KNN

February 8, 2021

[42]:
```
import pandas as pd
import numpy as np
import warnings
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.model_selection import GridSearchCV

warnings.filterwarnings('ignore')
```

1 Feature Selection Accuracy

[43]:
```
data = pd.read_csv("../Dataset/fe_leaf.csv", delimiter=",")
X_, y_ = data.iloc[:, :-1], data.iloc[:, -1:]
num_features = len(X_.columns)
```

[44]:
```
cv_strat = RepeatedStratifiedKFold(n_splits=5, n_repeats=4, random_state=42)
param = {'n_neighbors': [2, 4, 8, 12, 16, 20], 'p': [1, 2], 'weights':
 ↪['uniform', 'distance']}
```

[24]:
```
gs_KNN = GridSearchCV(estimator = KNeighborsClassifier(), param_grid = param,␣
 ↪cv=cv_strat, scoring='accuracy')
gs_KNN.fit(X_, y_)
print(gs_KNN.best_params_)
print(gs_KNN.best_score_)
```

```
{'n_neighbors': 2, 'p': 1, 'weights': 'distance'}
0.6720588235294118
```

[28]:
```
results_KNN = pd.DataFrame(gs_KNN.cv_results_['params'])
results_KNN['test_score'] = gs_KNN.cv_results_['mean_test_score']
results_KNN['metric'] = results_KNN['p'].replace([1,2,3], ["Manhattan",␣
 ↪"Euclidean", "Minkowski"
])
results_KNN
```

[28]:

	n_neighbors	p	weights	test_score	metric
0	2	1	uniform	0.596324	Manhattan
1	2	1	distance	0.672059	Manhattan
2	2	2	uniform	0.577941	Euclidean
3	2	2	distance	0.637500	Euclidean
4	4	1	uniform	0.584559	Manhattan
5	4	1	distance	0.648529	Manhattan
6	4	2	uniform	0.557353	Euclidean
7	4	2	distance	0.612500	Euclidean
8	8	1	uniform	0.572059	Manhattan
9	8	1	distance	0.616176	Manhattan
10	8	2	uniform	0.536765	Euclidean
11	8	2	distance	0.586765	Euclidean
12	12	1	uniform	0.538971	Manhattan
13	12	1	distance	0.597794	Manhattan
14	12	2	uniform	0.500735	Euclidean
15	12	2	distance	0.563235	Euclidean
16	16	1	uniform	0.509559	Manhattan
17	16	1	distance	0.601471	Manhattan
18	16	2	uniform	0.482353	Euclidean
19	16	2	distance	0.557353	Euclidean
20	20	1	uniform	0.469853	Manhattan
21	20	1	distance	0.595588	Manhattan
22	20	2	uniform	0.438235	Euclidean
23	20	2	distance	0.558824	Euclidean

2 Full Dataset Accuracy

[48]:
```
data_all = pd.read_csv("../Dataset/leaf.csv", delimiter=",")
X_all, y_all = data_all.iloc[:, :-1], data_all.iloc[:, -1:]
num_features_all = len(X_all.columns)
```

[46]:
```
gs_KNN_all = GridSearchCV(estimator = KNeighborsClassifier(), param_grid =␣
 ↪param, cv=cv_strat, scoring='accuracy')
gs_KNN_all.fit(X_all, y_all)
print(gs_KNN_all.best_params_)
print(gs_KNN_all.best_score_)
```

```
{'n_neighbors': 4, 'p': 1, 'weights': 'distance'}
0.6588235294117648
```

[47]:
```
results_KNN_all= pd.DataFrame(gs_KNN_all.cv_results_['params'])
results_KNN_all['test_score'] = gs_KNN_all.cv_results_['mean_test_score']
results_KNN_all
```

[47]:

	n_neighbors	p	weights	test_score
0	2	1	uniform	0.581618
1	2	1	distance	0.647794
2	2	2	uniform	0.532353
3	2	2	distance	0.593382
4	4	1	uniform	0.618382
5	4	1	distance	0.658824
6	4	2	uniform	0.565441
7	4	2	distance	0.605882
8	8	1	uniform	0.586765
9	8	1	distance	0.655147
10	8	2	uniform	0.533088
11	8	2	distance	0.586029
12	12	1	uniform	0.544853
13	12	1	distance	0.641912
14	12	2	uniform	0.491912
15	12	2	distance	0.574265
16	16	1	uniform	0.538235
17	16	1	distance	0.644118
18	16	2	uniform	0.486029
19	16	2	distance	0.585294
20	20	1	uniform	0.536765
21	20	1	distance	0.647059
22	20	2	uniform	0.452941
23	20	2	distance	0.575000

[ ]: