clustering_stability

June 7, 2021

[2]:
```
import pandas as pd
import numpy as np
import seaborn as sns

from matplotlib import pyplot as plt
from sklearn.cluster import KMeans, DBSCAN
from utils import get_data_train, get_columns
```

[3]:
```
df = get_data_train()
chosen_cols = get_columns(df, n_cols=25) + ['activity', 'subject']
```

[4]:
```
X = df[chosen_cols].drop(['activity', 'subject'], axis=1)
```

[44]:
```
from sklearn.metrics import adjusted_mutual_info_score

def cluster_stability(X, model, resamples_n:int, **kwargs):
    stability_scores = []
    # initialize instance of a model with passed key words args
    instance = model(**kwargs)
    # predict base labels
    labels = instance.fit_predict(X)

    for i in range(resamples_n):
        instance = model(**kwargs)
        # permutation of X
        sampled = X.sample(frac=1)

        # get index to retrive original order later
        sampled_index = sampled_index = sampled.index.to_list()

        # get predicted labels for permutated data
        predicted = instance.fit_predict(sampled)

        # retrive original order of observations
        new_labels = pd.concat(
            [pd.Series(sampled_index),
             pd.Series(predicted)],
```

```
            axis='columns').sort_values(by=0).loc[:,1]
        # get mutual info score between base labels and just predicted
        stability_scores.append(adjusted_mutual_info_score(labels, new_labels))

    return np.mean(stability_scores)
```

[46]:
```
print(cluster_stability(X, KMeans, 10, n_clusters=4))
print(cluster_stability(X, KMeans, 10, n_clusters=6))
print(cluster_stability(X, DBSCAN, 10, eps=0.5, min_samples=5))
print(cluster_stability(X, DBSCAN, 10, eps=0.5, min_samples=5))
```

```
1.0
0.9970639122443803
1.0
1.0
```

Wygląda na to, że klastrujemy stabilnie

[47]:
```
scores = []
for i in range(2,11):
    scores.append(cluster_stability(X, KMeans, 10, n_clusters=i))
```

[61]:
```
import plotly.express as px

fig = px.line(x=[x for x in range(2,11)], y=scores,
              labels={'x':'n_clusters', 'y':'Adjusted Mutual Info'},
              title='Clustering Stability for KMeans (mean of 10 resamplings)')
fig.show()
```