

K-means project

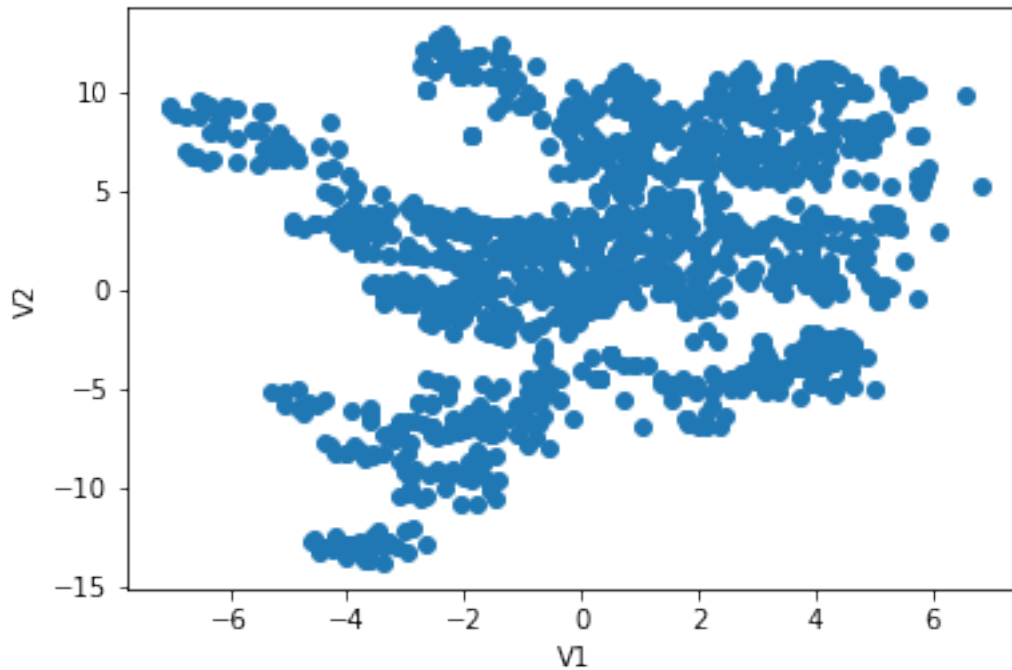
April 15, 2021

```
In [3]: import numpy as np
import pandas as pd
data=pd.read_csv('Banknote-authentication-dataset-.csv')
data=np.array(data)
mean=np.mean(data,0)
std_dev=np.std(data,0)
print("mean=",mean)
print("standard deviation=",std_dev)
```

```
mean= [0.43373526 1.92235312]
standard deviation= [2.84172641 5.86690749]
```

```
In [4]: import pandas as pd
import matplotlib.pyplot as plt
data=pd.read_csv('Banknote-authentication-dataset-.csv')
V1=data['V1']
V2=data['V2']
plt.xlabel('V1')
plt.ylabel('V2')
plt.scatter(V1,V2)
```

```
Out[4]: <matplotlib.collections.PathCollection at 0x7f4d5c4fc438>
```



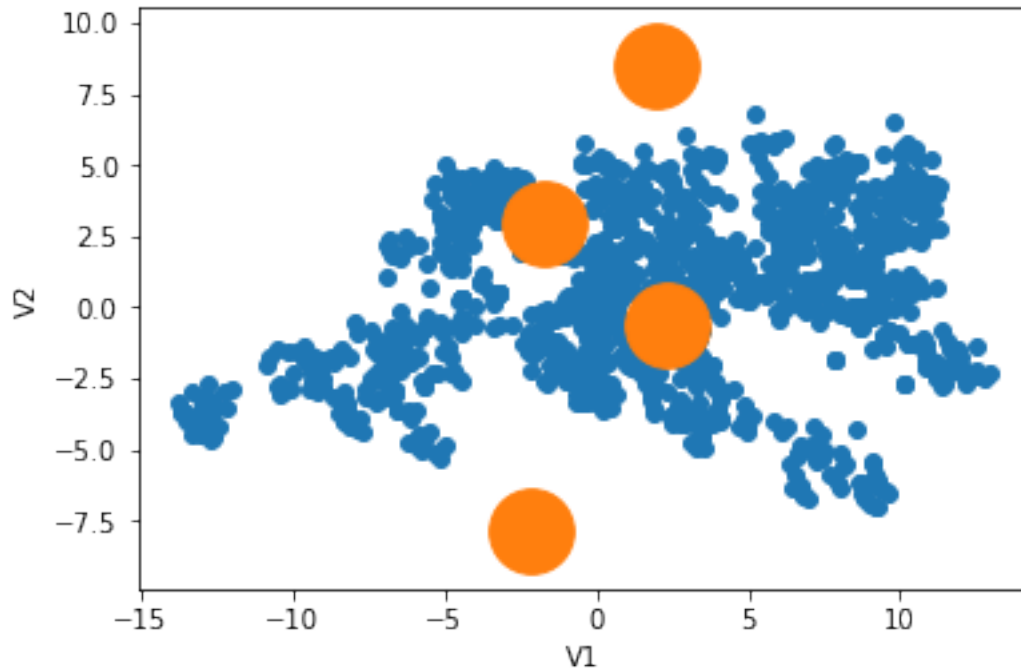
```
In [9]: from sklearn.cluster import KMeans
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

data=pd.read_csv('Banknote-authentication-dataset-.csv')
V1=data['V1']
V2=data['V2']

V12=np.column_stack((V1,V2))#converting V1 and V2 into a single 2-D list
kmean_res=KMeans(n_clusters=4).fit(V12)
plt.xlabel('V1')
plt.ylabel('V2')
plt.scatter(V2,V1)
clusters=kmean_res.cluster_centers_# gives centers of all the clusters(in this case we

plt.scatter(clusters[:,0],clusters[:,1],s=1000)

Out[9]: <matplotlib.collections.PathCollection at 0x7f222f4ce470>
```



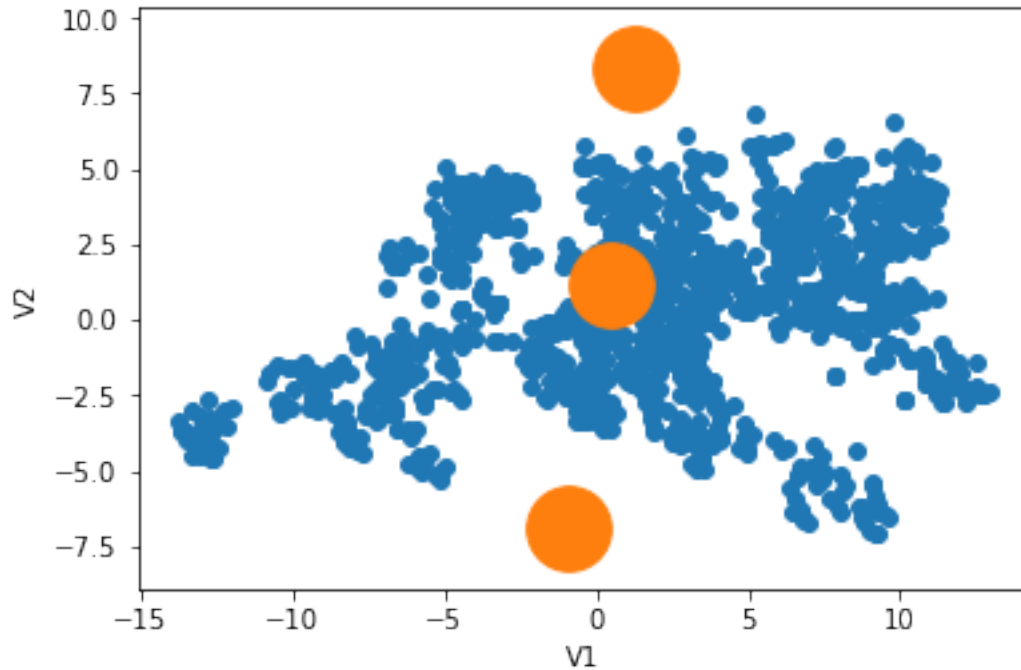
```
In [10]: from sklearn.cluster import KMeans
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

data=pd.read_csv('Banknote-authentication-dataset-.csv')
V1=data['V1']
V2=data['V2']

V12=np.column_stack((V1,V2))#converting V1 and V2 into a single 2-D list
kmean_res=KMeans(n_clusters=3).fit(V12)
plt.xlabel('V1')
plt.ylabel('V2')
plt.scatter(V2,V1)
clusters=kmean_res.cluster_centers_# gives centers of all the clusters(in this case w

plt.scatter(clusters[:,0],clusters[:,1],s=1000)

Out[10]: <matplotlib.collections.PathCollection at 0x7f222f42e390>
```



```
In [7]: from sklearn.cluster import KMeans
        from sklearn.preprocessing import MinMaxScaler
        import matplotlib.pyplot as plt
        %matplotlib inline
        import pandas as pd

        data=pd.read_csv('Banknote-authentication-dataset-.csv')

        km=KMeans(n_clusters=2)
        Y=km.fit_predict(data[['V1','V2']])
        km.cluster_centers_
        data['cluster']=Y

        data1=data[data.cluster==0]
        data2=data[data.cluster==1]

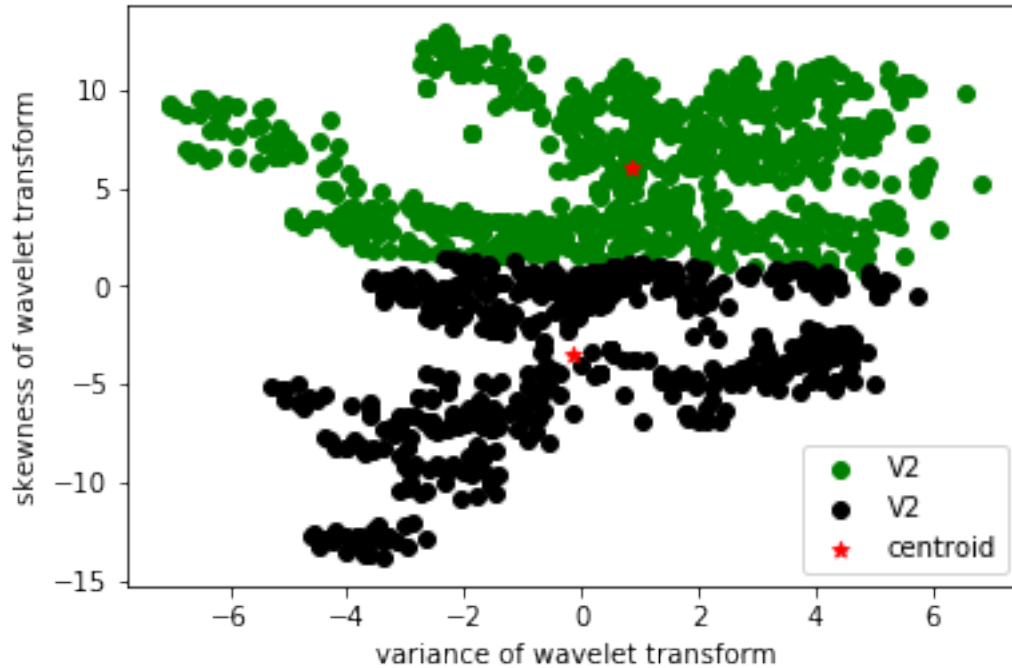
        plt.scatter(data1['V1'],data1['V2'],color='green')
        plt.scatter(data2['V1'],data2['V2'],color='black')

        plt.scatter(km.cluster_centers_[0],km.cluster_centers_[1],marker='*',label='centro')

        plt.xlabel('variance of wavelet transform')
```

```
plt.ylabel('skewness of wavelet transform')  
plt.legend()
```

Out[7]: <matplotlib.legend.Legend at 0x7eff1c3fc6a0>



- 1 Yes I found it is suitable for the k-Means clustering
- 2 Yes so by running k-means for several times I found it is stable for the given data set.