## In [1]:

```
import daal4py as d4p
   import os,sys,time
   import numpy as np
  from sklearn.metrics import accuracy score
5 from sklearn.metrics import f1_score
  from sklearn.metrics import recall score
7
   from sklearn.metrics import precision score
8 from sklearn.metrics import confusion matrix
   from sklearn.metrics import f1 score
9
   from sklearn.metrics import mean absolute error
10
   from sklearn.metrics import mean_squared_error
11
12
13
   from sklearn import datasets
14
   from sklearn.model selection import train test split
```

## In [2]:

```
#KNN
 1
 2
 3
   # let's try to use pandas' fast csv reader
 4
 5
       import pandas
 6
       read csv = lambda f, c, t=np.float64: pandas.read csv(f, usecols=c, delimite
 7
   except:
 8
       # fall back to numpy loadtxt
 9
       read csv = lambda f, c, t=np.float64: np.loadtxt(f, usecols=c, delimiter=',
10
11
   def main(readcsv=read csv, method='defaultDense'):
12
13
       start = time.clock() # 开始计时
14
15
       # Input data set parameters
       train_file = os.path.join('data', 'k_nearest_neighbors_train.csv')
16
       predict_file = os.path.join('data', 'k_nearest_neighbors_test.csv')
17
18
19
       # Read data. Let's use 5 features per observation
20
       nFeatures = 5
21
       nClasses = 5
22
       train data
                   = readcsv(train file, range(nFeatures))
23
       train labels = readcsv(train file, range(nFeatures, nFeatures+1))
24
       # Create an algorithm object and call compute
25
2.6
       train algo = d4p.kdtree knn classification training(nClasses=nClasses)
27
       # 'weights' is optional argument, let's use equal weights
       # in this case results must be the same as without weights
28
29
       weights = np.ones((train data.shape[0], 1))
30
       train_result = train_algo.compute(train_data, train_labels, weights)
31
32
       # Now let's do some prediction
33
       predict data = readcsv(predict file, range(nFeatures))
34
       predict_labels = readcsv(predict_file, range(nFeatures, nFeatures+1))
35
36
       # Create an algorithm object and call compute
37
       predict algo = d4p.kdtree knn classification prediction()
38
       predict result = predict algo.compute(predict data, train result.model)
39
       # We expect less than 170 mispredicted values
40
41
       assert np.count_nonzero(predict_labels != predict_result.prediction) < 170</pre>
42
43
       end=time.clock() #结束计时
44
       print("time", end-start)
45
46
       return (train_result, predict_result, predict_labels)
47
48
   if __name__ == "__main__":
49
50
       (train_result, predict_result, predict_labels) = main()
51
         print(predict result.prediction.shape)
         print(predict_labels.shape)
52
       print("accuracy_score", accuracy_score(predict_labels, predict_result.predict_
53
54
       print("precision score", precision score(predict labels, predict result.pred
55
       print("recall_score", recall_score(predict_labels, predict_result.prediction
56
       print("f1_score", f1_score(predict_labels, predict_result.prediction, average
57
       print("confusion_matrix\n", confusion_matrix(predict_labels, predict_result.
```

```
er.py:13: DeprecationWarning: time.clock has been deprecated in Pytho
n 3.3 and will be removed from Python 3.8: use time.perf_counter or t
ime.process_time instead
  del sys.path[0]

time 0.2330619999999988
accuracy_score 0.95775
precision score 0.9571618356058407
```

accuracy\_score 0.95775
precision\_score 0.9571618356058407
recall\_score 0.9574422555601799
f1\_score 0.9572627831746068
confusion\_matrix

[[	695	5 (	) (	) (	57]
[	0	823	14	8	0 ]
[	0	8	810	2	0 ]
[	0	7	1	760	0 ]
[	72	0	0	0	743]]

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche r.py:43: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf\_counter or time.process\_time instead

```
In [3]:
```

```
1
   #SVM
 2
 3
   # let's try to use pandas' fast csv reader
 4
5
       import pandas
6
       read csv = lambda f, c, t=np.float64: pandas.read csv(f, usecols=c, delimite
7
   except:
       # fall back to numpy loadtxt
8
9
       read csv = lambda f, c, t=np.float64: np.loadtxt(f, usecols=c, delimiter=',
10
11
   def main(readcsv=read csv, method='defaultDense'):
12
13
       start = time.clock()
14
       # input data file
       infile = "./data/svm two class train dense.csv"
15
16
       testfile = "./data/svm two class test dense.csv"
17
       # Configure a SVM object to use rbf kernel (and adjusting cachesize)
18
19
       kern = d4p.kernel function linear(method=method) # need an object that live
       train algo = d4p.svm training(doShrinking=True, kernel=kern, cacheSize=60000
20
21
22
       # Read data. Let's use features per observation
23
              = readcsv(infile, range(20))
24
       labels = readcsv(infile, range(20,21))
25
       train result = train algo.compute(data, labels)
26
27
       # Now let's do some prediction
       predict algo = d4p.svm prediction(kernel=kern)
2.8
29
       # read test data (with same #features)
30
       pdata = readcsv(testfile, range(20))
31
       plabels = readcsv(testfile, range(20,21))
       # now predict using the model from the training above
32
33
       predict result = predict algo.compute(pdata, train result.model)
34
35
       # Prediction result provides prediction
36
       assert(predict result.prediction.shape == (pdata.shape[0], 1))
37
38
       end=time.clock() #结束计时
       print("time", end-start)
39
40
41
       return (predict_result, plabels)
42
43
   if name == " main ":
44
       (predict result, predict labels) = main()
45
46
       #predict_labels = np.squeeze(predict_labels)
47
         print(predict labels.shape)
         print(predict result.prediction.shape)
48
49
         print(predict result.prediction)
          官方的包写错了,没有进行二值化,这里进行改正
50
51
       for i in range(len(predict_result.prediction)):
52
            if predict_result.prediction[i][0]>=0:
53
               predict result.prediction[i][0]=1
54
           else:
55
               predict_result.prediction[i][0]=-1
56
         print(predict result.prediction)
57
       print("accuracy_score", accuracy_score(predict_labels, predict_result.predict_
58
       print("precision score", precision score(predict labels, predict result.pred
59
       print("recall score", recall score(predict labels, predict result.prediction
```

```
print("f1_score", f1_score(predict_labels, predict_result.prediction, average print("confusion_matrix\n", confusion_matrix(predict_labels, predict_result.
```

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche
r.py:13: DeprecationWarning: time.clock has been deprecated in Python
3.3 and will be removed from Python 3.8: use time.perf\_counter or tim
e.process\_time instead
 del sys.path[0]

time 0.46292100000000014
accuracy\_score 0.971
precision\_score 0.971037103710371
recall\_score 0.9709918839675359
f1\_score 0.9709989559624146
confusion\_matrix
[[977 25]
[ 33 965]]

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche r.py:38: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf\_counter or time.process\_time instead

```
In [4]:
```

```
#朴素贝叶斯
 1
   # let's try to use pandas' fast csv reader
 2
 3
   try:
 4
       import pandas
 5
       read csv = lambda f, c, t=np.float64: pandas.read csv(f, usecols=c, delimite
 6
   except:
 7
       # fall back to numpy loadtxt
       read csv = lambda f, c, t=np.float64: np.loadtxt(f, usecols=c, delimiter=',
 8
 9
10
   def main(readcsv=read csv, method='defaultDense'):
11
       start = time.clock()
12
13
14
       # input data file
15
       infile = "./data/naivebayes train dense.csv"
       testfile = "./data/naivebayes test dense.csv"
16
17
       # Configure a training object (20 classes)
18
19
       talgo = d4p.multinomial_naive_bayes_training(20, method=method)
20
       # Read data. Let's use 20 features per observation
21
22
              = readcsv(infile, range(20))
23
       labels = readcsv(infile, range(20,21))
24
       tresult = talgo.compute(data, labels)
25
26
       # Now let's do some prediction
27
       palgo = d4p.multinomial naive bayes prediction(20, method=method)
       # read test data (with same #features)
28
29
       pdata = readcsv(testfile, range(20))
30
       plabels = readcsv(testfile, range(20,21))
       # now predict using the model from the training above
31
       presult = palgo.compute(pdata, tresult.model)
32
33
34
       # Prediction result provides prediction
35
       assert(presult.prediction.shape == (pdata.shape[0], 1))
36
37
       end=time.clock() #结束计时
       print("time", end-start)
38
39
       return (presult, plabels)
40
41
42
   if name == " main ":
43
44
       (predict_result, predict_labels) = main()
       print("accuracy_score", accuracy_score(predict_labels, predict_result.predict_
45
46
       print("precision_score", precision_score(predict_labels, predict_result.pred
47
       print("recall score", recall score(predict labels, predict result.prediction
       print("f1 score", f1 score(predict labels, predict result.prediction, average
48
49
       print("confusion_matrix\n", confusion_matrix(predict_labels, predict_result.
```

```
/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel_launche
r.py:12: DeprecationWarning: time.clock has been deprecated in Python
3.3 and will be removed from Python 3.8: use time.perf_counter or tim
e.process_time instead
  if sys.path[0] == '':

time 0.1957580000000001
accuracy_score 1.0
precision_score 1.0
```

0 0 0011

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche r.py:37: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf\_counter or time.process\_time instead

In [5]:

```
1
   # adaboost
 2
 3
   # let's try to use pandas' fast csv reader
 4
 5
       import pandas
 6
       read csv = lambda f, c, t=np.float64: pandas.read csv(f, usecols=c, delimite
 7
   except:
       # fall back to numpy loadtxt
 8
 9
       read csv = lambda f, c, t=np.float64: np.loadtxt(f, usecols=c, delimiter=',
10
11
   def main(readcsv=read csv, method='defaultDense'):
12
13
       start = time.clock()
14
15
       infile = "./data/adaboost train.csv"
       testfile = "./data/adaboost test.csv"
16
17
       nClasses = 2
18
19
       # Configure a adaboost training object
       train algo = d4p.adaboost training(nClasses=nClasses)
20
21
       # Read data. Let's have 20 independent, and 1 dependent variable (for each of
22
23
       indep data = readcsv(infile, range(20))
24
       dep data
                  = readcsv(infile, range(20,21))
25
       # Now train/compute, the result provides the model for prediction
       train result = train algo.compute(indep data, dep data)
26
27
       # Now let's do some prediction
28
29
       predict algo = d4p.adaboost prediction(nClasses=nClasses)
30
       # read test data (with same #features)
       pdata = readcsv(testfile, range(20))
31
       # now predict using the model from the training above
32
       predict result = predict algo.compute(pdata, train result.model)
33
34
35
       # The prediction result provides prediction
36
       assert predict result.prediction.shape == (pdata.shape[0], dep data.shape[1]
37
       predict labels = np.loadtxt(testfile, usecols=range(20,21), delimiter=',', r
38
       assert np.allclose(predict result.prediction, predict labels)
39
       end=time.clock() #结束计时
40
       print("time", end-start)
41
42
       return (train result, predict result, predict labels)
43
44
45
46
   if __name__ == "__main__":
47
       (train result, predict result, predict labels) = main()
       print("accuracy score", accuracy score(predict labels, predict result.predict
48
49
       print("precision_score", precision_score(predict_labels, predict_result.pred
50
       print("recall_score", recall_score(predict_labels, predict_result.prediction
51
       print("f1_score", f1_score(predict_labels, predict_result.prediction, average
       print("confusion_matrix\n", confusion_matrix(predict_labels, predict_result.
52
```

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche
r.py:13: DeprecationWarning: time.clock has been deprecated in Python
3.3 and will be removed from Python 3.8: use time.perf\_counter or tim
e.process\_time instead
 del sys.path[0]

```
time 0.2239170000000006
accuracy_score 1.0
precision_score 1.0
recall_score 1.0
f1_score 1.0
confusion_matrix
[[1599 0]
[ 0 401]]
```

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche r.py:40: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf\_counter or time.process\_time instead

## In [6]:

```
# 随机森林RandomFroest
 1
 2
 3
   # let's try to use pandas' fast csv reader
 4
   try:
5
       import pandas
       read csv = lambda f, c, t=np.float64: pandas.read csv(f, usecols=c, delimite
6
7
   except:
8
       # fall back to numpy loadtxt
9
       read csv = lambda f, c, t=np.float64: np.loadtxt(f, usecols=c, delimiter=',
10
   # Get Intel(R) Data Analytics Acceleration Library (Intel(R) DAAL) version
11
   from daal4py import daal link version as dv
12
13
   daal\_version = tuple(map(int, (dv[0:4], dv[4:8])))
14
15
   def main(readcsv=read csv, method='defaultDense'):
16
       start = time.clock()
17
       # input data file
       infile = "./data/df classification train.csv"
18
19
       testfile = "./data/df classification test.csv"
20
       # Configure a training object (5 classes)
21
       train algo = d4p.decision forest classification training(5,
22
23
                                                                  nTrees=10,
24
                                                                  minObservationsInLe
25
                                                                  featuresPerNode=3,
26
                                                                  engine = d4p.engine
27
                                                                  varImportance='MDI
28
                                                                  bootstrap=True,
29
                                                                  resultsToCompute='c
30
       # Read data. Let's use 3 features per observation
31
32
              = readcsv(infile, range(3), t=np.float32)
33
       labels = readcsv(infile, range(3,4), t=np.float32)
       train result = train algo.compute(data, labels)
34
35
       # Traiing result provides (depending on parameters) model, outOfBagError, ol
36
37
       # Now let's do some prediction
38
       predict algo = d4p.decision forest classification prediction(nClasses=5)
39
   #
         if daal version < (2020,1):
             predict algo = d4p.decision forest classification prediction(nClasses=
40
   #
41
         else:
   #
             predict algo = d4p.decision forest classification prediction(nClasses
42
43
                  resultsToEvaluate="computeClassLabels|computeClassProbabilities",
44
       # read test data (with same #features)
       pdata = readcsv(testfile, range(3), t=np.float32)
45
46
       plabels = readcsv(testfile, range(3,4), t=np.float32)
47
       # now predict using the model from the training above
       predict result = predict algo.compute(pdata, train result.model)
48
49
50
       # Prediction result provides prediction
       assert(predict_result.prediction.shape == (pdata.shape[0], 1))
51
52
       end=time.clock() #结束计时
53
54
       print("time", end-start)
55
       return (train result, predict result, plabels)
56
57
   if name == " main ":
58
59
       (train result, predict result, predict labels) = main()
```

```
print("accuracy_score", accuracy_score(predict_labels, predict_result.predict_print("precision_score", precision_score(predict_labels, predict_result.predict_print("recall_score", recall_score(predict_labels, predict_result.prediction_print("fl_score", fl_score(predict_labels, predict_result.prediction, average print("confusion_matrix\n", confusion_matrix(predict_labels, predict_result.)
```

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel launche r.py:16: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf counter or tim e.process time instead app.launch new instance() time 3.0253419999999993 accuracy score 0.972 precision score 0.9724331546028709 recall score 0.9724483374081128 f1 score 0.9723101652373585 confusion matrix [[161 2 0 0 01 2 155 3 0 0 ] 0 8 310 2 0] [ 0 0 5 193 0] [ 0 0 0 6 15311 ſ

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche r.py:53: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf\_counter or time.process time instead

```
In [7]:
```

```
# EM算法
 1
 2
 3
    # let's try to use pandas' fast csv reader
 4
 5
        import pandas
        read csv = lambda f, c=None, t=np.float64: pandas.read csv(f, usecols=c, del
 6
 7
    except:
        # fall back to numpy loadtxt
 8
 9
        read csv = lambda f, c=None, t=np.float64: np.loadtxt(f, usecols=c, delimite
10
11
    def main(readcsv=read csv, method='defaultDense'):
12
13
        start = time.clock()
14
        nComponents = 2
15
        infile = "./data/em gmm.csv"
16
        # We load the data
17
        data = readcsv(infile)
18
19
        # configure a em_gmm init object
        algo1 = d4p.em gmm init(nComponents)
20
21
        # and compute initial model
22
        result1 = algo1.compute(data)
23
24
        # configure a em_gmm object
25
        algo2 = d4p.em gmm(nComponents)
26
27
        # and compute em qmm using initial weights and means
28
        result2 = algo2.compute(data, result1.weights, result1.means, result1.covari
29
30
        end=time.clock() #结束计时
        print("time", end-start)
31
32
        # implicit als prediction result objects provide covariances, goalFunction,
33
        return result2
34
35
36
37
    if name == " main ":
38
        res = main()
39
        # daal库的api不支持预测
        print("Weights:\n", res.weights)
40
        print("Means:\n", res.means)
41
        for c in res.covariances:
42
43
            print("Covariance:\n", c)
44
        print('All looks good!')
time 0.0050040000000004525
Weights:
```

```
[ 0.08127243 -0.10165961 2.57295252]] All looks good!
```

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche
r.py:13: DeprecationWarning: time.clock has been deprecated in Python
3.3 and will be removed from Python 3.8: use time.perf\_counter or tim
e.process\_time instead
 del sys.path[0]

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche r.py:30: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf\_counter or time.process\_time instead

## In [8]:

```
# 随机森林回归
 1
 2
 3
   # let's try to use pandas' fast csv reader
 4
 5
       import pandas
       read csv = lambda f, c, t=np.float64: pandas.read csv(f, usecols=c, delimite
 6
 7
   except:
       # fall back to numpy loadtxt
 8
 9
       read csv = lambda f, c, t=np.float64: np.loadtxt(f, usecols=c, delimiter=',
10
11
   def main(readcsv=read csv, method='defaultDense'):
12
13
       start = time.clock()
14
       infile = "./data/df regression train.csv"
15
       testfile = "./data/df regression test.csv"
16
17
       # Configure a Linear regression training object
       train algo = d4p.decision forest regression training(nTrees=100,
18
19
                                                              varImportance='MDA Raw
20
                                                              bootstrap=True,
21
                                                              engine = d4p.engines mt
22
                                                              resultsToCompute='compu
23
24
       # Read data. Let's have 13 independent, and 1 dependent variables (for each
25
       indep data = readcsv(infile, range(13), t=np.float32)
26
                  = readcsv(infile, range(13,14), t=np.float32)
27
       # Now train/compute, the result provides the model for prediction
28
       train result = train algo.compute(indep data, dep data)
29
       # Traiing result provides (depending on parameters) model, outOfBagError, ol
30
       # Now let's do some prediction
31
32
       predict algo = d4p.decision forest regression prediction()
33
       # read test data (with same #features)
34
       pdata = readcsv(testfile, range(13), t=np.float32)
35
       ptdata = readcsv(testfile, range(13,14), t=np.float32)
36
       # now predict using the model from the training above
37
       predict result = predict algo.compute(pdata, train result.model)
38
39
       # The prediction result provides prediction
       assert predict_result.prediction.shape == (pdata.shape[0], dep_data.shape[1]
40
41
       end=time.clock() #结束计时
42
       print("time", end-start)
43
44
45
       return (train result, predict result, ptdata)
46
47
   if name == " main ":
48
49
       from daal4py import __daal_link_version__ as dv
50
       daal\_version = tuple(map(int, (dv[0:4], dv[4:8])))
51
       if daal version < (2019, 1):
52
           print("Need Intel(R) DAAL 2019.1 or later")
53
       else:
            (train result, predict result, ptdata) = main()
54
55
           print("MAE", mean_absolute_error(ptdata, predict_result.prediction))
56
            print("MSE", mean squared error(ptdata, predict result.prediction))
57
58
```

time 0.2121190000000004 MAE 2.7084110872930176 MSE 17.652745413976753

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche
r.py:13: DeprecationWarning: time.clock has been deprecated in Python
3.3 and will be removed from Python 3.8: use time.perf\_counter or tim
e.process\_time instead
 del sys.path[0]

/opt/anaconda3/envs/daal/lib/python3.7/site-packages/ipykernel\_launche r.py:42: DeprecationWarning: time.clock has been deprecated in Python 3.3 and will be removed from Python 3.8: use time.perf\_counter or time.process\_time instead

n [ ]:	
1	
n [ ]:	
1	