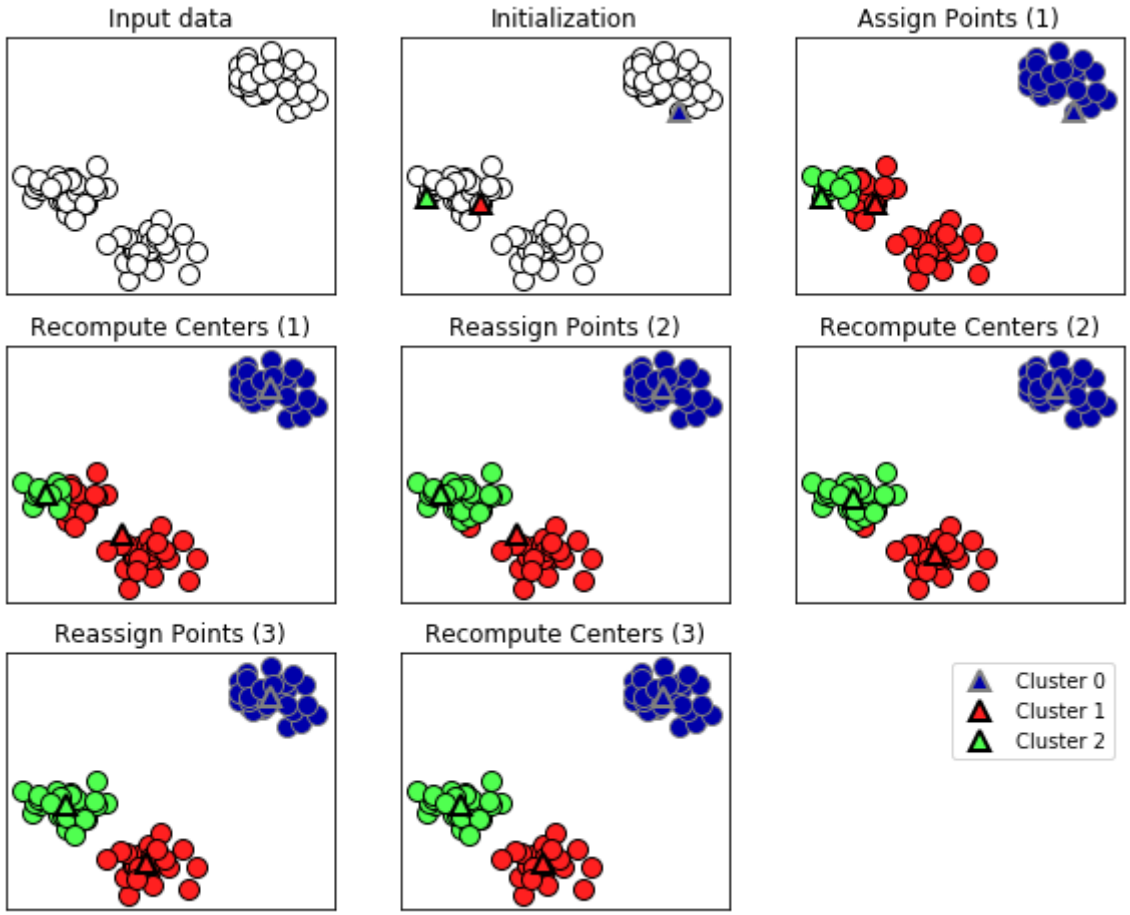
Clustering - Kmeans

01. 군집의 목적

한 클러스터 안의 데이터 포인트끼리는 매우 비슷. 다른 클러스터의 데이터 포인터와는 구분되도록 데이터를 나눈다.

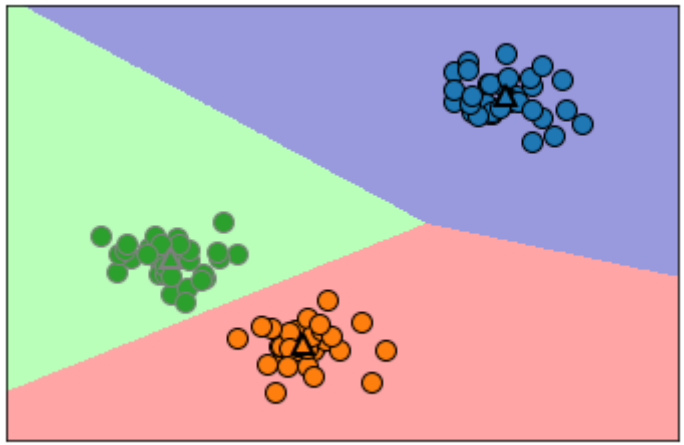
In [3]:
```
import mglearn
%matplotlib inline
```

In [4]:
```
mglearn.plots.plot_kmeans_algorithm()
```



02. k-평균 알고리즘으로 찾은 클러스터 중심과 클러스터 경계

In [6]:
```
mglearn.plots.plot_kmeans_boundaries()
```



03. k-means 알고리즘 적용

In [12]:
```
from sklearn.datasets import make_moons
from sklearn.cluster import KMeans

X, y = make_moons(n_samples=200, noise=0.05, random_state=0)  # 데이터 만들기(2차원 데이터)

# 두 개의 클러스터로 데이터에 KMeans 알고리즘 적용
kmeans = KMeans(n_clusters=2)
kmeans.fit(X)
y_pred = kmeans.predict(X)
y_pred
```

Out[12]:
```
array([1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0,
       1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0,
       1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0,
       0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1,
       1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,
       1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0,
       1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0,
       1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1,
       1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1])
```

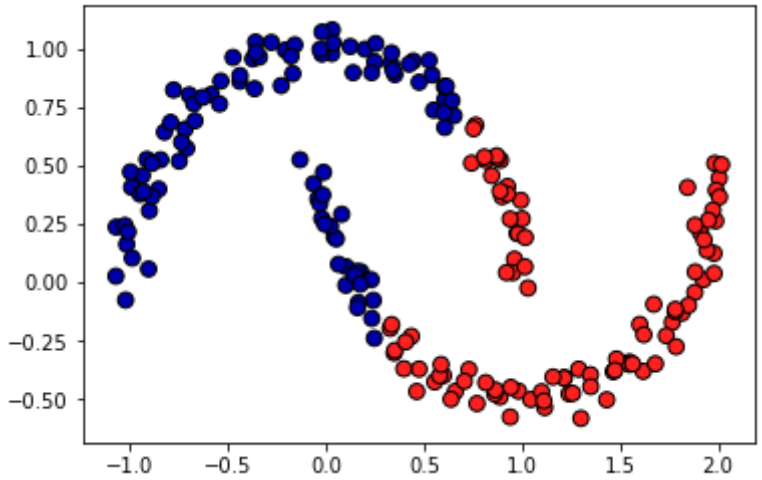
In [21]:
```
print(X.shape, y.shape)
print(X[1:5])
print(X[:, 0][1:5])   # X의 첫번째 열에서 1행부터~4행까지
print(X[:, 1][1:5])   # X의 두번째 열에서 1행부터~4행까지
```

```
(200, 2) (200,)
[[ 1.61859642 -0.37982927]
 [-0.02126953  0.27372826]
 [-1.02181041 -0.07543984]
 [ 1.76654633 -0.17069874]]
[ 1.61859642 -0.02126953 -1.02181041  1.76654633]
[-0.37982927  0.27372826 -0.07543984 -0.17069874]
```

In [23]:
```
import matplotlib.pyplot as plt

# 클러스터 할당과 클러스터 중심을 표시한다.
col1 = X[: ,0]
col2 = X[: ,1]
plt.scatter(col1, col2, c=y_pred, cmap=mglearn.cm2, s=60, edgecolors='k')
```

Out[23]:
```
<matplotlib.collections.PathCollection at 0x1d6d2bb6f60>
```



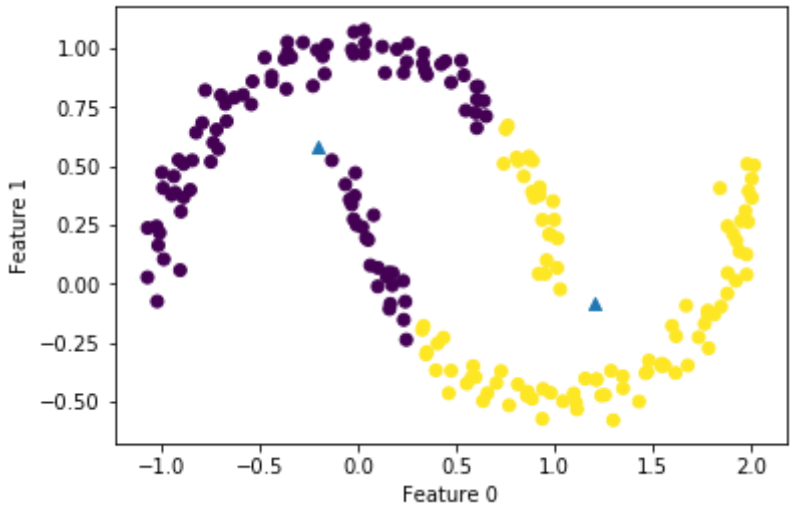
In [28]:
```
# 클러스터의 중심
print(kmeans.cluster_centers_)
print(kmeans.cluster_centers_[ : , 0])  # x 좌표
print(kmeans.cluster_centers_[ : , 1])  # y 좌표
```

```
[[-0.2003285   0.58035606]
 [ 1.20736718 -0.0825517 ]]
[-0.2003285   1.20736718]
[ 0.58035606 -0.0825517 ]
```

In [38]:
```
## 그래프 위에 클러스터의 중심 표시
centerX = kmeans.cluster_centers_[ : , 0]
centerY = kmeans.cluster_centers_[ : , 1]

plt.scatter(col1, col2, c=y_pred)
plt.scatter(centerX, centerY, marker="^")
plt.xlabel("Feature 0")
plt.ylabel("Feature 1")
```

Out[38]: <matplotlib.text.Text at 0x1d6d2330a20>



In [ ]: