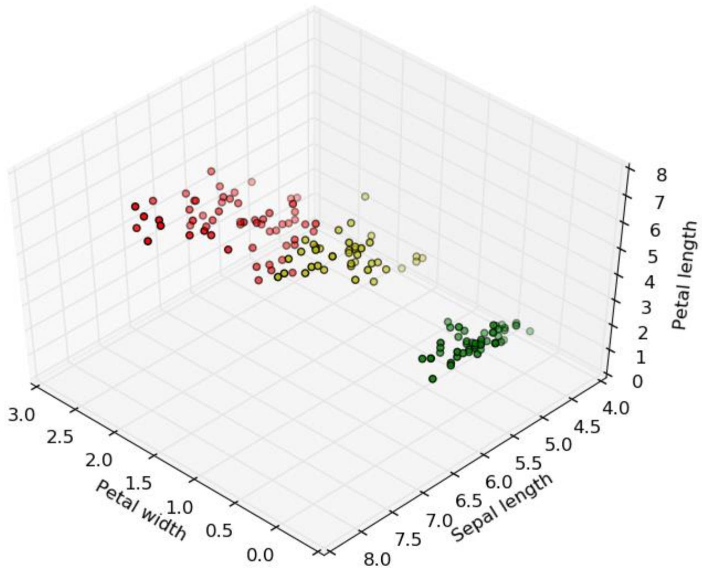
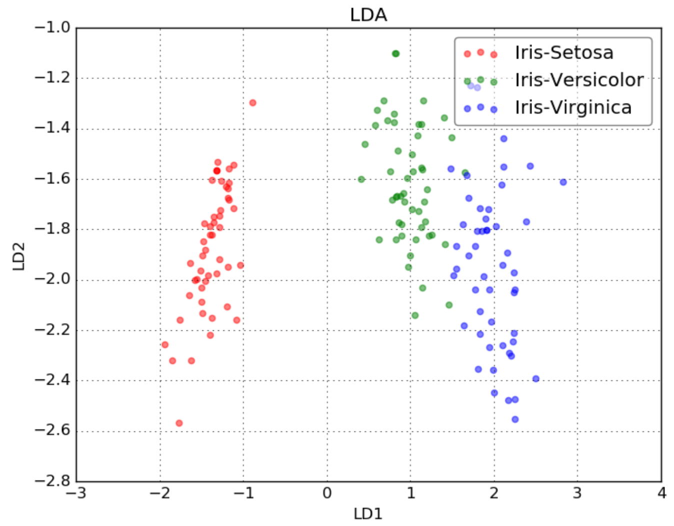
K-means && LDA

machine learning basic





Vision@OUC

Wang Chao

Group of DL

Overview

- K-means
- LDA
- Useful tools
- Q&A

What is clustering

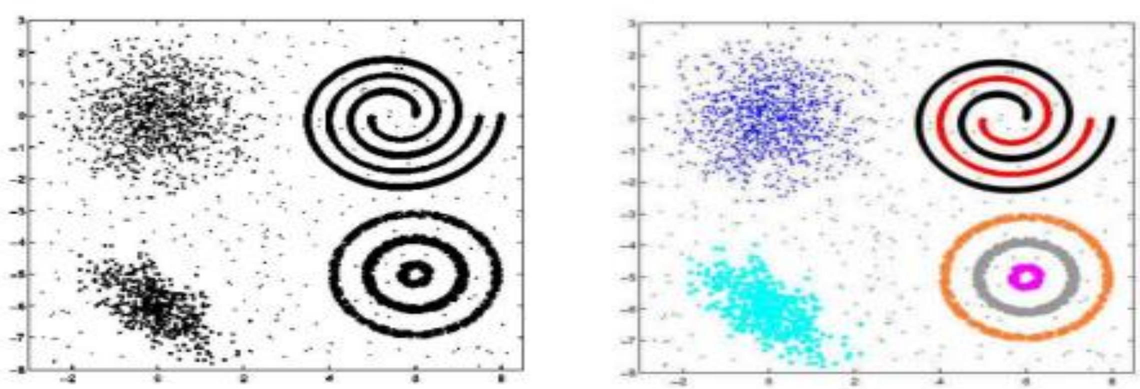
- Clustering is an unsupervised learning algorithm
- Goal: Automatically segment data into groups of similar points
- The only information clustering uses is the similarity between samples
- Clustering groups examples based of their mutual similarities
- A good clustering:
  - High within-cluster similarity
  - Low inter-cluster similarity

When and why we want to do this?

- Automatically organizing data
- Understanding hidden structure in some data
- Representing high-dimensional data in a low-dimensional space

K-means

- Different clustering algorithms use the data and distance measurements in different ways.
- K-means : the simplest clustering algorithm
  - The basic idea is to describe each cluster by its mean value.
  - The goal of K-means is to group the samples into K partitions



(a) Input data (b) Desired clustering

K-means algorithm

- Dataset: *Iris* flower data set

Fisher's *Iris* Data

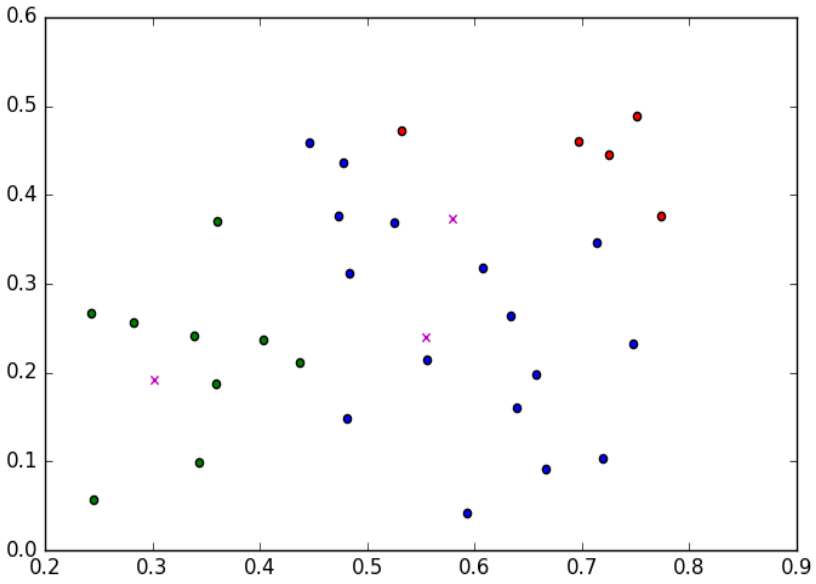
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

Input: *150 samples* $\{x_1, x_2, x_3, x_4\}$

R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics.

Initialization

- Randomly initialized anywhere in $\mathbb{R}^D$ ($D=4$)
- Choose any K examples as the cluster centers



Iterate

- Assign each of examples x_n to its closest cluster center

$$C_k = \{n\colon\ k = \arg\min \left|\left|x_n - \mu_k\right|\right|^2\}$$

(C_k is the set of samples closest to μ_k)

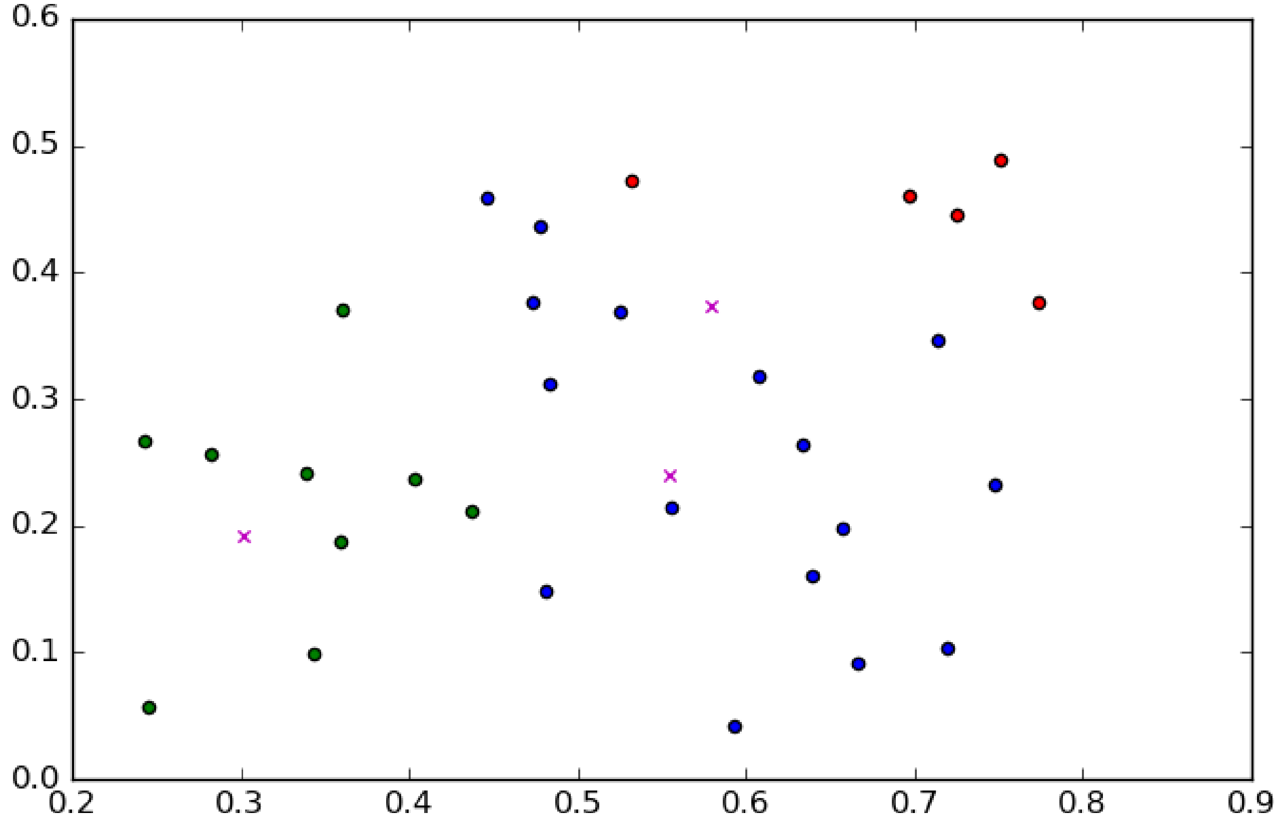
- Recompute the new cluster centers $\mu_k(mean\ of\ centroid\ of\ set\ C_k)$ to its closest cluster center

$$\mu_k = \frac{1}{|C_k|}\sum_n x_n$$

- Repeat while not converged

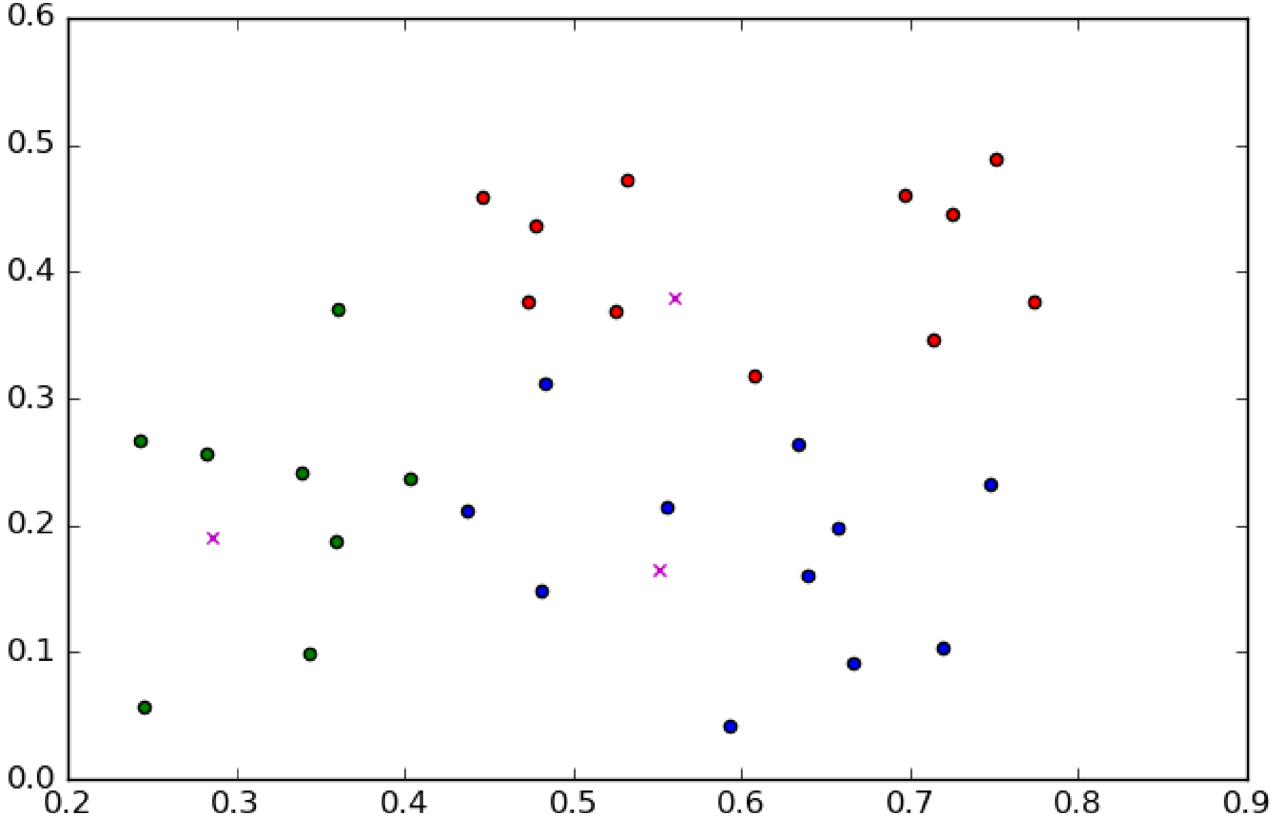
convergence criteria: cluster centers do not changes any more

K-means: Initialization(assume K=3)

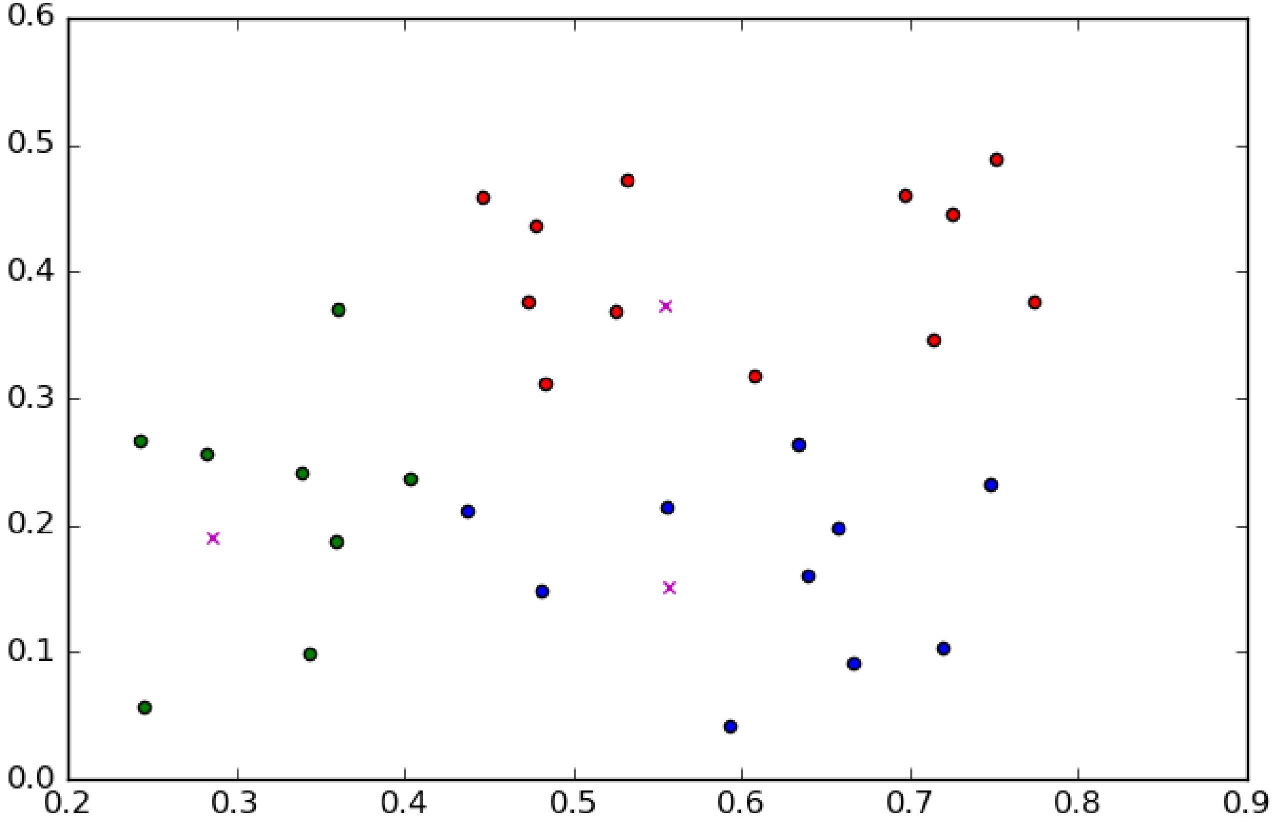


dataset from 《Machine learning》（Zhihua Zhou P202）

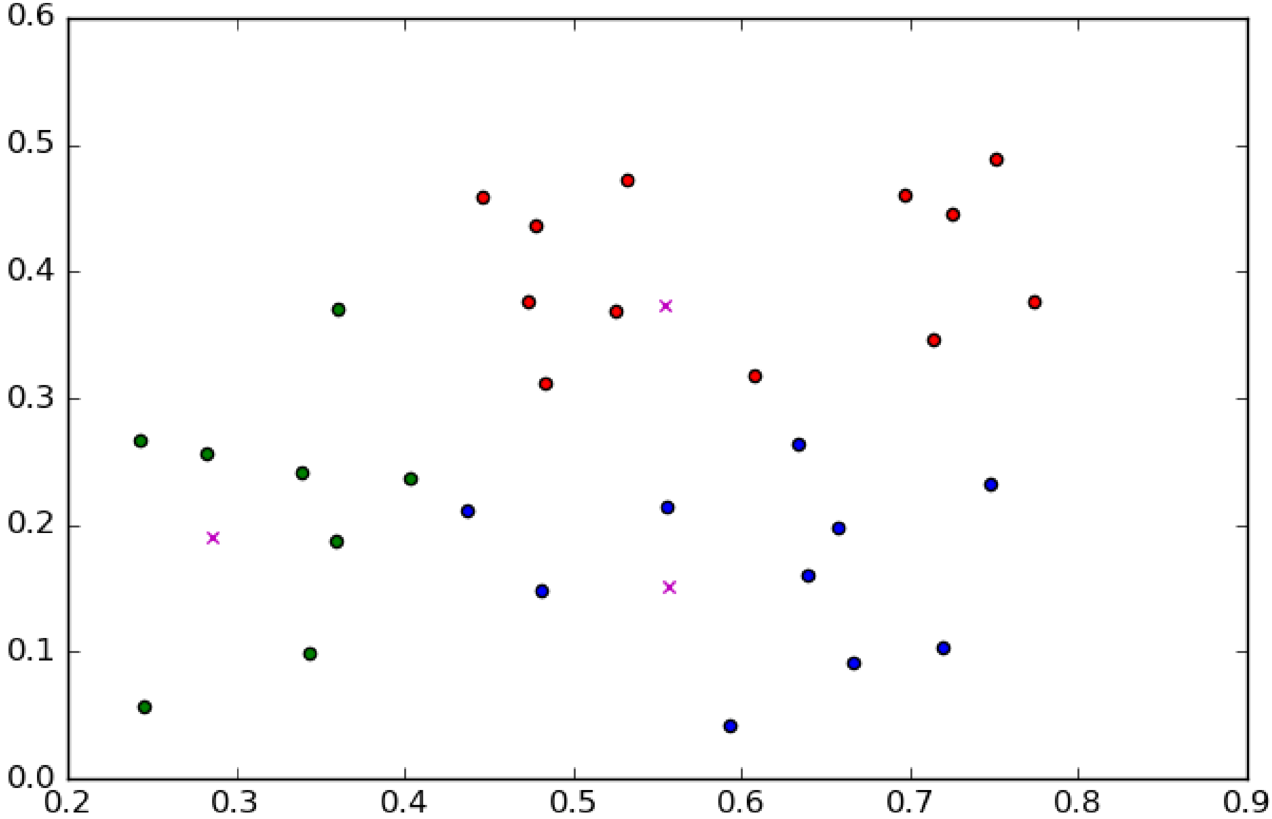
K-means: Iteration1



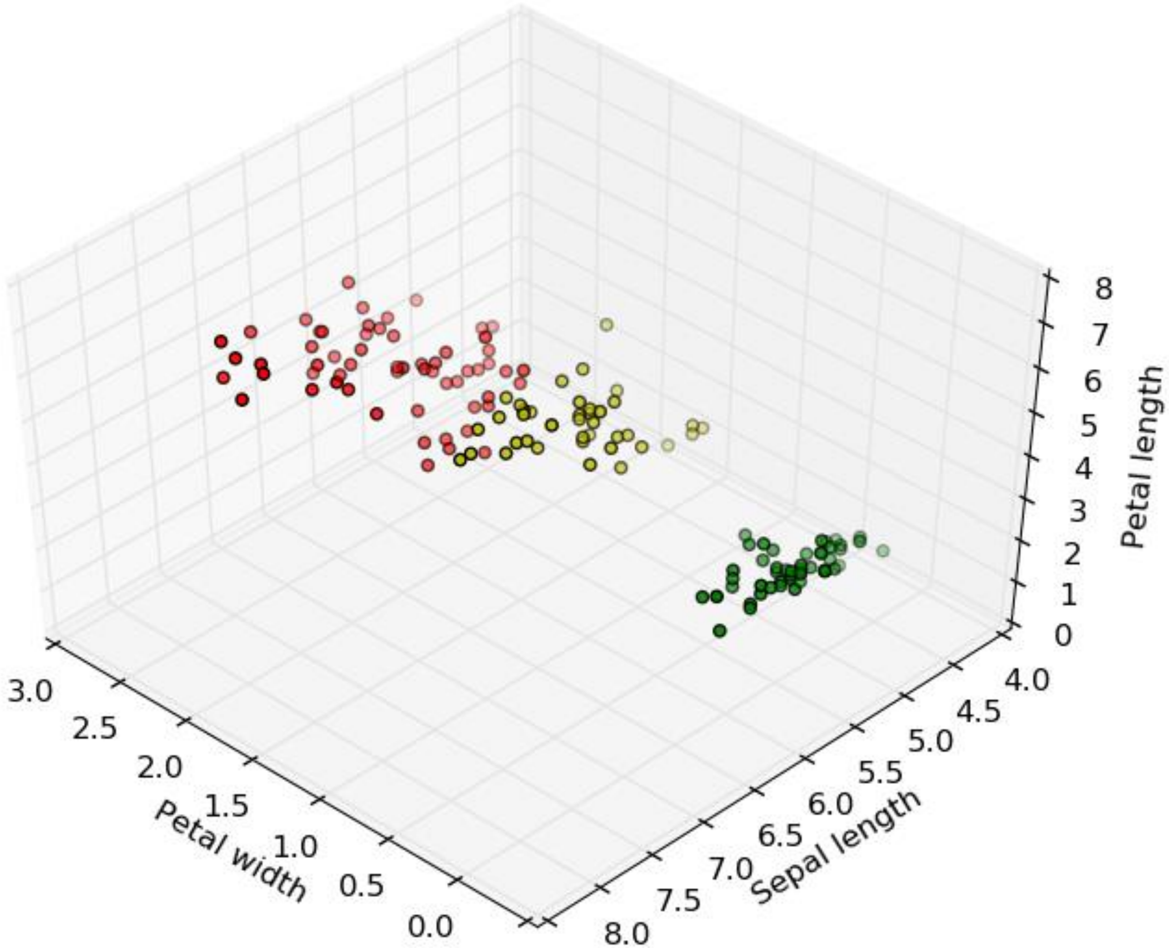
K-means: Iteration2



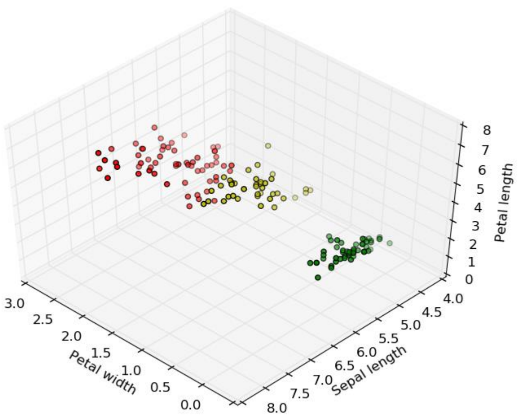
K-means: Iteration3



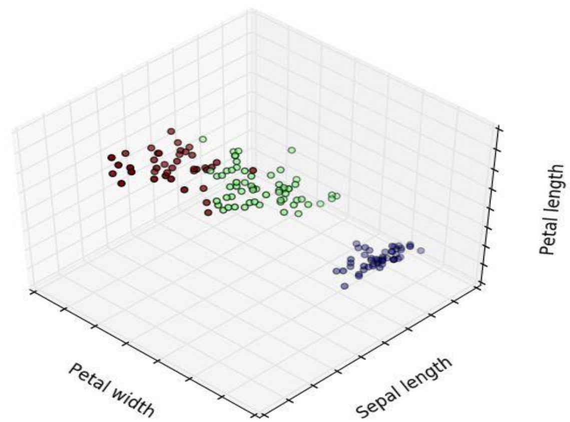
Examples on Iris dataset (iteration=10)



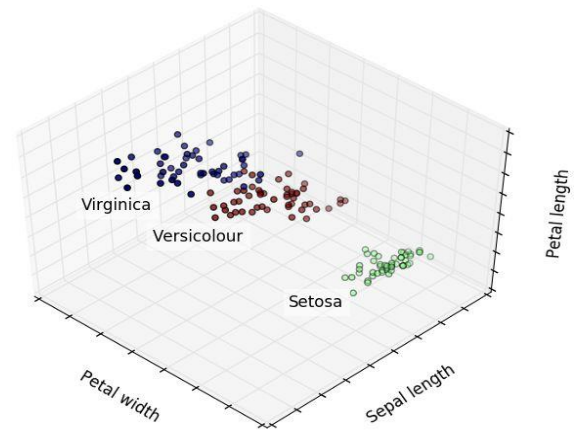
K-means VS scikit-learn VS ground truth



My result

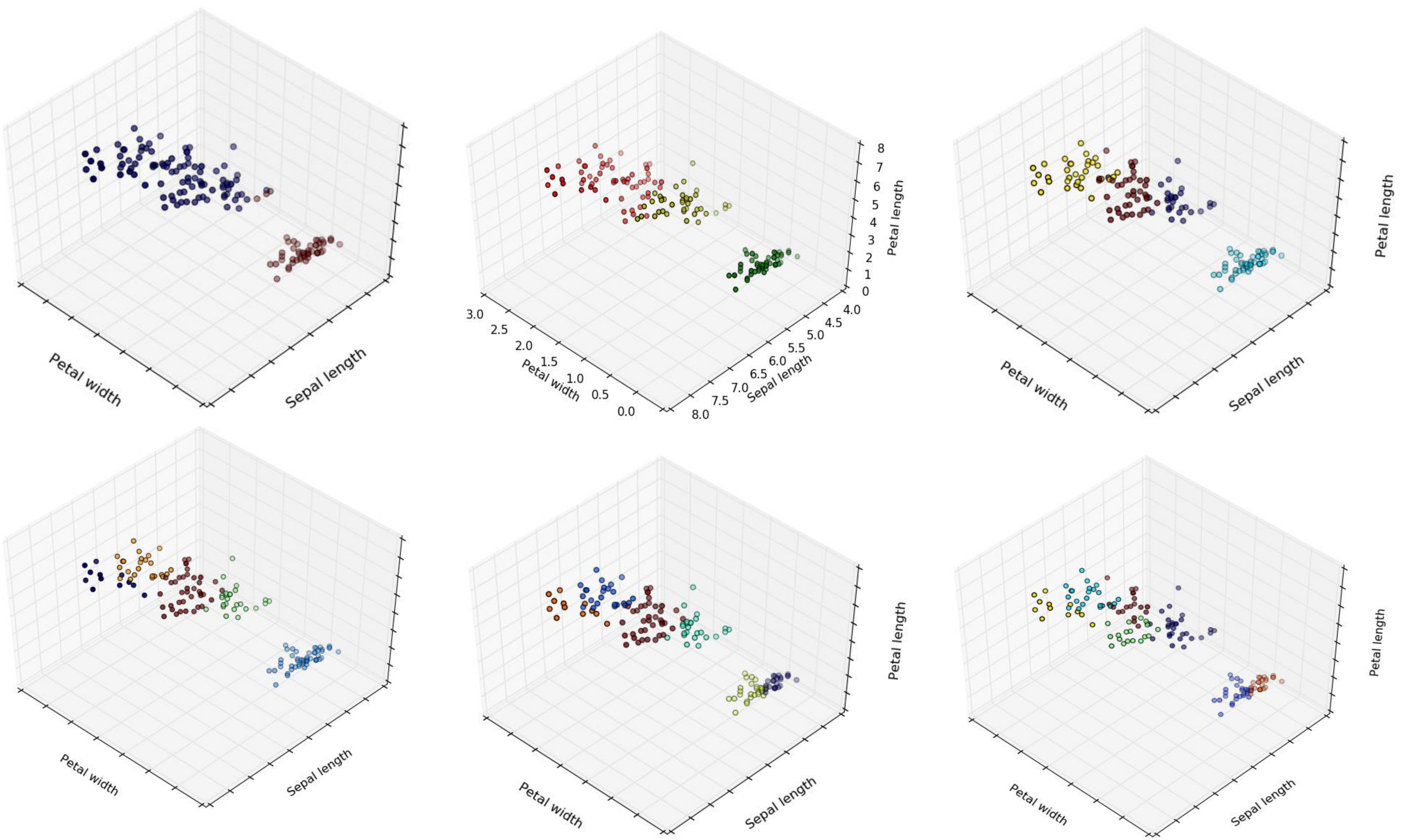


Result on scikit-learn



Ground truth

Initialize with different number of cluster center



Summary

- Advantages:
  - Computationally faster than hierarchical clustering
  - Fast to converge
  - Easy to relize

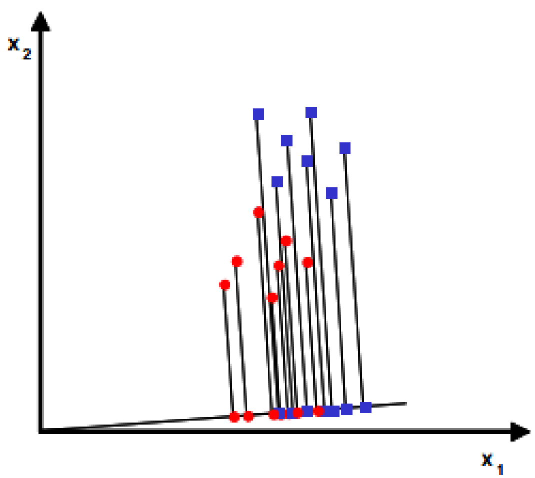
- Limitations:
  - Makes hard assignments of points to clusters
  - Sensitive to outlier samples(affect mean a lot)
  - Works well only for round shape

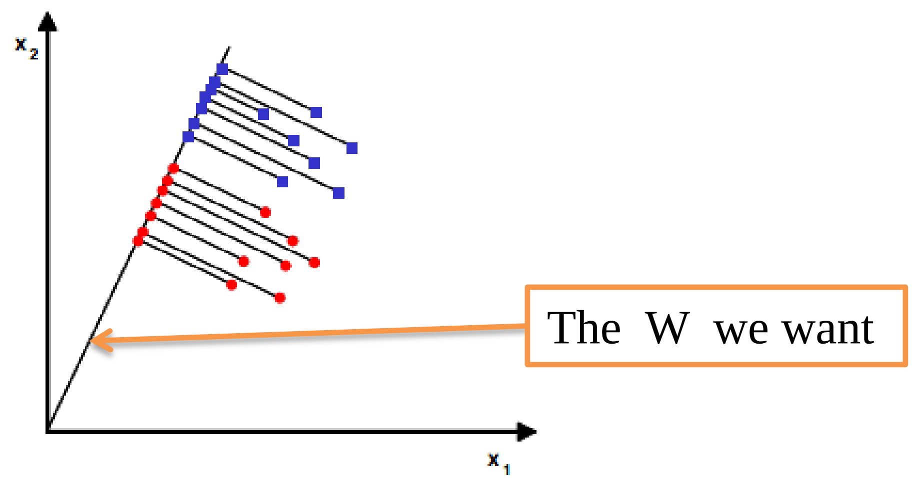
LDA（Linear Discriminant Analysis）

Ronald Fisher

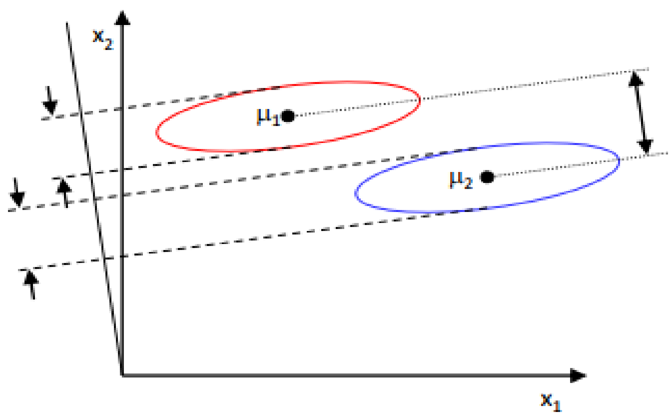
Objective

- LDA seeks to reduce dimensionality while preserving as much of class discriminatory information as possible
- Assume we have a set of D-dimensional samples $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots \boldsymbol{x_n}\}$, which include two class $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$
- We seek to obtain a scalar **y** by projecting the samples x onto a line $\boldsymbol{y} = \boldsymbol{w}^T \boldsymbol{x}$
- Of all the possible lines we would like to select the one that maximizes the separability of scalars





- **In order to find a good projection vector, we need to define a measure of separation**
- Fisher's solution
  - Fisher suggested maximizing difference between the means, normalized by a measure of within-class scatter
  - So the criterion function: $J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$
  - Therefore, we are looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible



R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics

To find the optimum w^*, we must express $J(w)$ as a function of w

– First, we define a measure of the scatter in feature space x

$$S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$
$$S_1 + S_2 = S_W$$

- where S_W is called the within-class scatter matrix

– The scatter of the projection y can then be expressed as a function of the scatter matrix in feature space x

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)^2 =$$
$$= \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w$$

– Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T \underbrace{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}_{S_B} w = w^T S_B w$$

- The matrix S_B is called the between-class scatter. Note that, since S_B is the outer product of two vectors, its rank is at most one

– We can finally express the Fisher criterion in terms of S_W and S_B as

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Maximum of J(w) use Lagrange Multiplier

- After a series of derivations, we get:

$$\boldsymbol{S_w}^{-1}\boldsymbol{S_b w} = \boldsymbol{\lambda w}$$

Examples on Iris datasets

Iris setosa

Iris versicolor

Petal(花瓣)

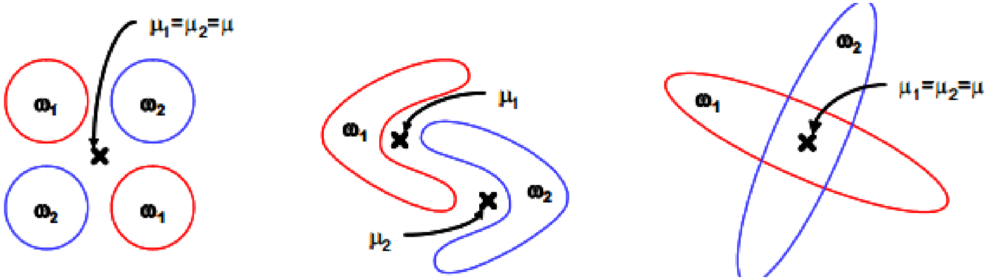
Sepal(花萼)

Iris virginica

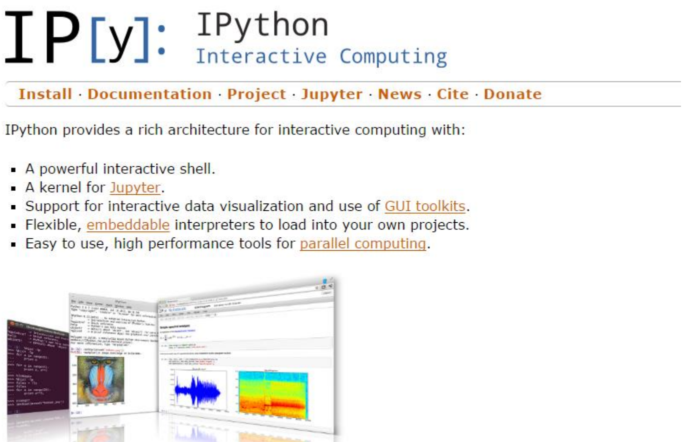
Summary

- Advantages:
  - Clear to reflect the difference in samples
  - supervised

- Limitations:
  - Produces at most C-1 feature projections
  - LDA is a parametric method since it assumes unimodal Gaussian likelihoods



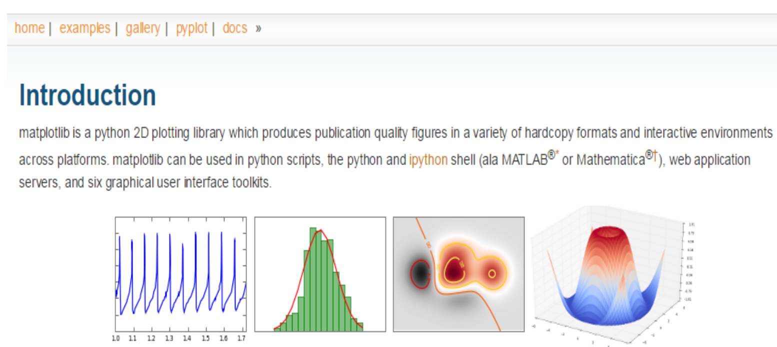
Useful tools



\+

jupyter

matplotlib



Q & A