FLORIDAPOLY

Machine learning Application Mini-project 2 – Support Vector Machines

Rubric

Report format and quality 30%
Presentation 20 %
Code and documentation 50%

Submission deadline: Nov 8, 2018 11:59 pm
Presentation deadline: Nov 8, 2018, 2018 (class time)

Participants name	*Code Section*	*Report Section*	*Documentation Sec*	*Presentation Sec*

Goals:

- Get the student familiar with the different stages of real machine learning project.
- Expose the student to a commercial API for machine learning
- Practicing concepts such as training set, validation set, test set, parameters, and hyper-parameter, data exploration, feature engineering, error analysis and evaluations.

FLORIDAPOLY

Resources

- Data: https://archive.ics.uci.edu/ml/datasets/iris
  Dataset description

  1. sepal length in cm
  2. sepal width in cm
  3. petal length in cm
  4. petal width in cm
  5. class:
  -- Iris Setosa
  -- Iris Versicolour
  -- Iris Virginica

- Data: Canvas -> Modules -> Resources ->Housing dataset
- Chapter 7 textbook: Support Vector Machines
- https://github.com/ageron/handson-ml (Textbook Jupyter-Notebooks)

Project description.

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Fisher in the 1936 as an example of discriminant analysis. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor), so 150 total samples. Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Project steps

1. Read the data.
2. Use just the features: petal length and petal width, and drop any other feature.
3. Split your data into a training set and a testing set. You can use 80%, 20% for training and testing, something like that:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)
```

4. Train the model for classification by using the traditional fit command. You can use something like:

from sklearn.svm import SVC and then the fit command check chaper 7 text book

5. Train the model using 5 fold cross validation. You will get two models, let's called the two models: traditional, and 5-fold

6. Make predictions using both models (using the testing set)

7. Evaluate your prediction by computing:
    a. Confusion matrix.
    b. Precision
    c. Recall
    d. F1

8. Use Gridsearch to find the best parameters C and gamma, you can use a combination like that:

```
grid = GridSearchCV(SVC(),param_grid,refit=True,verbose=2)
grid.fit(X_train,y_train)
```

9. Predict again (step 6)

10. Evaluate again (step 6)