Breast Cancer Wisconsin Dataset HMM Classifier

KATHIA TERAN

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

dataset = pd.read_csv('/Users/kathiateran/Documents/data.csv')
dataset
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smo
0	842302	M	17.99	10.38	122.80	1001.0	
1	842517	M	20.57	17.77	132.90	1326.0	
2	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	
...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	
565	926682	M	20.13	28.25	131.20	1261.0	
566	926954	M	16.60	28.08	108.30	858.1	
567	927241	M	20.60	29.33	140.10	1265.0	
568	92751	B	7.76	24.54	47.92	181.0	

569 rows × 33 columns

- Importing the data and making sure everything is in order.

- Splitting the data and using 70% in each class for training, and 30% to test the classifier performance.

```
X = dataset.iloc[:, 2:32].values
y = dataset.iloc[:, 1].values
```

```
#Encoding categorical data values
from sklearn.preprocessing import LabelEncoder
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)
```

```
# Splitting the dataset:
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.7, random_state=42)
```

```
from sklearn import metrics
from hmmlearn import hmm
from hmmlearn.hmm import GaussianHMM

hmm_clf = GaussianHMM(n_components=3, covariance_type='diag', n_iter=10000).fit(X)

pred_train_hmm = hmm_clf.predict(X_train)
pred_test_hmm = hmm_clf.predict(X_test)


print('\nPrediction accuracy for the training dataset')
print('{:.2%}\n'.format(metrics.accuracy_score(y_train, pred_train_hmm)))


print('Prediction accuracy for the test dataset')
print('{:.2%}\n'.format(metrics.accuracy_score(y_test, pred_test_hmm)))

print('Confusion Matrix')
print(metrics.confusion_matrix(y_test, hmm_clf.predict(X_test)))
```

- Using HMM Gaussian classifier from hmmlearn.
- Please consider the sklearn classifier has been deprecated in Python and hmmlearn has to be installed via terminal separately. Results may be affected due to this exception.

RESULTS

```
Prediction accuracy for the training dataset
82.94%

Prediction accuracy for the test dataset
84.96%

Confusion Matrix
[[234  20   5]
 [  7 105  28]
 [  0   0   0]]
```

	precision	recall	f1-score	support
0	0.97	0.90	0.94	259
1	0.84	0.75	0.79	140
2	0.00	0.00	0.00	0
accuracy			0.85	399
macro avg	0.60	0.55	0.58	399
weighted avg	0.93	0.85	0.89	399

	HMM Classifier in Python
Overall Accuracy	(234+105)/(234+20+5+7+105+28) = 339/399 = **84.96%**
Sensitivity/Recall	**55%**
Specificity	(234+20+7+105)/(399) = 366/399 = **91.72%**

Thanks!