CH 3.5.2 병합 군집(agglomerative clustering)

병합 군집 간단하게 알아보기

- (1) 각 포인트를 하나의 클러스터로 지정한다.
- (2) 어떤 종료 조건(클러스터 개수)을 만족할 때가지 **비슷한 두 클러스터를 합쳐**나간다.
- (3) 하나의 포인트에서 시작하여 마지막 클러스터까지 이동

병합군집

- (1) 병합 군집은 계층적 군집(hierarchical clustering)을 만든다.
- (2) 병합 군집은 predict 메서드가 없다.
- (3) 훈련 세트 모델을 만들고 **클러스터 소속 정보를 얻기 위해 fit_predict 메서드 사용**한다.
- (4) sklearn.cluster.AgglomerativeClustering 클래스 사용

scikit-learn 옵션

- (3) linkage 옵션에서 가장 비슷한 클러스터를 측정하는 방법을 지정한다.
  - ward : 모든 클러스터 내의 분산을 가장 작게 증가시키는 두 클러스터를 합친다. (대부분 이를 사용)
  - average : 클러스터 포인트 사이의 평균 거리가 가장 짧은 두 클러스터를 합친다.
  - complete : 클러스터 포인트 사이의 최대 거리가 가장 짧은 두 클러스터를 합친다.

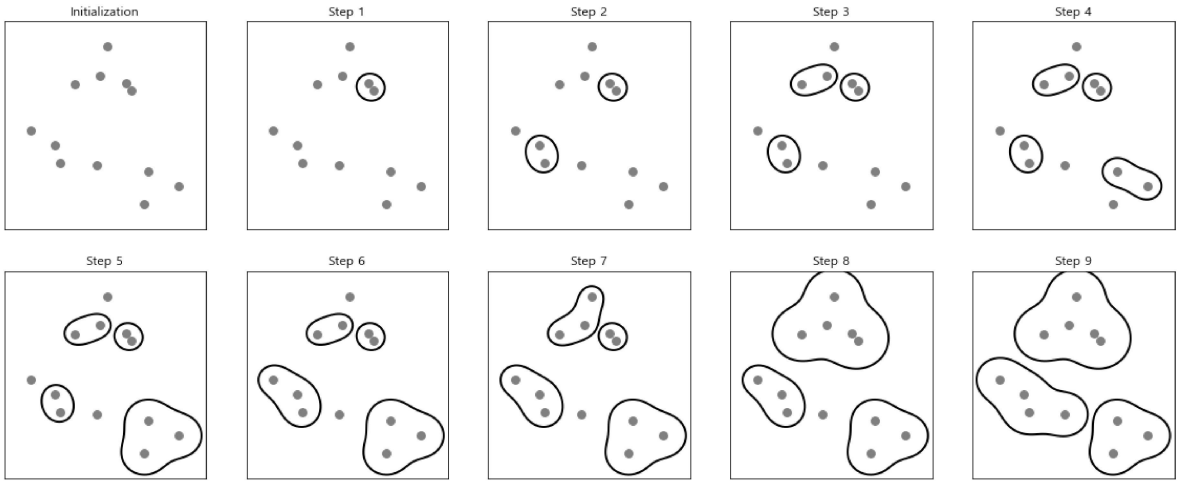
In [1]:

```
import mglearn
import matplotlib.pyplot as plt

### 한글
import matplotlib
from matplotlib import font_manager, rc
font_loc = "C:/Windows/Fonts/malgunbd.ttf"
font_name = font_manager.FontProperties(fname=font_loc).get_name()
matplotlib.rc('font', family=font_name)
```

In [2]:

```
mglearn.plots.plot_agglomerative_algorithm()
```



- 초기의 각 포인트가 하나의 클러스터이다.
- 각 단계에서 가장 가까운 두 클러스터가 합쳐진다.
- 네 번째 단계까지는 포인트가 하나뿐인 클러스터가 두 개가 선택되어 두포인트를 가진 클러스터가 되었다.
- 단계 5에서는 두 개의 포인트를 가진 클러스터 중 하나가 세 개의 포인트로 확장됩니다.

In [3]:

```
from sklearn.datasets import make_blobs
```

In [4]:

```
from sklearn.cluster import AgglomerativeClustering
X, y = make_blobs(random_state=1)

agg = AgglomerativeClustering(n_clusters=3)
assignment = agg.fit_predict(X)   # 클러스터의 소속 정보 얻기
assignment
```

Out[4]:

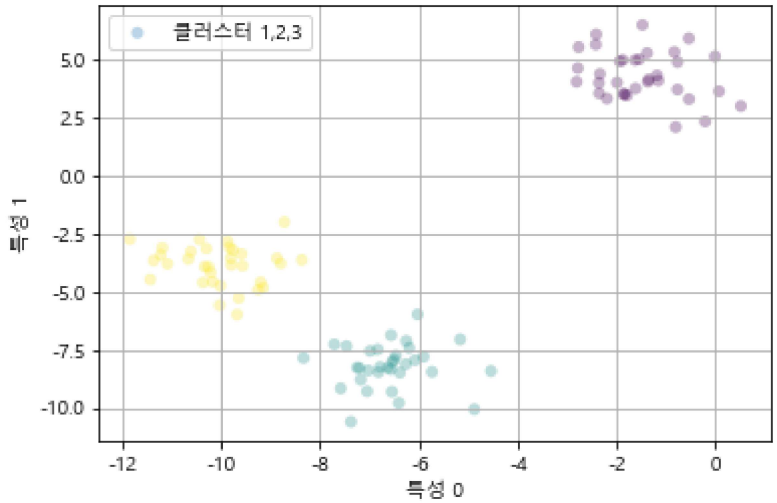
```
array([0, 2, 2, 2, 1, 1, 1, 2, 0, 0, 2, 2, 1, 0, 1, 1, 1, 0, 2, 2, 1, 2,
       1, 0, 2, 1, 1, 0, 0, 1, 0, 0, 1, 0, 2, 1, 2, 2, 2, 1, 1, 2, 0, 2,
       2, 1, 0, 0, 0, 0, 2, 1, 1, 1, 0, 1, 2, 2, 0, 0, 2, 1, 1, 2, 2, 1,
       0, 1, 0, 2, 2, 2, 1, 0, 0, 2, 1, 1, 0, 2, 0, 2, 2, 1, 0, 0, 0, 0,
       2, 0, 1, 0, 0, 2, 2, 1, 1, 0, 1, 0], dtype=int64)
```

In [5]:

```
import matplotlib
matplotlib.rcParams['axes.unicode_minus'] = False
```

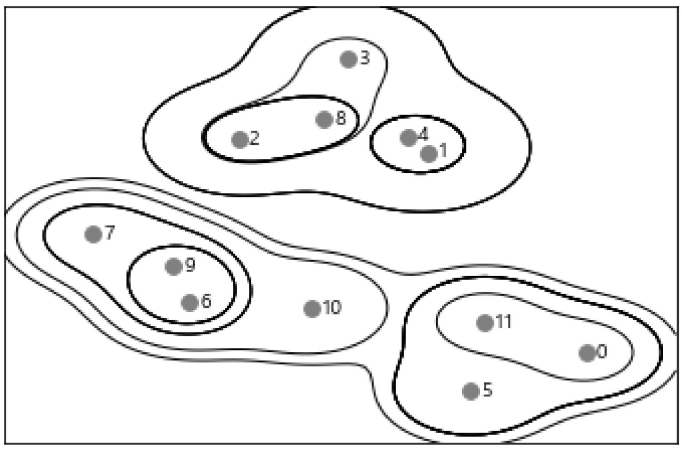
In [6]:

```
plt.scatter(X[:,0],X[:,1], label="클러스터 1,2,3",
            alpha=0.3, edgecolors='none', c=assignment)
plt.legend()
plt.grid(True)
plt.xlabel("특성 0")
plt.ylabel("특성 1")
plt.show()
```



In [7]:

```
mglearn.plots.plot_agglomerative()
```



위의 내용 포함 다른 거리 개념

군집방법	두 군집 사이의 거리 정의
single linkage	한 군집의 점과 다른 군집의 점 사이의 가장 짧은 거리(shortest distance)
complete linkage	한 군집의 점과 다른 군집의 점 사이의 가장 긴 거리(longest distance)
average linkage	한 군집의 점과 다른 군집의 점 사이의 평균 거리. UPGMA(unweighted pair group mean averaging)이라고도 한다.
centroid	두 군집의 centroids(변수 평균의 벡터) 사이의 거리.관측치가 하나인 경우 centroid는 변수의 값이 된다
Ward	모든 변수들에 대하여 두 군집의 ANOVA sum of square를 더한 값

기타 알아보기

시각화 도구 : 3차원 이상의 데이터 시각화 (덴드로 그램-dendrogram)- 현재 scikit learn에서 제공하지 않음.

REF

- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html)