K-Means Clustering Algorithm Example-1

Scikit-learn library နှင့် random data များဒေတာအဖြစ် အသုံးပြု၍ K-means clustering ဥပမာ တစ်ခုကို ရှင်းပြထားသည်။

Step 1: Import libraries

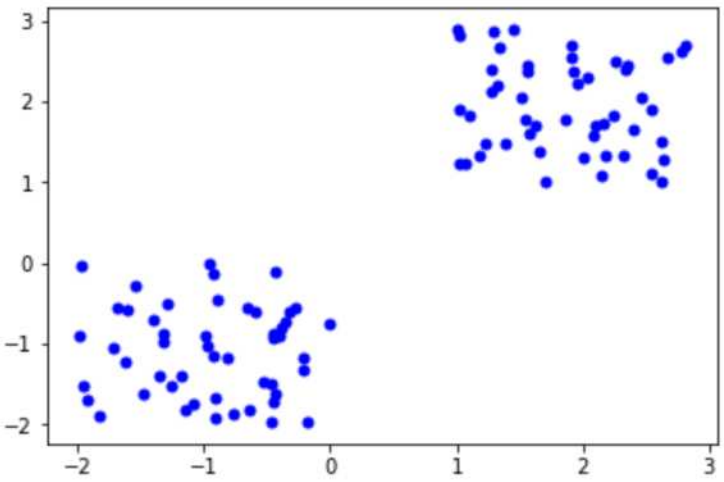
pandas , numpy , matplotlib နှင့် sklearn တို့ကို လုပ်သည်။
Pandas for reading and writing spreadsheets
Numpy for carrying out efficient computations
Matplotlib for visualization of data

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
%matplotlib inline
```

Step 2: Generate random data

Two-dimensional random data များရရှိရန် generate လုပ်သည်။ ထိုဒေတာများကို သုံး၍ ပုံဆွဲသည်။wo-dimensional random data များရရှိရန် generate လုပ်သည်။

In [2]:
```
X= -2 * np.random.rand(100,2)
X1 = 1 + 2 * np.random.rand(50,2)
X [50:100, :] = X1
plt.scatter(X[ : , 0], X[ :, 1] , s = 25 , c = 'b')
plt.show()
```



sklearn.cluster မှ KMeans ဖြင့် fit လုပ်သည်။ n_clusters=2 သည် ဒေတာများကို (၂)စု ခွဲခြားမည်ဟုဆိုလိုသည်။ ဂရပ်ကို ကြည့်ရုံဖြင့် အစု ဘယ်နှစ်ခု ခွဲရမည်ကို သိနိုင်သည်။

In [3]:
```
from sklearn.cluster import KMeans
Kmean = KMeans(n_clusters=2)
Kmean.fit(X)
```

Out[3]:
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=2, n_init=10, n_jobs=None, precompute_distances='auto',
    random_state=None, tol=0.0001, verbose=0)
```

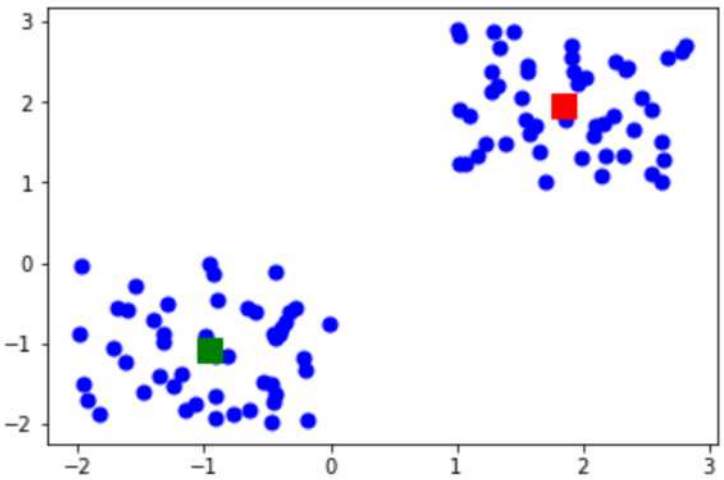
max_iter=300 သည် itration အများဆုံး အကြိမ် (၃၀၀) အထိ လုပ်မည်ဟု ဆိုလိုသည်။
centroids တည်ရှိရာ နေရာကိုဖော် ပြရန်အတွက် တန်ဖိုးများကို ဖတ်ယူသည်။

In [4]:
```
centroids = Kmean.cluster_centers_
centroids
```

Out[4]:
```
array([[-0.95456533, -1.08966598],
       [ 1.85253191,  1.9484689 ]])
```

Centroid တည်ရှိရာ နေရာကို ရရှိပြီးနောက် ဒေတာများနှင့် အတူ centroid ကို ပါ ထည့်သွင်း၍ ဂရပ် ဆွဲသည်။

In [10]:
```
plt.scatter(X[ : , 0], X[ : , 1], s =50, c= 'b')
plt.scatter(-0.95456533, -1.08966598, s=150, c='g', marker='s')
plt.scatter(1.85253191,  1.9484689, s=150, c='r', marker='s')
plt.show()
```



Clustering လုပ်ပြီးနောက် ဒေတာများ အားလုံးတွင် label ရရှိသွားပြီဖြစ်သည်။ တစ်နည်းအားဖြင့် မည်သည့်ဒေတာသည် မည်သည့် cluster တွင် ပါဝင်သည်ကို သတ်မှတ်ပြီးဖြစ်သည်။ အနီရောင် centroid ရှိသည့် အစုကို အနီရောင် အစု အဖြစ် အစိမ်းရောင် centroid ရှိသည့် အစုကို အစိမ်းရောင် အစု အဖြစ် သတ်မှတ်သည်။

In [6]:
```
Kmean.labels_
```

Out[6]:
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
```

လုပ်ပြီးသည့် clustering မှန်မမှန် စစ်ဆေးရန် test point တစ်ခုကို ဖန်တီးသည်။

In [7]:
```
sample_test = np.array([-2.0,-1.0])
sample_test
```

Out[7]:
```
array([-2., -1.])
```

test point တစ်ခုကို reshape() လုပ်သည်။

In [8]:
```
second_test=sample_test.reshape(1, -1)
second_test
```

Out[8]:
```
array([[-2., -1.]])
```

Kmean.predict() ဖြင့် test point သည် မည်သည့် ထဲတွင် ပါဝင်သည်ကို စစ်ဆေးသည်။ 0 ဟု အဖြေထွက်လျှင် အနီရောင် အစုထဲတွင် ပါဝင်သည်။ 1 ဟု အဖြေထွက်လျှင် အစိမ်းရောင် အစုထဲတွင် ပါဝင်သည်။

In [9]:
```
Kmean.predict(second_test)
```

Out[9]:
```
array([0])
```

test point သည် အနီရောင် အစုထဲတွင် ပါဝင်သည်။

Ref: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1 (https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1)

Ref: Understanding K-means Clustering in Machine Learning by Dr. Michael J. Garbade, Sep 13, 2018