LING 506 - TOPICS IN COMPUTATIONAL LINGUISTICS

Introductory Machine Learning

Yan Tang

Department of Linguistics, UIUC

Week 3

Last week…

- Clustering
  - To identify any natural patterns of groupings in data based on their "similarity"
  - Unsupervised ML
- Data visualisation
  - A quick way to find out explicit groups or patterns in data
  - Difficulty with data having over 3 dimensions

Last week…

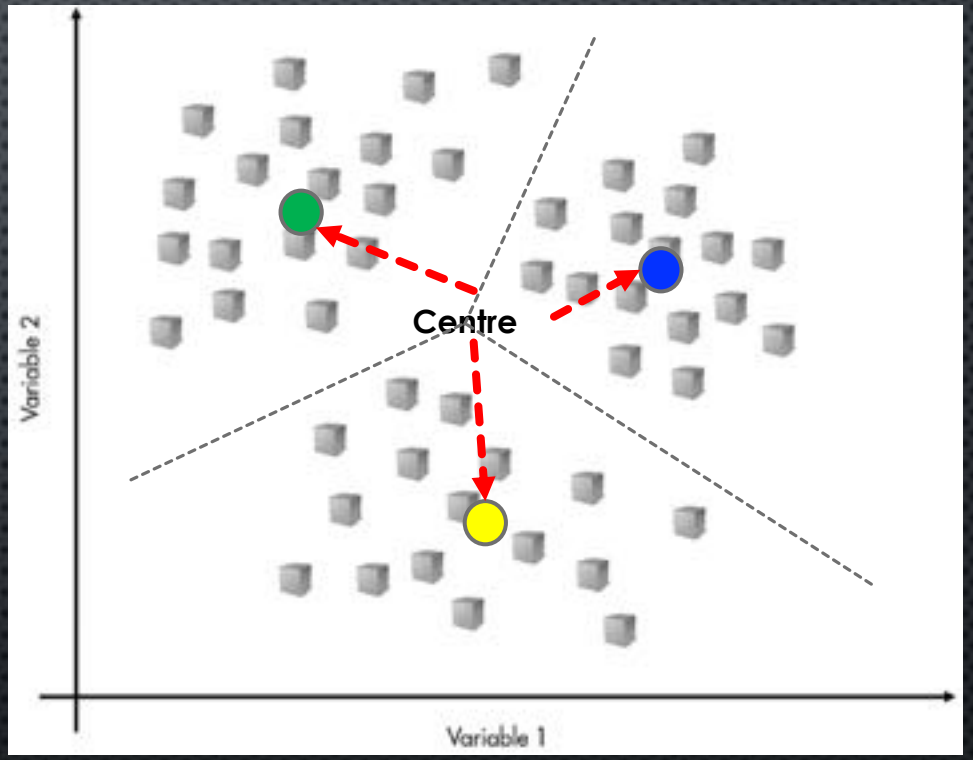
- Dimensionality reduction
  - Principal Component Analysis (PCA)
    - Transform data from one N-D space to another N-D space while preserving covariance of data
    - Widely used in statistics and other areas
  - Multiple-dimensional Scaling (MDS)
    - Transform data from one N-D space into smallest space while preserving pair-wise distance
    - Equivalent to PCA when using Euclidean distance

K-means clustering

- A partition method
  - Clusters/groups data by trying to separate samples in k groups of equal variance.
  - k must be specified before clustering
- K and mean?
  - K: the number of centroid
  - Means: a centroid which is considered as the central point of a cluster.
- Very good scalability when handling data

K-means: principle

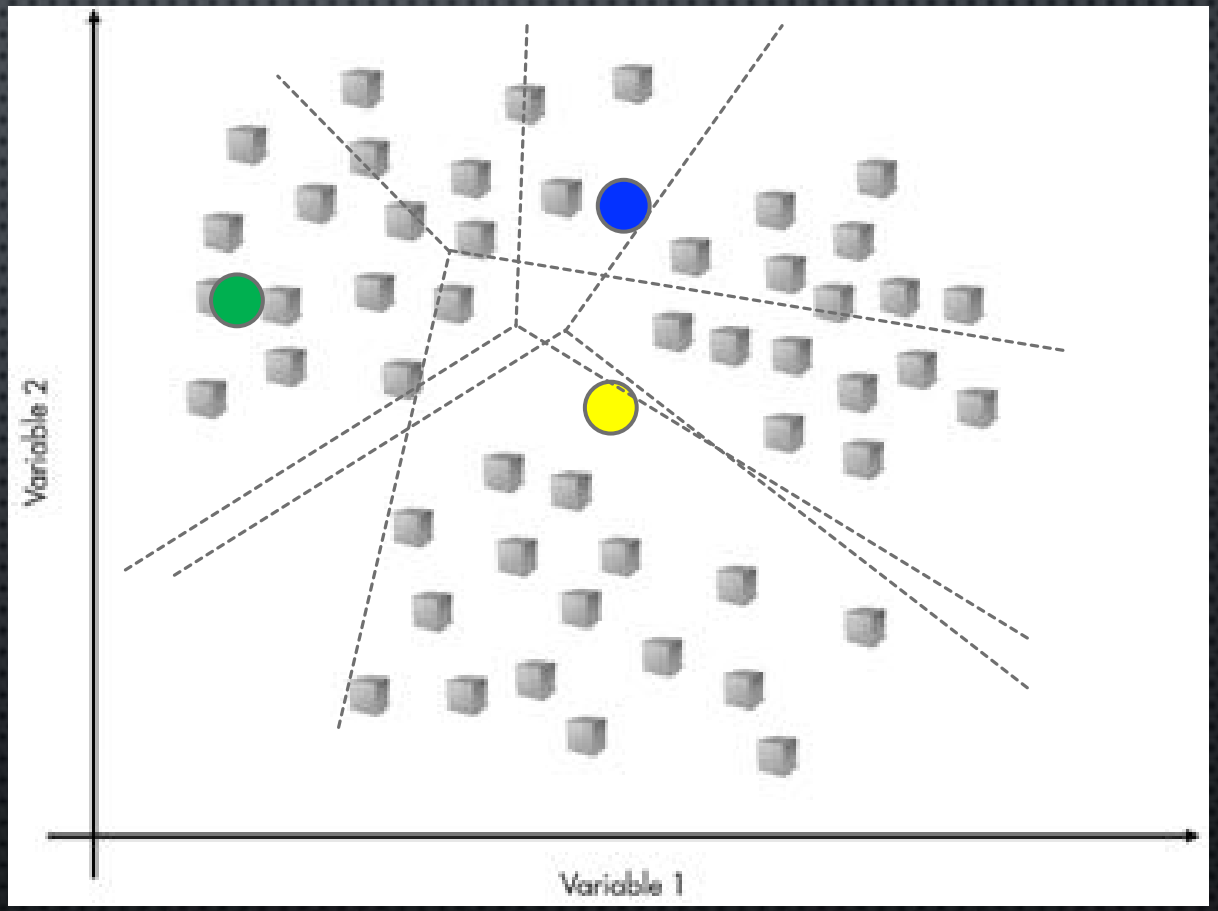
- K-means clustering operates on actual observations
- Each observation/data point treated as an object that has a location in space
- Find the centres for the k groups; assign each observation to the group whose centre is nearest to the observation



How are the centres determined?

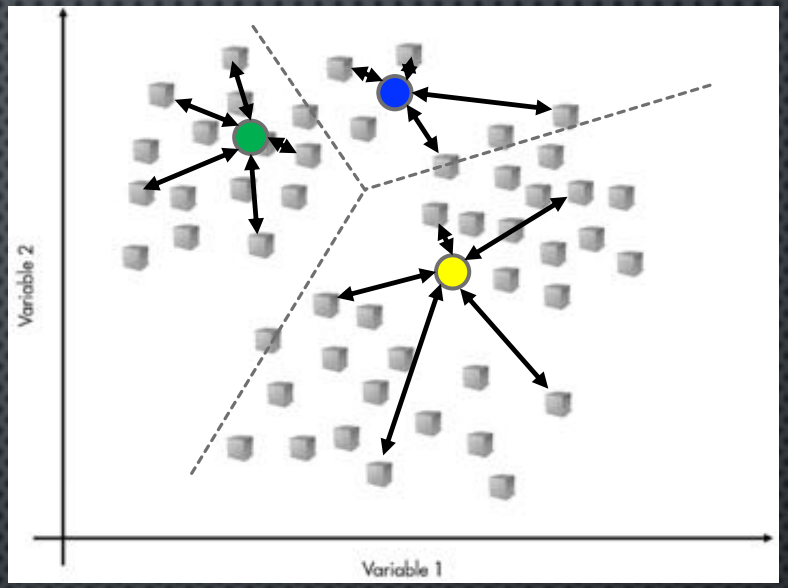
K-means: steps

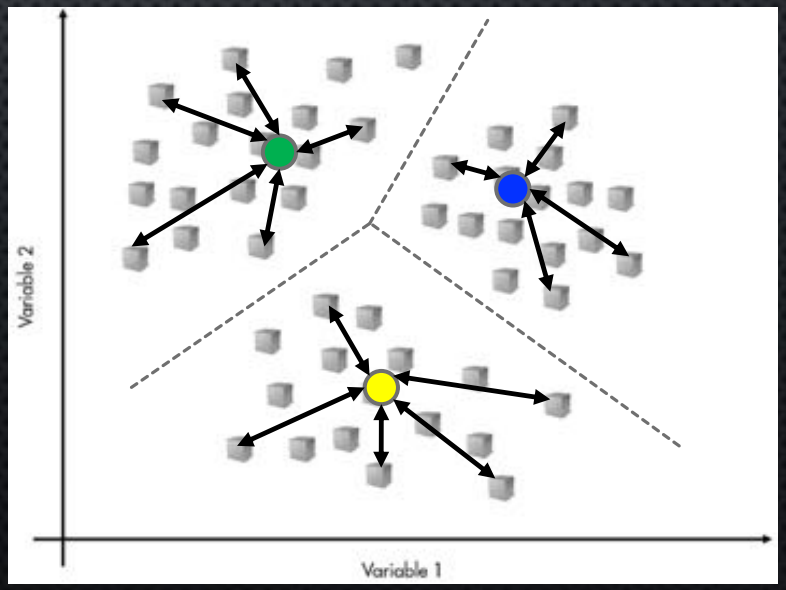
- Randomly select k data points as centroids
- Check the distance between data point and the centroids.
- Create new centroids by taking the mean value of all of the points assigned to each previous centroid.
- Reassign the data points to the cluster whose centroid is at a minimum distance from the data point.
- Repeat until the centroids do not move significantly



K-means: issues

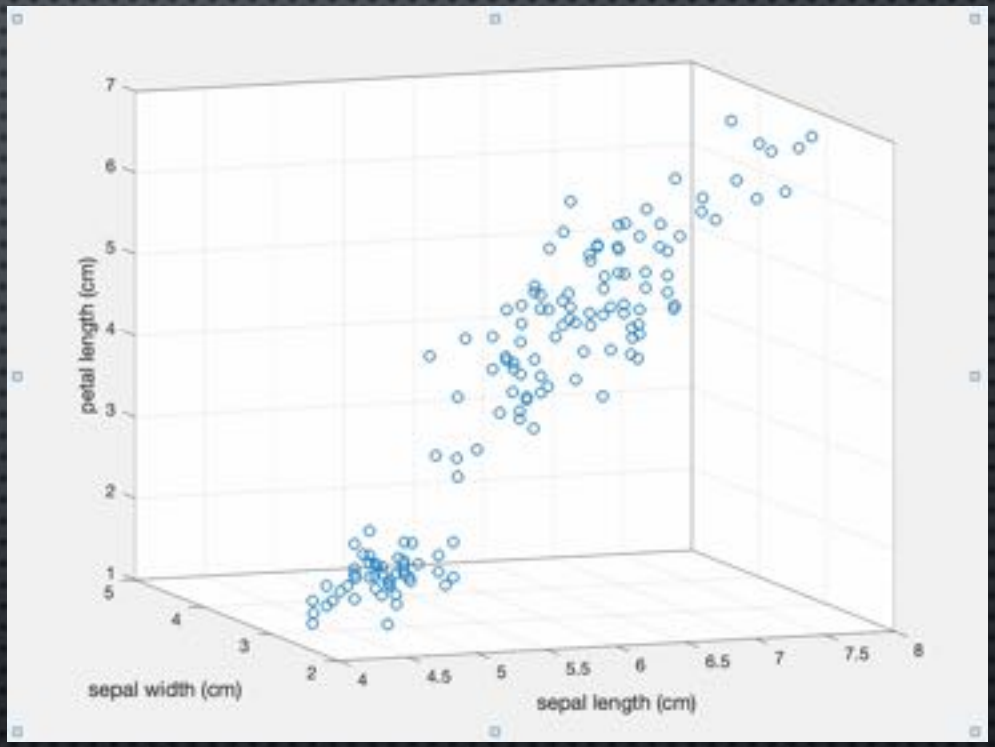
- Run several clustering trials using different randomly starting centroids
- The optimal solution leads to the lowest total sum of squared distance among all the trials
- Perform PCA before K-means to speed up computation for high dimensional data

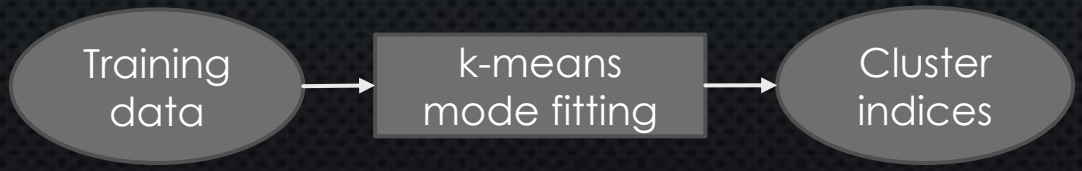




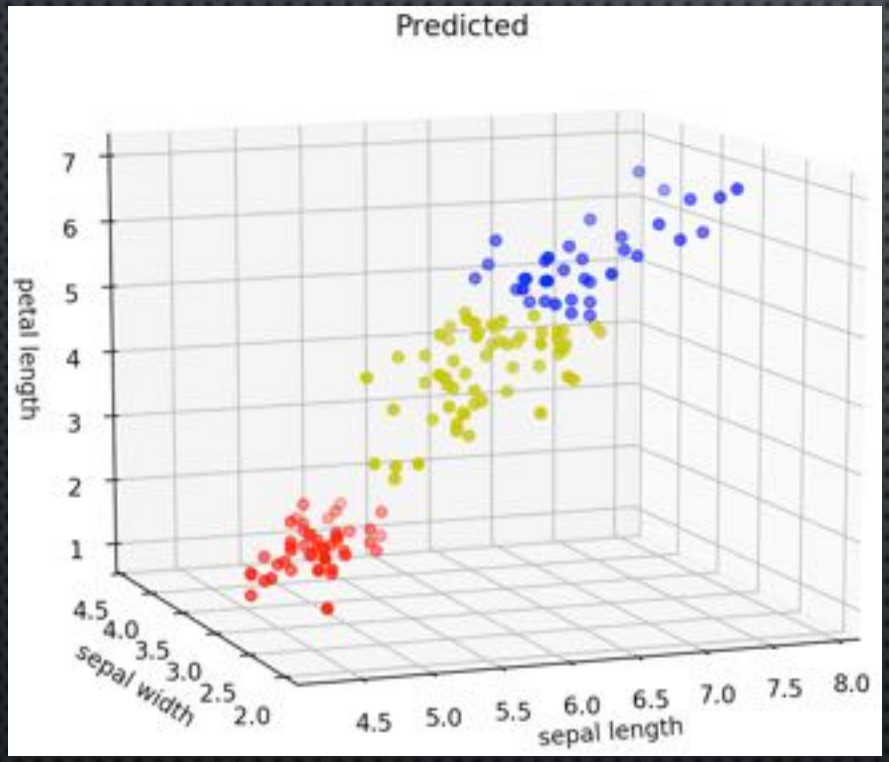
K-means: Iris data

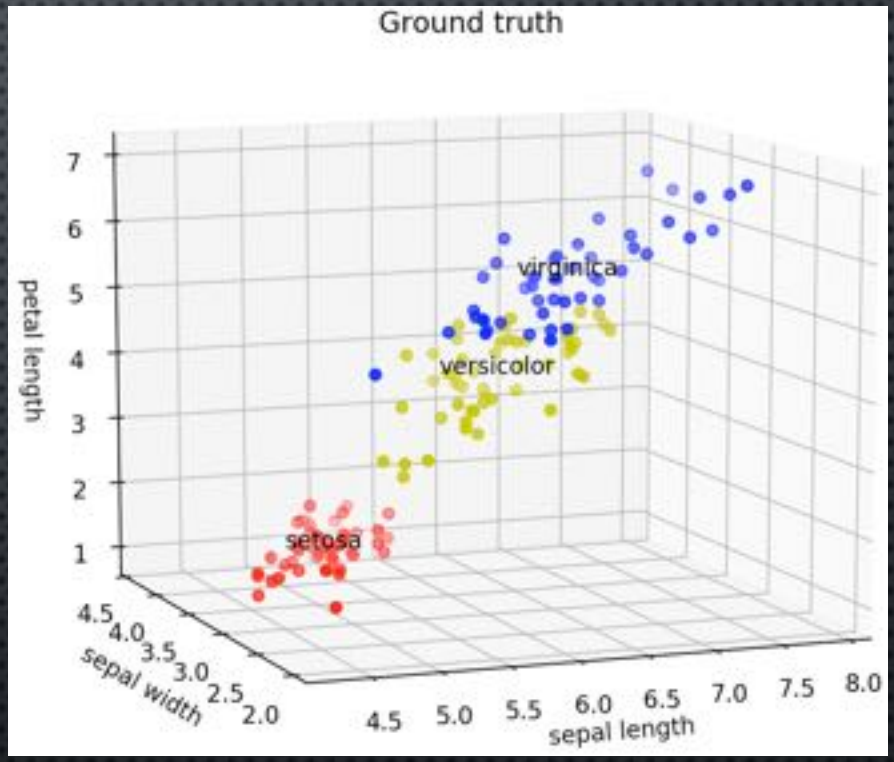
- Data statistics
  - Four variables: length & width of sepal and petal
  - Sample size: 150
- How many type of Iris flowers are in the data?
  - i.e. for k-means, the value of k?





K-means: clusters in Iris data



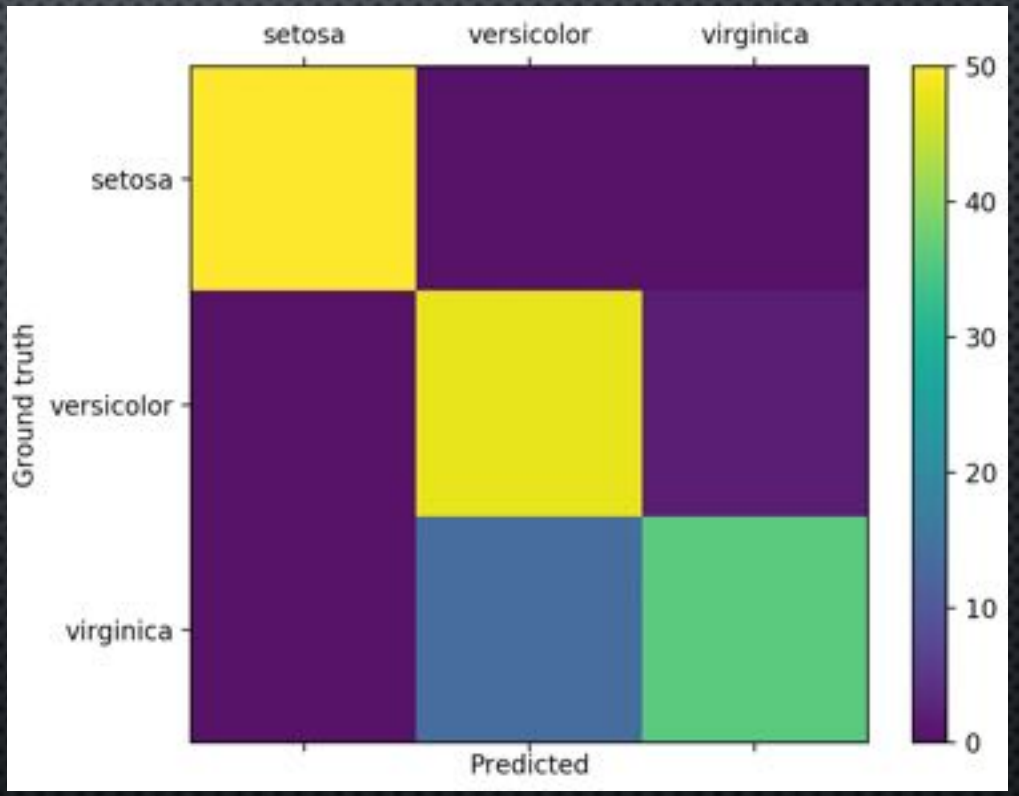


- Performance
  - Overall error rate: 10.7%

K-means: performance on Iris data

- Error rate on each label:
  - "setosa": 50/50
  - "versicolor": 48/50, 2 → "virginica"
  - "virginica": 36/50, 14 → "versicolor"

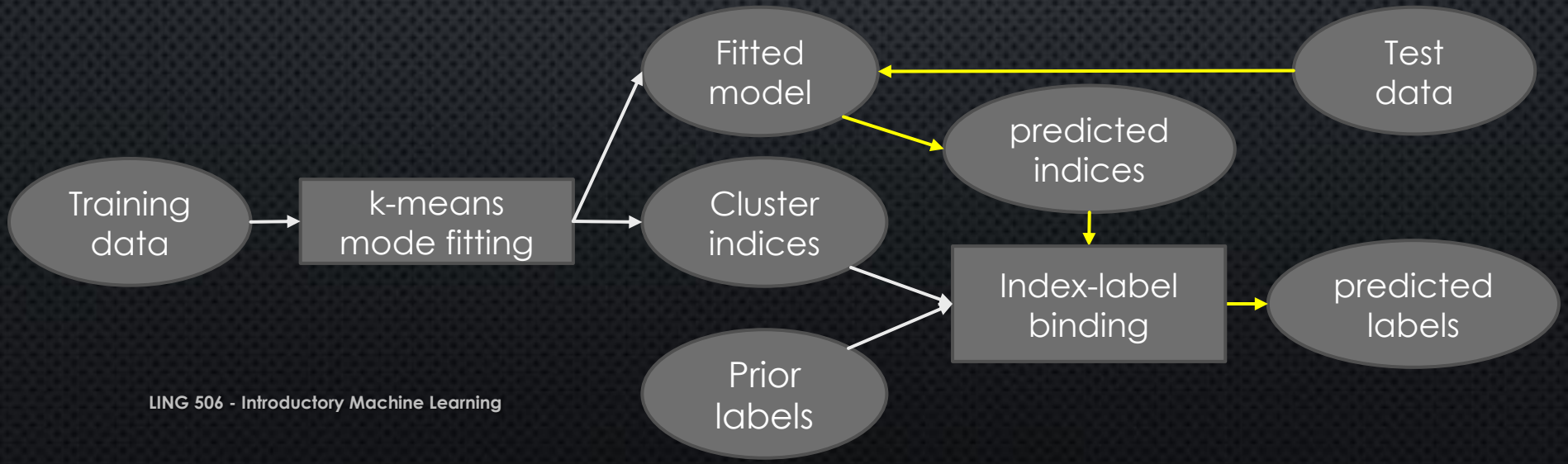
setosa	versicolor	virginica
0.0%	4.0%	18.0%



Confusion matrix of labels

K-means: performance on Iris data

- What if the labels are available for some of the samples, and we want to find the correct label for other samples?
  - Similar to a supervised ML problem
  - But labels used to match the cluster indices after training rather than the feedback during training.



K-means: performance on Iris data

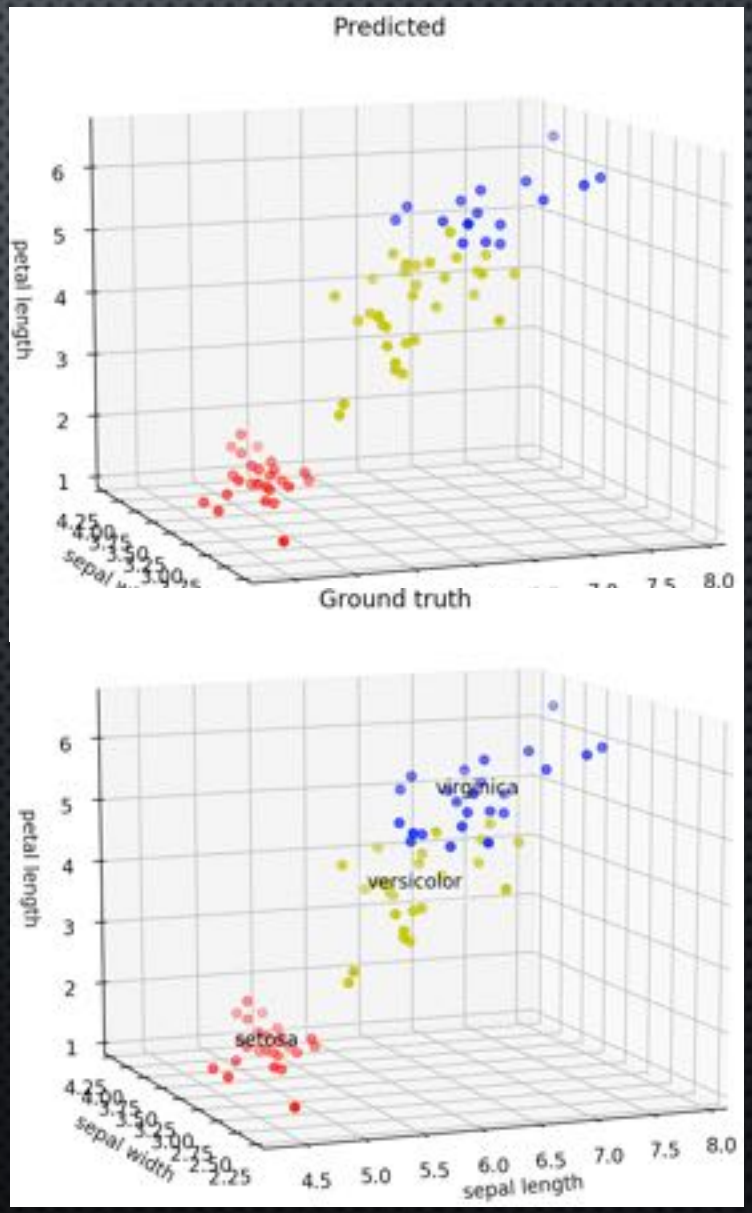
- Data division
  - Training vs testing: 50%:50%, i.e. 75 samples for each
- Error rate
  - Training data: 9.3%
  - Test data: 10.7%

setosa	versicolor	virginica
0.0%	0.0%	32.0%

Training

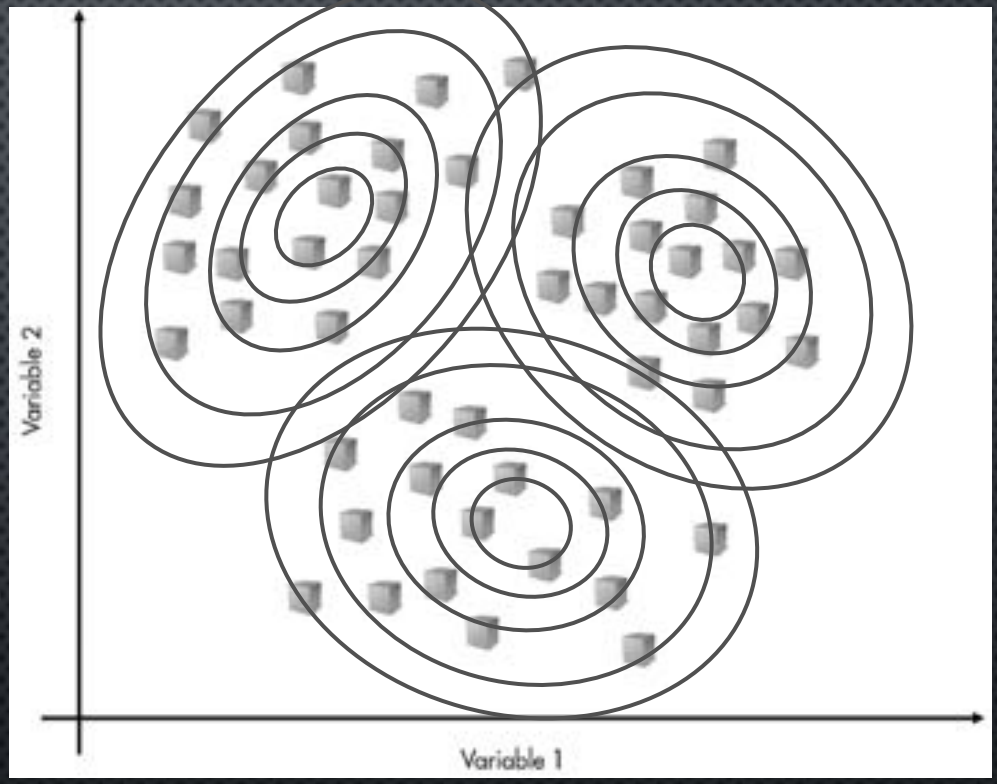
setosa	versicolor	virginica
0.0%	0.0%	28.0%

Test



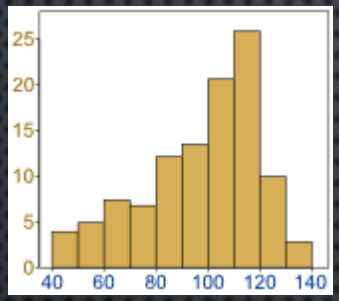
Gaussian Mixture Model (GMM) clustering

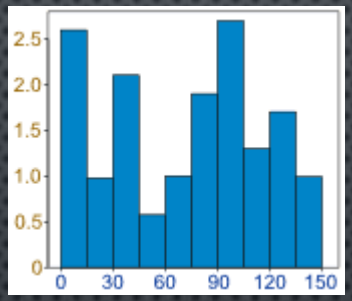
- Fits k n-dimensional normal distributions to the data
- Data is grouped according to which of the Gaussians they most probably belong to
- k must be specified before clustering

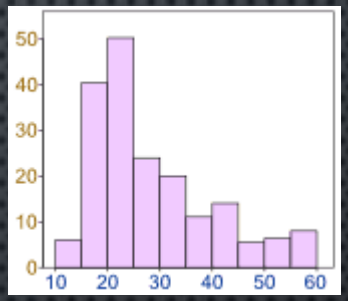


Gaussian/normal distribution

- Types of data distribution:



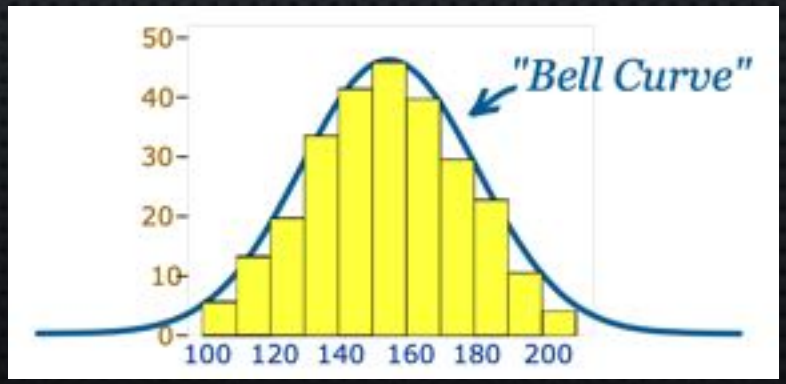




- Gaussian/normal distribution
  - A case how real-valued random data is distributed
  - A continuous probability distribution

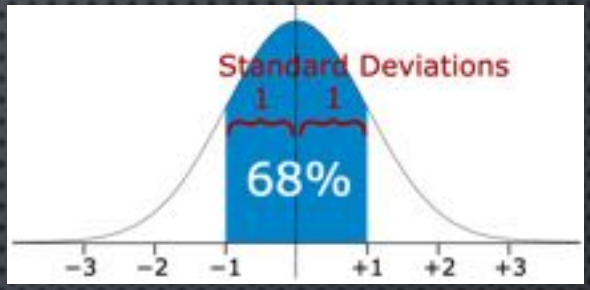
$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}$$

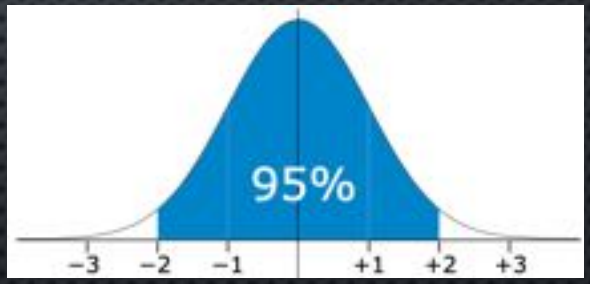
When $\mu = 0$ and $\sigma = 1$, $f(x)$ is standard normal distribution

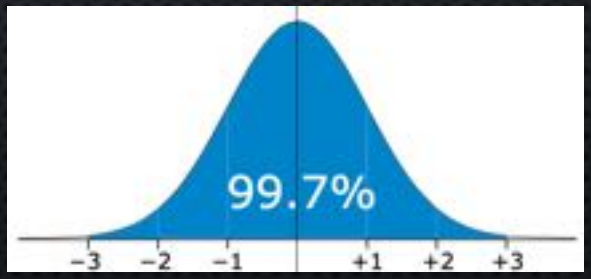


Gaussian/normal distribution: some properties

- Normal distribution in practice
  - Blood pressure, grade of a test, height of human, errors in measurements
- When a set of data is ***strictly*** normally distributed
  - Mean = median = mode
  - Symmetric to the mean, i.e. each half of the data less and greater than the mean



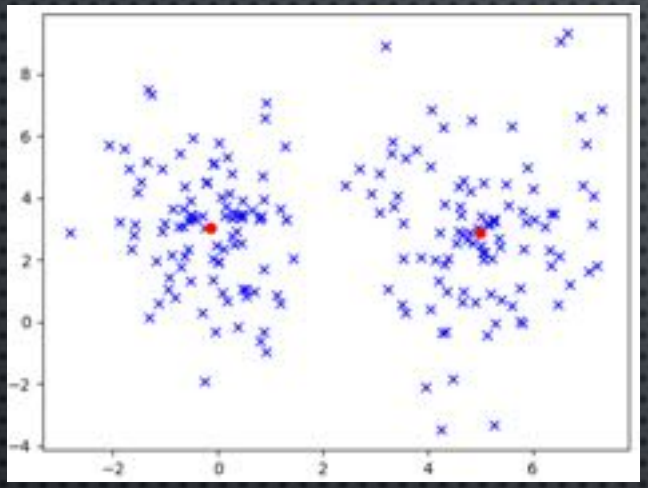


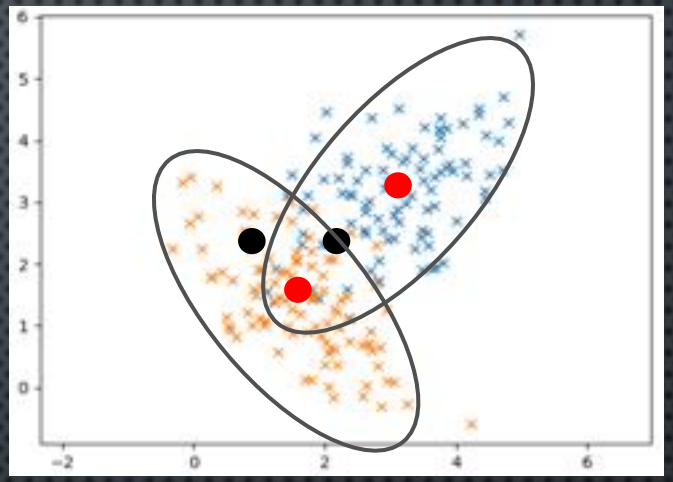


GMM

- A probabilistic model
  - Assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.
- Can be thought of as generalised k-means clustering
  - Incorporates information about the covariance structure of the data and the centres of the latent Gaussians.
  - Allows soft clustering

GMM clustering: principle





$$g(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}$$

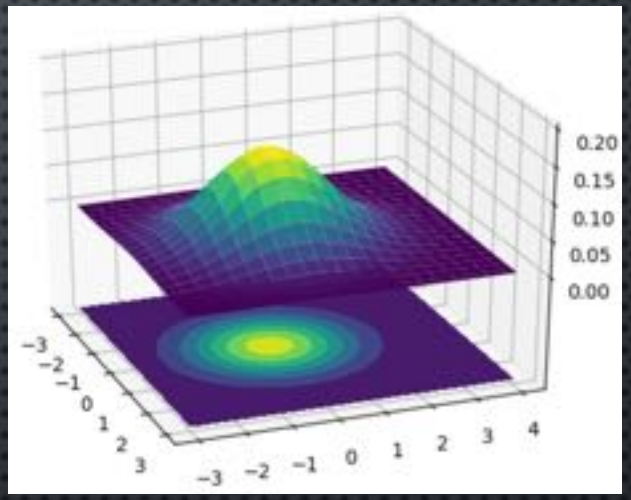
$$p(x) = \sum_{i=1}^{K} w_i g(x)$$

w: mixing coefficient
Σ: covariance matrix
n: number of dimensions

- GMM is to find the parameters (Σ, μ and w) of the Gaussians that best explain each group of data

GMM clustering: Expectation-Maximization (EM)

- EM attempts to determine the optimum values for the parameters in each Gaussian from existing data.
- E-step: calculate the probabilities for each sample belonging to each group – resembles assigning data to a cluster in k-means
- M-step: use the values computed rom E-step to update the parameters (Σ, μ and Π) – resembles updating cluster centroid in k-means
- Repeat EM to maximise the log-likelihood function



GMM clustering: Hard vs Soft clustering

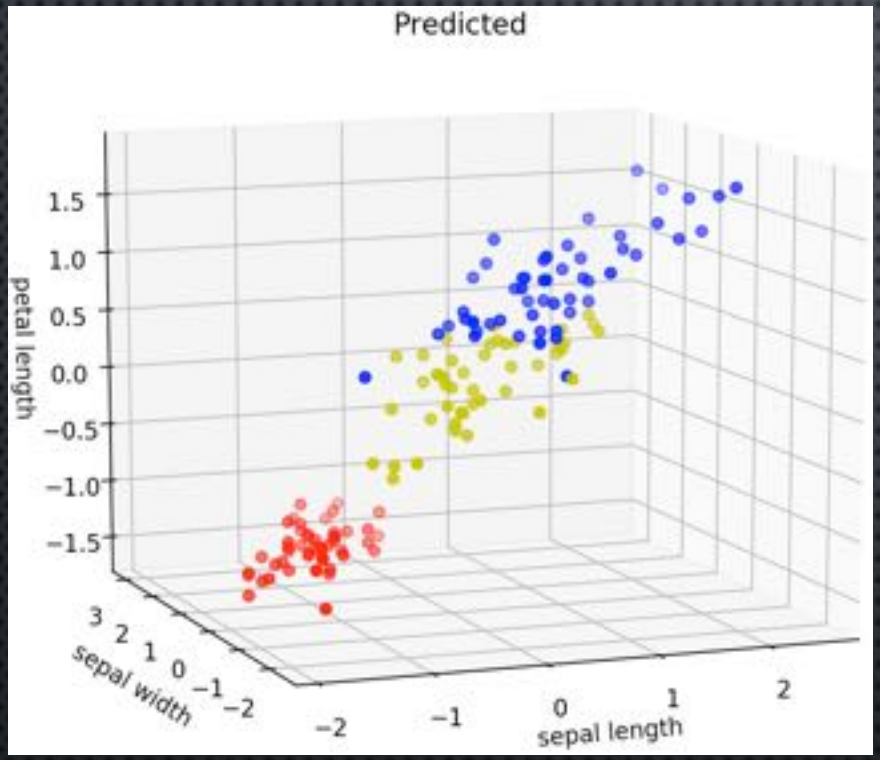
	Cluster 1	Cluster 2	Cluster 3
Individual 1	1	0	0
Individual 2	0	1	0
Individual 3	0	1	0
Individual 4	1	0	0
Individual 5	...	...	...
Individual 6	...	...	...

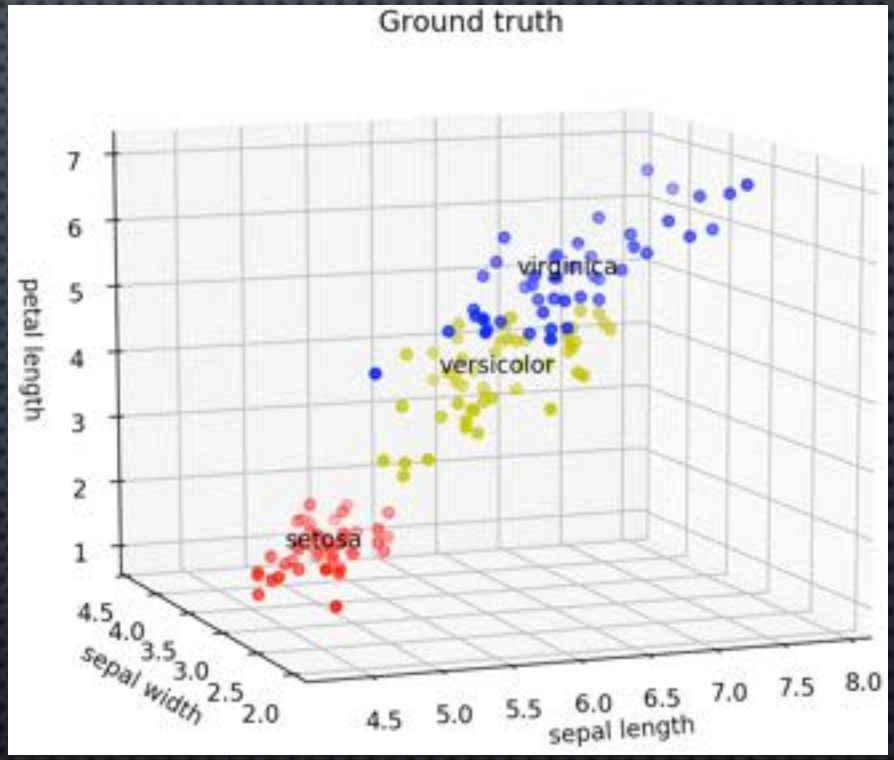
Probability of	Cluster 1	Cluster 2	Cluster 3	Sum
Individual 1	0.1	0.4	0.5	1
Individual 2	0.8	0.1	0.1	1
Individual 3	0.7	0.2	0.1	1
Individual 4	0.10	0.05	0.85	1
Individual 5	...	...	...	1
Individual 6	...	...	...	1

- Each cluster is defined by its centre variable (centroid or mean)
- Each observation can be assigned to more than one cluster with a probability
- Target cluster is determined according to the highest probability

```
[[6.68458864e-044 1.00000000e+000 7.44269825e-035]
 [6.88281903e-031 1.00000000e+000 2.97316198e-028]
 [7.26674155e-036 1.00000000e+000 4.82530460e-030]
 [1.16681891e-031 1.00000000e+000 3.06088222e-026]
 [2.43353714e-046 1.00000000e+000 3.11573125e-035]
 [6.27253973e-045 1.00000000e+000 3.86997452e-035]
 [6.66095706e-036 1.00000000e+000 7.09594114e-029]
 [4.46454786e-040 1.00000000e+000 7.18242595e-032]
 [1.76283223e-027 1.00000000e+000 3.07012240e-024]
 [1.53326187e-035 1.00000000e+000 1.08368342e-028]
 [2.65105565e-049 1.00000000e+000 7.27403367e-038]
 [1.42018409e-038 1.00000000e+000 3.44616063e-029]
 [9.31681897e-034 1.00000000e+000 2.37927764e-028]
```

GMM: clusters in Iris data

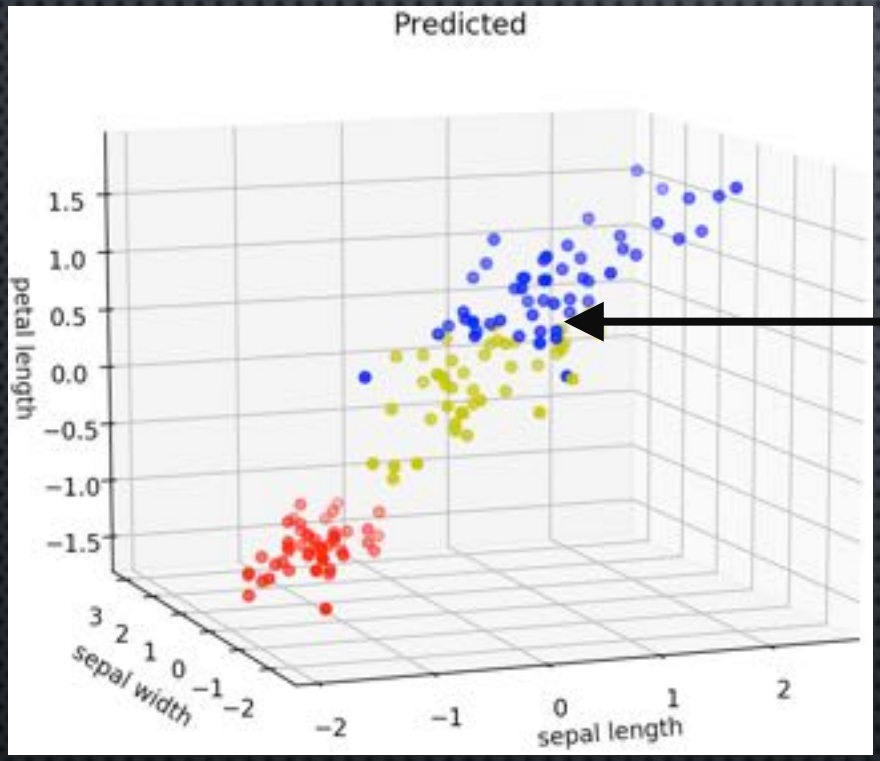




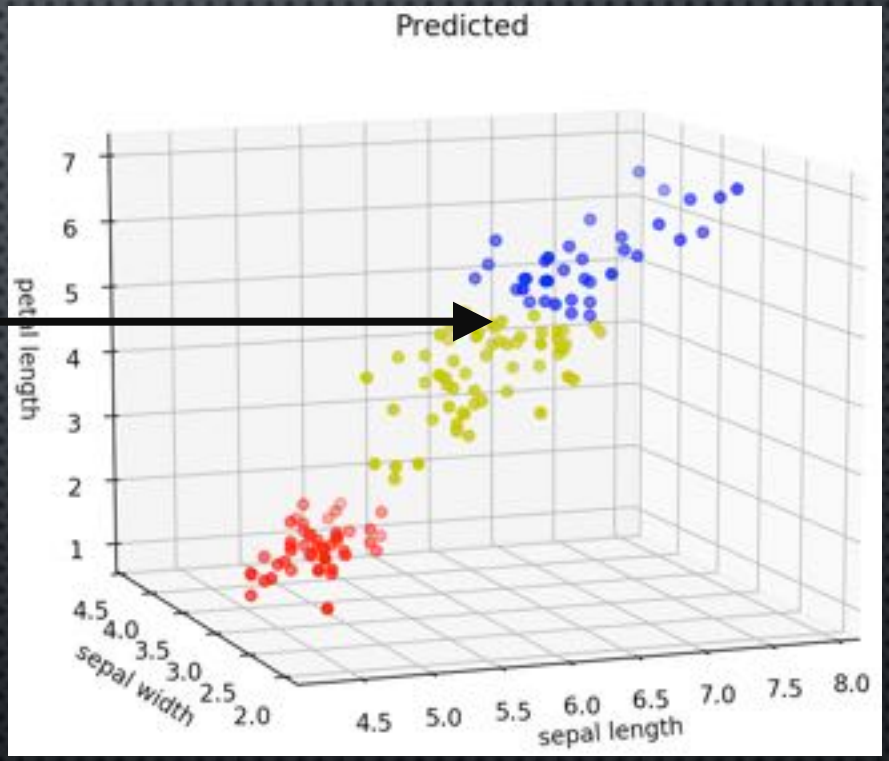
- Performance
  - Overall error rate: 3.3%

setosa	versicolor	virginica
0.0%	10.0%	0.0%

GMM: clusters in Iris data



GMM: **3.3%**



k-means: **10.7%**

Random initialisation

- Performance may be affected by random initialisation
- Some initial points may result in convergence to local minimal/maximal, leading to sub-optimal clustering
- Repeat the computation for multiple times
  - The implementations usually allow to do so
- Use the results from different method as the initial points

K-means vs GMM

	Parameters	Scalability	Use case	D-metric
K-means	No. of cluster	Very large	Similar cluster size, flat geometry, with small k	Distance between observations
GMM	No. of cluster + several, e.g. Σ, μ and Π	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centres

More info at
https://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods