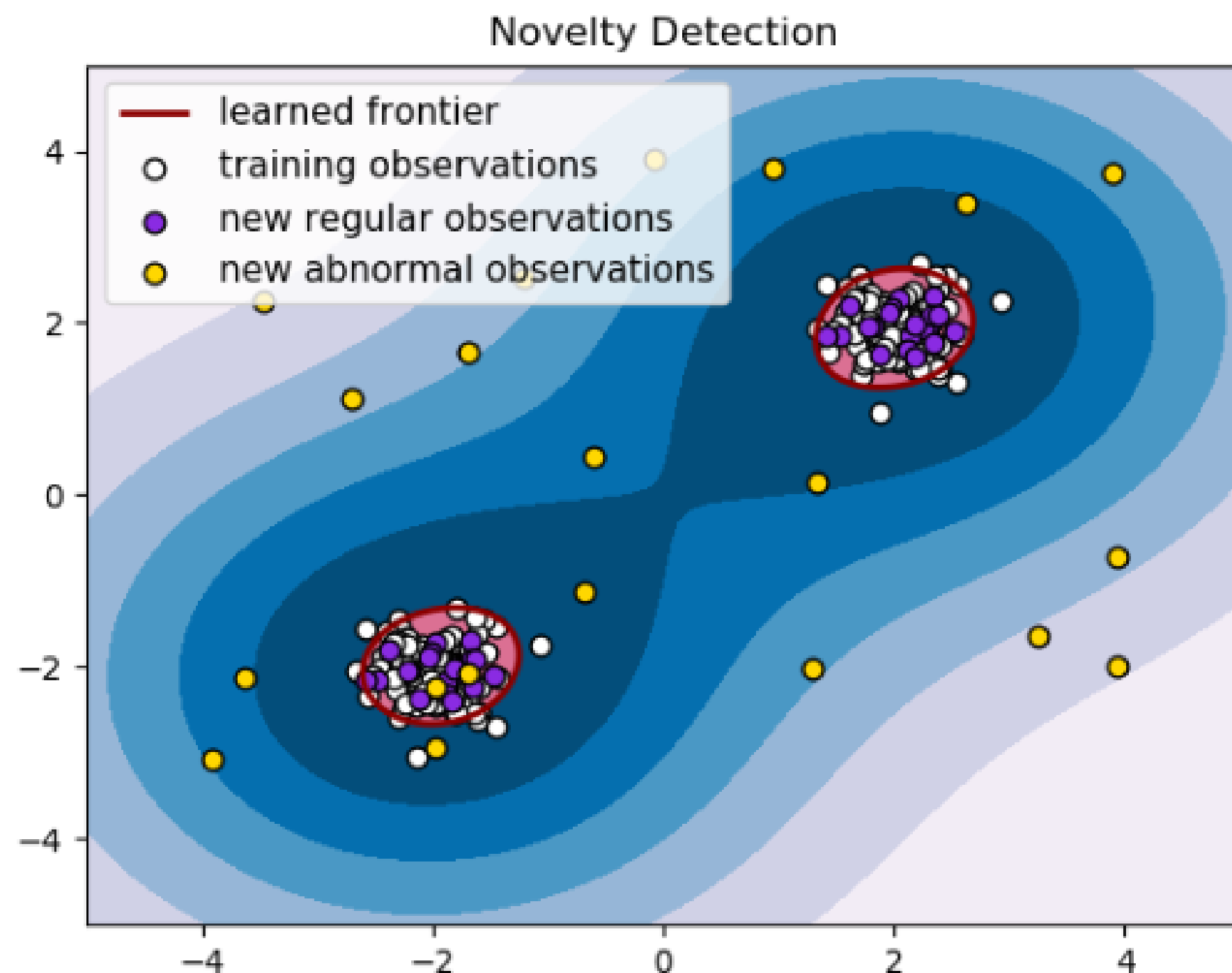


Practice 10

Novelty Detection

Problem

- Classify which data is normal or pathologic(outlier) using Novelty Detection with Spark.
- Use predefined *sklearn.svm.OneClassSVM()*



<https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVM>

<https://www.semanticscholar.org/paper/LGND%3A-a-new-method-for-multi-class-novelty-Tang-Tian/2b18f73596e24b8587eed014f1c9f242e8e5f727/figure/0>

Dataset

➤ Cardiotocography

- The dataset consists of measurements of fetal heart rate and uterine contraction features on **cardiotocograms** classified by expert obstetricians

➤ Explanation

- Data point has **21 features, and label**
- The label value was changed:

From

Label 0: *inliers(normal)* data points

Label 1: *outlier(pathologic)* data points

To

Label 1: *inliers(normal)* data points

Label -1: *outlier(pathologic)* data points

- Because it is convenient to compare the predicted label of Novelty Detection and Real label

You can download dataset and see data description from below links.

Practice 10

1. Calculate accuracy, and f1score of prediction using Novelty Detection algorithm to test data points and get Confusion matrix of the result.

⌘ Note that training dataset is all normal, but test dataset is half normal and half abnormal.

You can download this dataset on I-Campus.

Please use dataset from I-Campus, not from UCI or Stonybrook.

2. Use predefined classes in *sklearn.svm.OneClassSVM*

Parameters for the method

- nu: 0.1, gamma: 0.1, kernel: 'rbf' (Don't change the other parameters)

Practice 10

3. How to train the model using RDD data format

- Before training the model, you need to save data into your memory using **cache()** function.

- For example

```
trRDDs.cache()  
tsRDDs.cache()
```

- In this example, **trRDDs**: training data points & **tsRDDs**: test data points

- Then, you can easily train NoveltyDetection model provided by scikit-learn using **fit()** function

- For example

```
novel = Novelty(nu=nu, kernel="rbf", gamma=gamma)  
novel.fit(trRDDs.collect())
```

- In this example, **nu & gamma**: parameters for Novelty Detection algorithm

Practice 10

4. After training the models, get the accuracy & F1 score for test data points
5. Get confusion matrix of the result
6. You need to use predefined arguments we suggests

- Number of partitions: 30

You can split data when you make it RDDs.

For example, “ ***RDD = sc.parallelize(Data, numPartition)*** ”

Submission

➤ You need to submit result.txt

Write accuracy score of NoveltyDetection result, using *sklearn.metrics.accuracy_score* library

Then, write F1 score of NoveltyDetection result, using *sklearn.metrics.f1_score* library

Also, write confusion matrix of NoveltyDetection result, using *sklearn.metrics.confusion_matrix* library

When you calculate F1 score, you need to use parameter *average = 'macro'*

➤ Result

Novelty Detection Results:

ACC: 0.9345, F1Score: 0.9343

Confusion Matrix

172 20

3 156

```
Novelty Detection Results:
```

```
ACC: 0.9345, F1Score: 0.9343
```

```
Confusion Matrix
```

```
172 20
```

```
3 156
```

Windows

Linux