Recitation

Minfa Wang

mwang5@caltech.edu

Outline

- Tutorial on Kaggle
- Decision Tree with Scikit-Learn

Steps

- Go to https://www.kaggle.com/join/csee155
- Sign up with Caltech Account
- Read the instructions
- Click ‘Make a submission’
- Read the rules and click accept button
- Choose compete as individual/team

Rules

- Team size limit is 3 people
- Submit a maximum of 5 entries per day
- Select up to 2 final submissions for judging

Evaluation

- Categorization Accuracy
- Public vs. private score
- Final grade:
  - 80% Report
  - 20% Model performance

Submission Format

The submission file should contain two columns: Id and Prediction. The Id in the submission file should match the Id of the test file.

The file should contain a header and have the following format:

```
Id,Prediction
1,0
2,0
3,1
4,0
...
9867,1
9868,0
```

Due Date

- Feb 24th (Tuesday) – Scoring file due
- Feb 26th (Thursday) – Report due via Moodle

Outline

- Tutorial on Kaggle
- Decision Tree with Scikit-Learn

Installation

- Python 2.7 or Python 3
  - https://www.python.org/downloads/
- Numpy 1.9
  - http://www.scipy.org/scipylib/download.html
- Scikit-Learn 0.15
  - http://scikit-learn.org/stable/install.html
- Matplotlib 1.4
  - http://matplotlib.org/downloads.html

Read Data

```
from sklearn import tree
import csv
import numpy as np
import matplotlib.pyplot as plt

NUM_TRAININGS = 200
fin_name = 'haberman.data'

with open(fin_name, 'r') as fin:
    data = np.array(list(csv.reader(fin))).astype(int)

X_train = data[:NUM_TRAININGS, :-1]
Y_train = data[:NUM_TRAININGS, -1]
X_test = data[NUM_TRAININGS:, :-1]
Y_test = data[NUM_TRAININGS:, -1]
```

Read Data

```
from sklearn import tree
import csv
import numpy as np
import matplotlib.pyplot as plt

NUM_TRAININGS = 200
fin_name = 'haberman.data'

with open(fin_name, 'r') as fin:
    data = np.array(list(csv.reader(fin))).astype(int)

X_train = data[:NUM_TRAININGS, :-1]
Y_train = data[:NUM_TRAININGS, -1]
X_test = data[NUM_TRAININGS:, :-1]
Y_test = data[NUM_TRAININGS:, -1]
```

Error Function

```
def get_error(G, Y):
    error = 0
    for i in range(len(G)):
        if G[i] != Y[i]:
            error += 1
    return 1.0 * error / len(G)
```

Model Training

```
min_samples_leafs = [i for i in range(1, 25)]
test_errors = []
train_errors = []
for min_samples_leaf in min_samples_leafs:
    # initialize the tree model
    clf = tree.DecisionTreeClassifier(criterion='gini',
        min_samples_leaf=min_samples_leaf)
    # train the model
    clf = clf.fit(X_train, Y_train)

    # make prediction
    G_train = clf.predict(X_train)
    G_test = clf.predict(X_test)

    # compute error
    train_error = get_error(G_train, Y_train)
    train_errors.append(train_error)
    test_error = get_error(G_test, Y_test)
    test_errors.append(test_error)
```

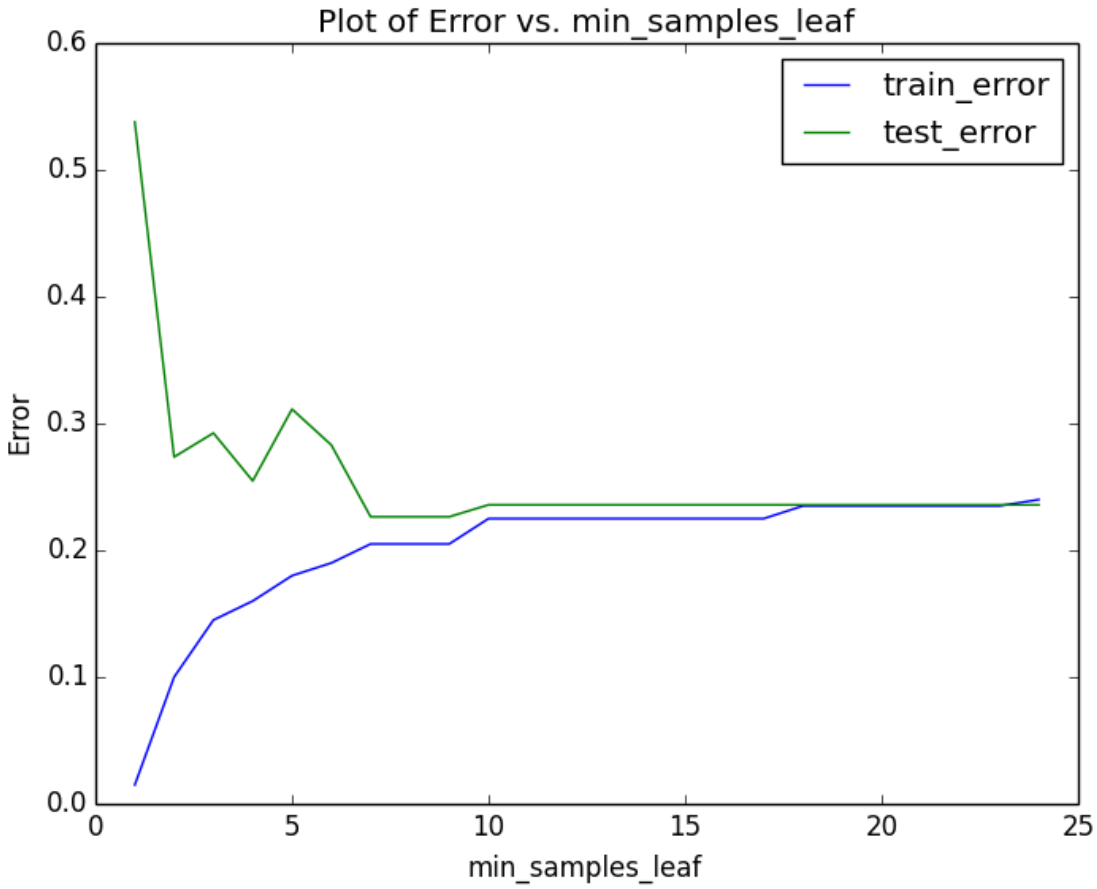
For more details of decision tree model of Scikit-Learn, please go to
http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier

Draw Plot

```
# draw the plot
plt.plot(min_samples_leafs, train_errors)
plt.plot(min_samples_leafs, test_errors)
plt.xlabel('min_samples_leaf')
plt.ylabel('Error')
plt.title('Plot of Error vs. min_samples_leaf')
plt.legend(['train_error', 'test_error'])
# plt.show()
plt.savefig('your_destination.png', bbox_inches='tight')
```

For more details, please refer to official Pyplot tutorial:
http://matplotlib.org/users/pyplot_tutorial.html

Draw Plot


Plot of Error vs. min_samples_leaf
train_error
test_error
Error
min_samples_leaf
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25

Cross Validation

```
# cross validation
K = 5
from sklearn import cross_validation
scores = cross_validation.cross_val_score(clf, X_train, Y_train,
    cv=K, scoring='accuracy')
avg_score = sum(scores) / len(scores)
print('Scores = {}'.format(scores))
print('avg_score = {}'.format(avg_score))
```

```
Minfas-MacBook-Pro:data voiceup$ python clean_decision_tree.py
Scores = [ 0.70731707  0.75        0.7         0.725       0.74358974]
avg_score = 0.725181363352
```

K-fold Cross Validation:
http://scikit-learn.org/stable/modules/cross_validation.html#k-fold
Cross_val_score:
http://scikit-learn.org/stable/modules/generated/sklearn.cross_validation.cross_val_score.html#sklearn.cross_validation.cross_val_score

- Link to the script:

https://www.dropbox.com/s/hzssv4pdmab6h2g/decision_tree_tutorial.zip?dl=0