Python scikit svm “ValueError: X has 62 features per sample; expecting 337”

4

Playing around with Python's scikit SVM Linear Support Vector Classification and I'm running into an error when I attempt to make predictions:

```
ten_percent = len(raw_routes_data) / 10

# Training
training_label = all_labels[ten_percent:]
training_raw_data = raw_routes_data[ten_percent:]
training_data = DictVectorizer().fit_transform(training_raw_data).toarray()


learner = svm.LinearSVC()
learner.fit(training_data, training_label)

# Predicting
testing_label = all_labels[:ten_percent]
testing_raw_data = raw_routes_data[:ten_percent]
testing_data = DictVectorizer().fit_transform(testing_raw_data).toarray()

testing_predictions = learner.predict(testing_data)


m = metrics.classification_report(testing_label, testing_predictions)
```

The raw_data is represented as a Python dictionary with categories of arrival times for various travel options and categories for weather data:

```
{'72_bus': '6.0 to 11.0', 'uber_eta': '2.0 to 3.5', 'tweet_delay': '0',
'c_train': '1.0 to 4.0', 'weather': 'Overcast', '52_bus': '16.0 to 21.0',
'uber_surging': '1.0 to 1.15', 'd_train': '17.6666666667 to 21.8333333333',
'feels_like': '27.6666666667 to 32.5'}
```

When I train and fit the training data I use a Dictionary Vectorizer on 90% of the data and turning it into an array.

The provided testing_labels are represented as:

```
[1,2,3,3,1,2,3, ... ]
```

It's when I attempt to use the LinearSVC to predict that I'm informed:

```
ValueError: X has 27 features per sample; expecting 46
```

What am I missing here? Obviously it is the way I fit and transform the data.

python machine-learning scikit-learn svm

edited Feb 5 '16 at 22:28 asked Feb 5 '16 at 20:43


Farseer
2,478 1 27 44

Lorena Nicole
157 1 4 16

1 Answer

7

The problem is that you creating and fitting different `DictVectorizer` for train and for test.

You should create and fit only one `DictVectorizer` using train data and use `transform` method of this object on your testing data to create feature representation of your test data.

answered Feb 5 '16 at 21:19


Farseer
2,478 1 27 44