random_forest_size_training_data

January 30, 2019

In [13]:
```
import numpy
from sklearn.utils import shuffle
from sklearn.model_selection import train_test_split
X = numpy.loadtxt("./data/Train/X_train.txt")
y = numpy.loadtxt("./data/Train/y_train.txt")
```

In [15]:
```
X, y = shuffle(X, y)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.10)
size_training = len(X_train)
```

In [16]:
```
from models import random_forest
from sklearn.metrics import accuracy_score
ks = [.05, .10, .20, .50, 1]
report = []
for k in ks:
    random_forest_model = random_forest.without_penalty(X_train[:int(k*size_training)],
    y_pred = random_forest_model.predict(X_test)
    score = accuracy_score(y_test, y_pred)
    data={
        'k': k,
        'score': score
    }
    report.append(data)
```

In [17]:
```
report
```

Out[17]:
```
[{'k': 0.05, 'score': 0.8558558558558559},
 {'k': 0.1, 'score': 0.888030888030888},
 {'k': 0.2, 'score': 0.9086229086229086},
 {'k': 0.5, 'score': 0.9446589446589446},
 {'k': 1, 'score': 0.9588159588159588}]
```