6. K-means/KNN and PCA

Table of Contents

- Using Iris data
  - K-means clustering
  - KNN classification
  - PCA analysis

SNU BIOINFORMATICS INSTITUTE

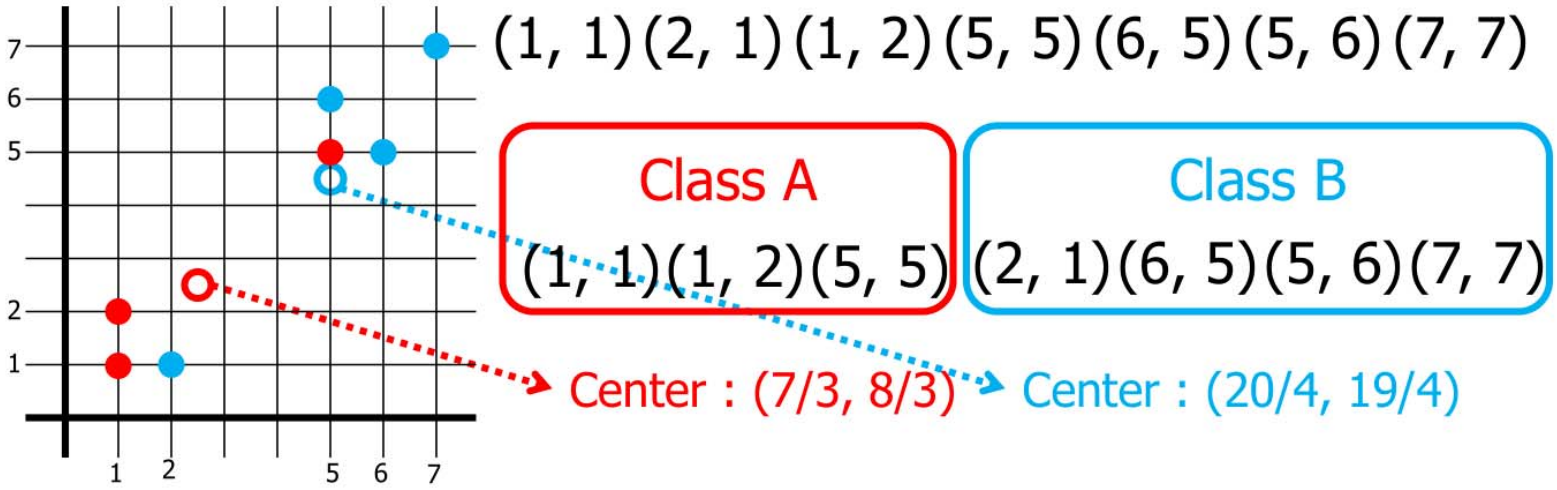
K-means clustering using sklearn

- class sklearn.cluster.Kmeans
  - KMeans (*n_clusters=8*,
    init='k-means++',
    n_init=10,
    max_iter=300,
    tol=0.0001,
    precompute_distances='auto',
    verbose=0,
    random_state=None,
    copy_x=True,
    n_jobs=1,
    algorithm='auto')

SNU BIOINFORMATICS INSTITUTE

K-means clustering example

- For 2 dimensional data, initialize classes of each data point
- This is to initialize the centroids (K=2)
- The centroids are set as K_1=(7/3, 8/3) and K_2=(20/4, 19/4)

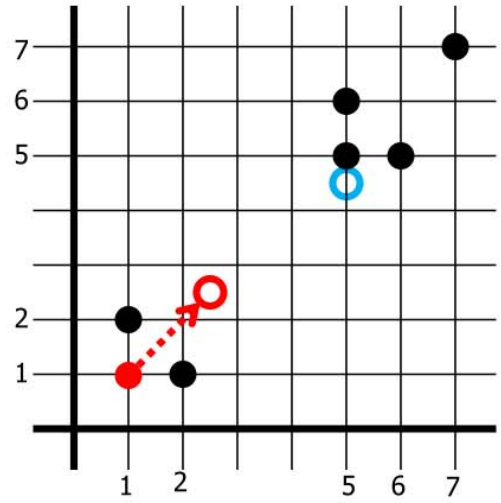


Credit to Professor 심규석

SNU BIOINFORMATICS INSTITUTE

K-means clustering example

- Re-assign each data point to the closest centroid



(1, 1) (2, 1) (1, 2) (5, 5) (6, 5) (5, 6) (7, 7)

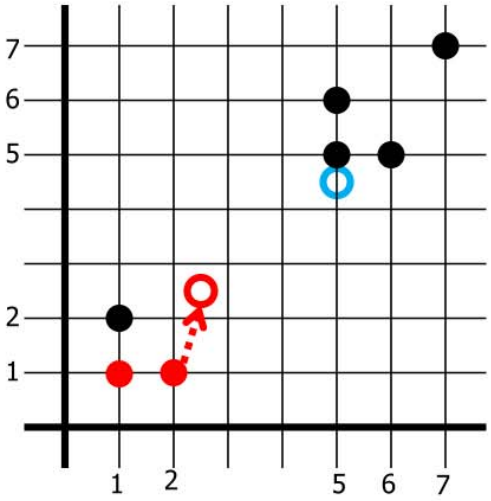
Class A	Class B
(1, 1)	

Center : (7/3, 8/3) Center : (20/4, 19/4)

Credit to Professor 심규석

SNU BIOINFORMATICS INSTITUTE

K-means clustering example



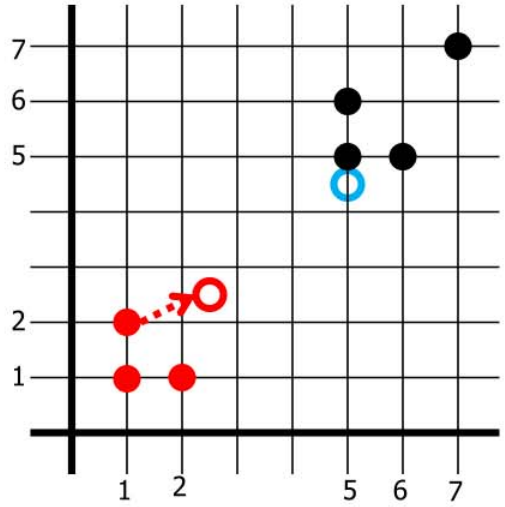
(1, 1) *(2, 1)* (1, 2) (5, 5) (6, 5) (5, 6) (7, 7)

Class A	Class B
(1, 1)(2, 1)	
Center : (7/3, 8/3)	Center : (20/4, 19/4)

Credit to Professor 심규석

SNU BIOINFORMATICS INSTITUTE

K-means clustering example



(1, 1) (2, 1) *(1, 2)* (5, 5) (6, 5) (5, 6) (7, 7)

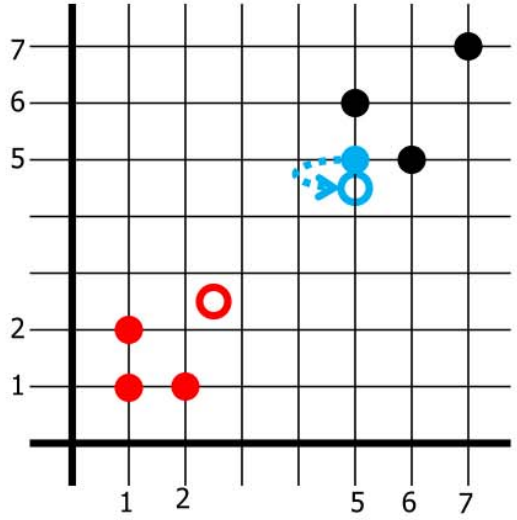
Class A	Class B
(1, 1) (2, 1) (1, 2)	

Center : (7/3, 8/3) Center : (20/4, 19/4)

Credit to Professor 심규석

SNU BIOINFORMATICS INSTITUTE

K-means clustering example



(1, 1) (2, 1) (1, 2) *(5, 5)* (6, 5) (5, 6) (7, 7)

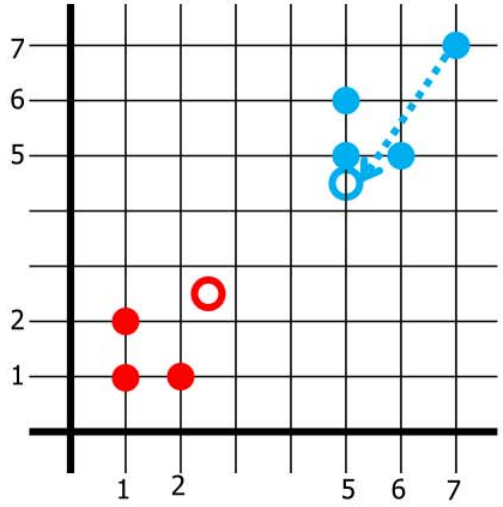
Class A	Class B
(1, 1) (2, 1) (1, 2)	(5, 5)

Center : (7/3, 8/3) Center : (20/4, 19/4)

Credit to Professor 심규석

SNU BIOINFORMATICS INSTITUTE

K-means clustering example



(1, 1)(2, 1)(1, 2)(5, 5)(6, 5)(5, 6) *(7, 7)*

Class A	Class B
(1, 1)(2, 1)(1, 2)	(5, 5)(6, 5)(5, 6)(7, 7)

Center : (7/3, 8/3) Center : (20/4, 19/4)

Credit to Professor 심규석

SNU BIOINFORMATICS INSTITUTE

Disadvantages of K-means

- The k-means algorithm is sensitive to outliers !
  - Since an object with an extremely large value may substantially distort the distribution of the data.
- K-Medoids: Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.
- K-Means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-spherical shapes

Credit to Professor 심규석

SNU BIOINFORMATICS INSTITUTE

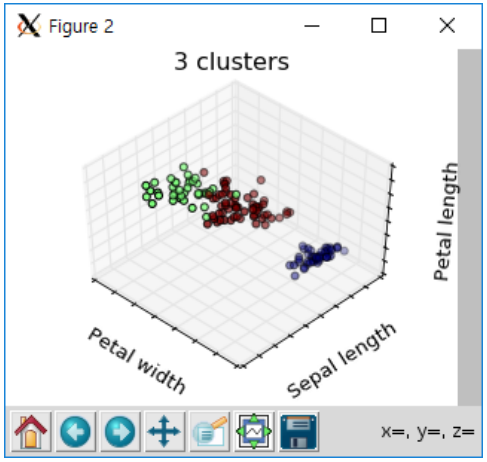
K-means using sklearn KMeans

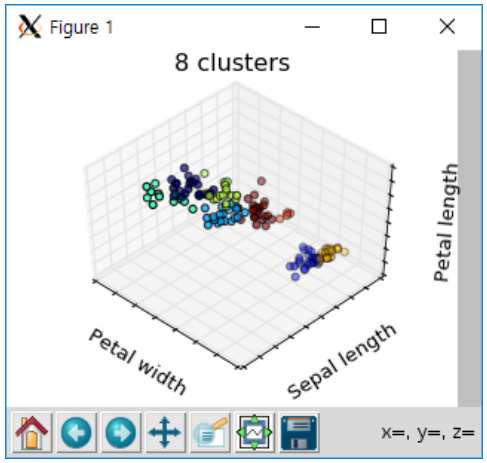
- Simply fit the data using KMeans function

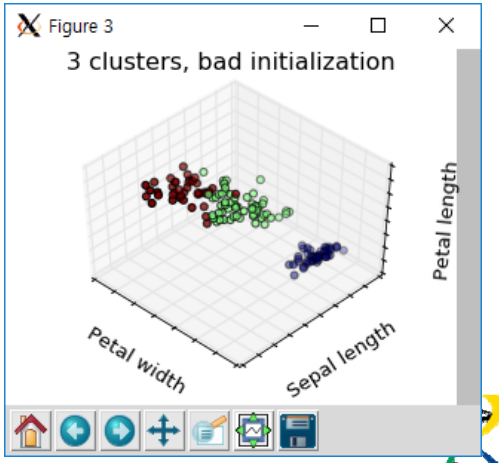
```
from sklearn.cluster import KMeans
import numpy as np
X = np.array([[1, 2], [1, 4], [1, 0],
              [4, 2], [4, 4], [4, 0]])
kmeans = KMeans(n_clusters=2, random_state=0).fit(X)

print "Clusters %s"%(kmeans.labels_)
print "Cluster centroids: %s"%(kmeans.cluster_centers_)
print "Prediction cluster of [0, 0], [4, 4]: %s"%(kmeans.predict([[0,
0], [4, 4]]))
```

K-means using Iris data

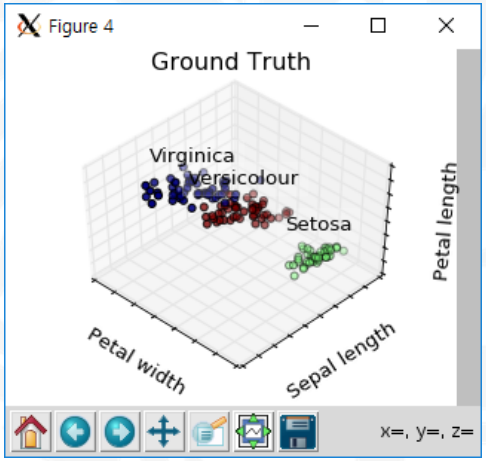
- Load iris data
- Fit data using KMeans with three different settings
  - K=3, 8 and bad initialization (random)







Answer



SNU BIOINFORMATICS INSTITUTE

Measure cluster quality with different scores

- There are several score metrics for measuring clustering quality
  - Homogeneity
  - Completeness
  - V-means
  - Adjusted rand index (ARI)
  - Adjusted mutual information (AMI)
  - Silhouette score

Estimator	Homogeneity	Completeness	V-means	ARI	AMI	Silhouette
k=8	0.926	0.51	0.658	0.456	0.498	0.363
k=3	0.751	0.765	0.758	0.73	0.748	0.553
k=3(random init)	0.736	0.747	0.742	0.716	0.733	0.551

SNU BIOINFORMATICS INSTITUTE

K-Nearest Neighbor (KNN) classification using sklearn

- class sklearn.neighbors.KNeighborsClassifier(

  n_neighbors=5,

  weights='uniform',

  algorithm='auto',

  leaf_size=30,

  p=2,

  metric='minkowski',

  metric_params=None,

  n_jobs=1,

  **kwargs)

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

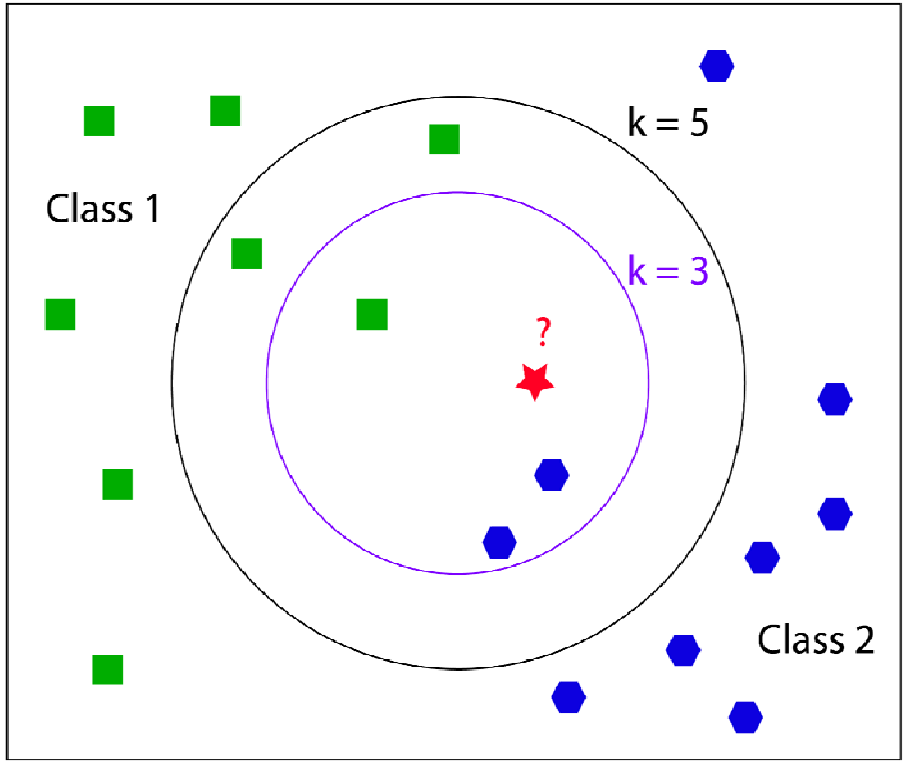
Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sqrt{\sum_{i=1}^{k}(|x_i - y_i|)^q}\right)^{1/q}$

SNU BIOINFORMATICS INSTITUTE

K-Nearest Neighbor (KNN)

- The class of input data is determined by voting of K-neighbors
- Actually, a classification model is not generated by KNN

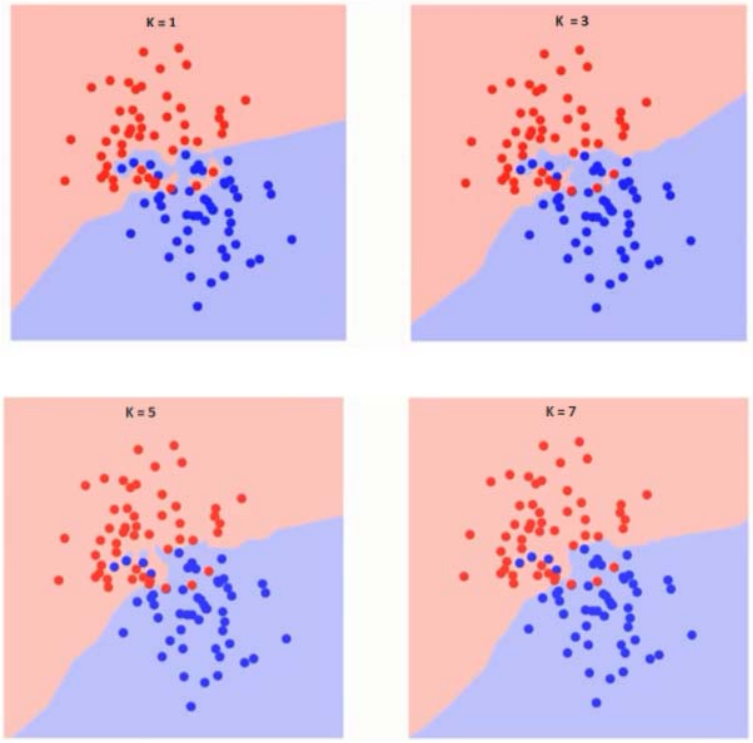


SNU BIOINFORMATICS INSTITUTE

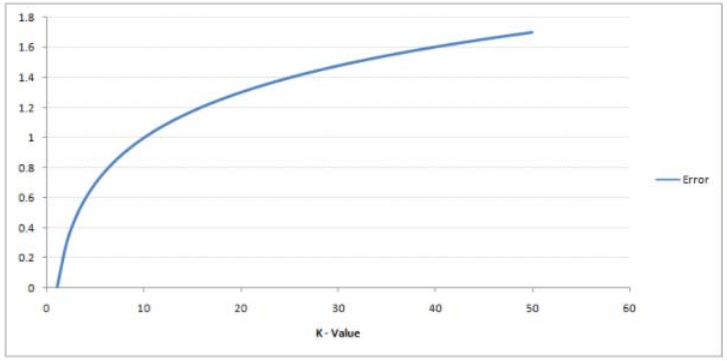
Disadvantages of KNN

- The classification is done on the fly
- For each input data, KNN must compute K-neighbor distances
- This can be very slow for a large input data

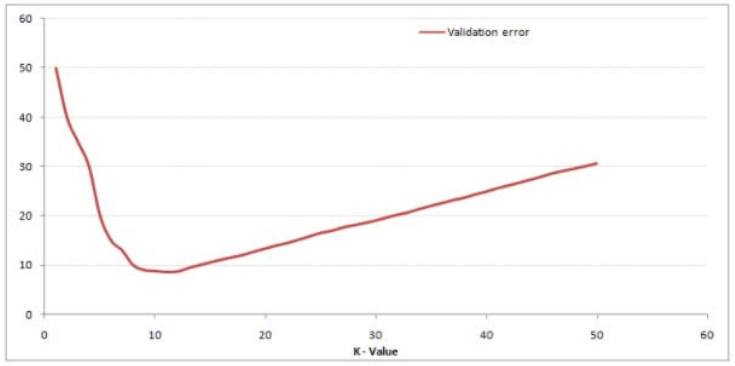
How do we choose K for KNN?



Error



Validation error

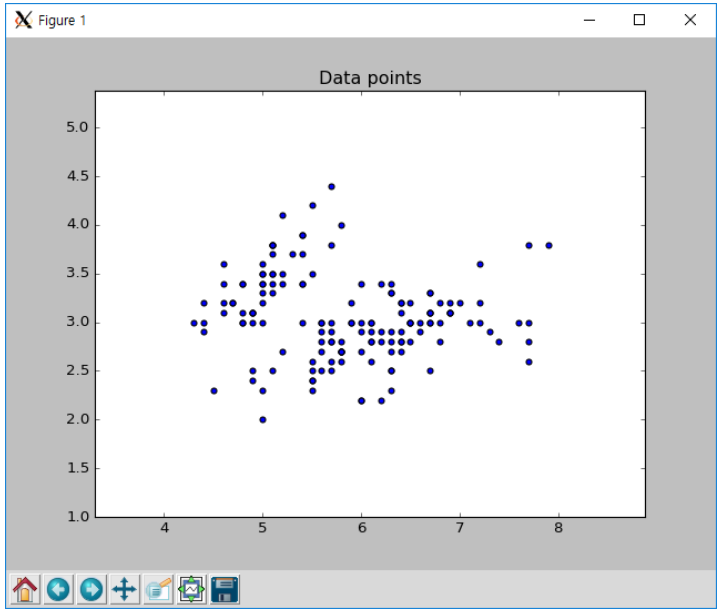


SNU BIOINFORMATICS INSTITUTE

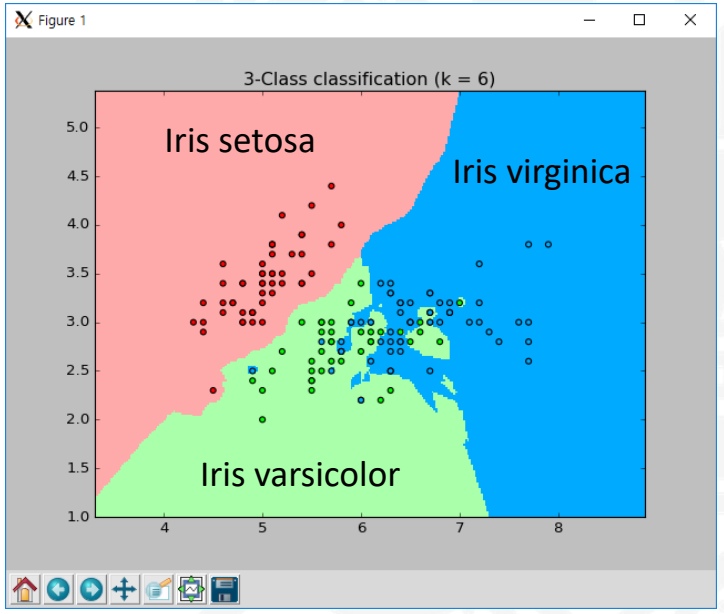
KNN example using Iris data

- The colored regions show the classification area for future input data

Input data



Classifier



SNU BIOINFORMATICS INSTITUTE

PCA analysis

- Principal component analysis is a method that reduces dimensionality to project the data onto a lower dimensional space

SNU BIOINFORMATICS INSTITUTE

PCA using sklearn decomposition class

- class sklearn.decomposition.PCA(

  n_components=None,

  copy=True,

  whiten=False,

  svd_solver='auto',

  tol=0.0,

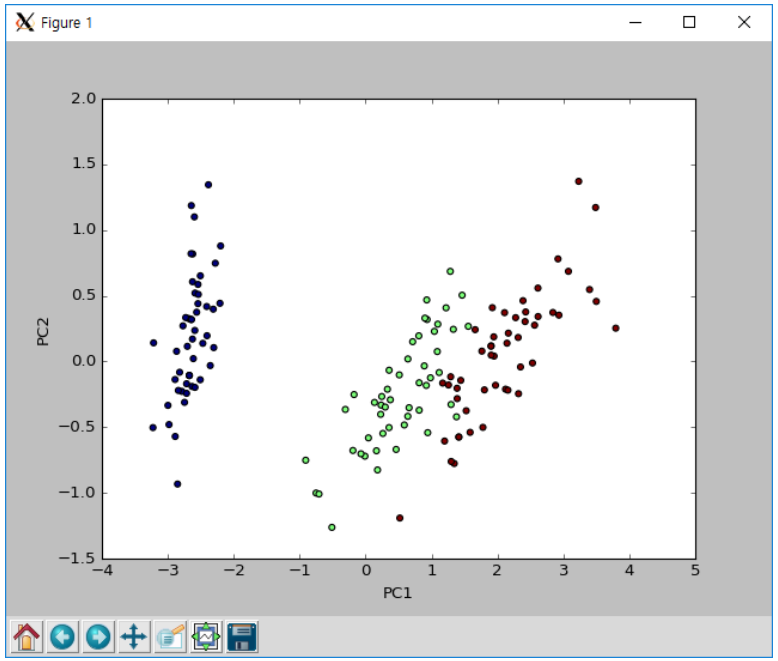
  iterated_power='auto',

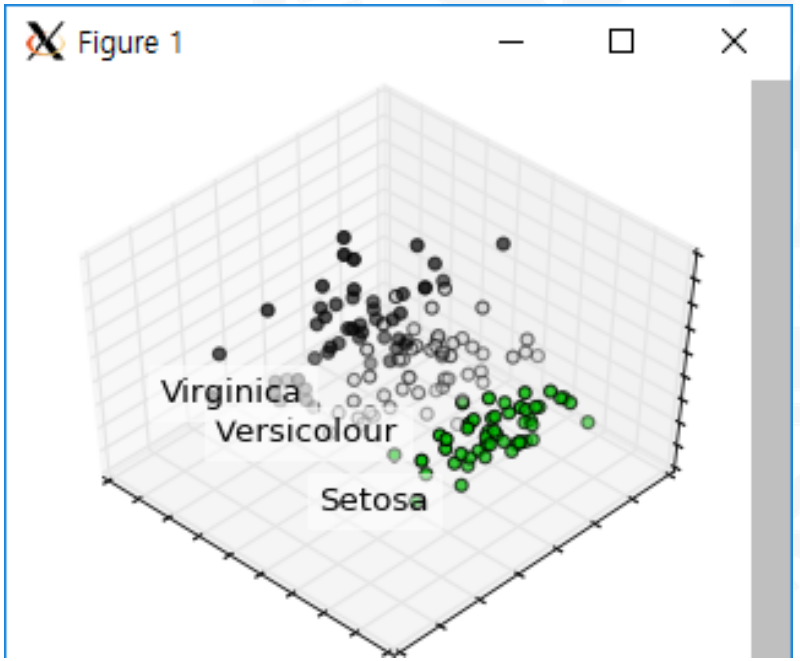
  random_state=None)

SNU BIOINFORMATICS INSTITUTE

PCA example using Iris data

- Load iris data
- Perform PCA and observe (PC1, PC2, PC3)
- Compare with (PC2, PC3) plot





SNU BIOINFORMATICS INSTITUTE