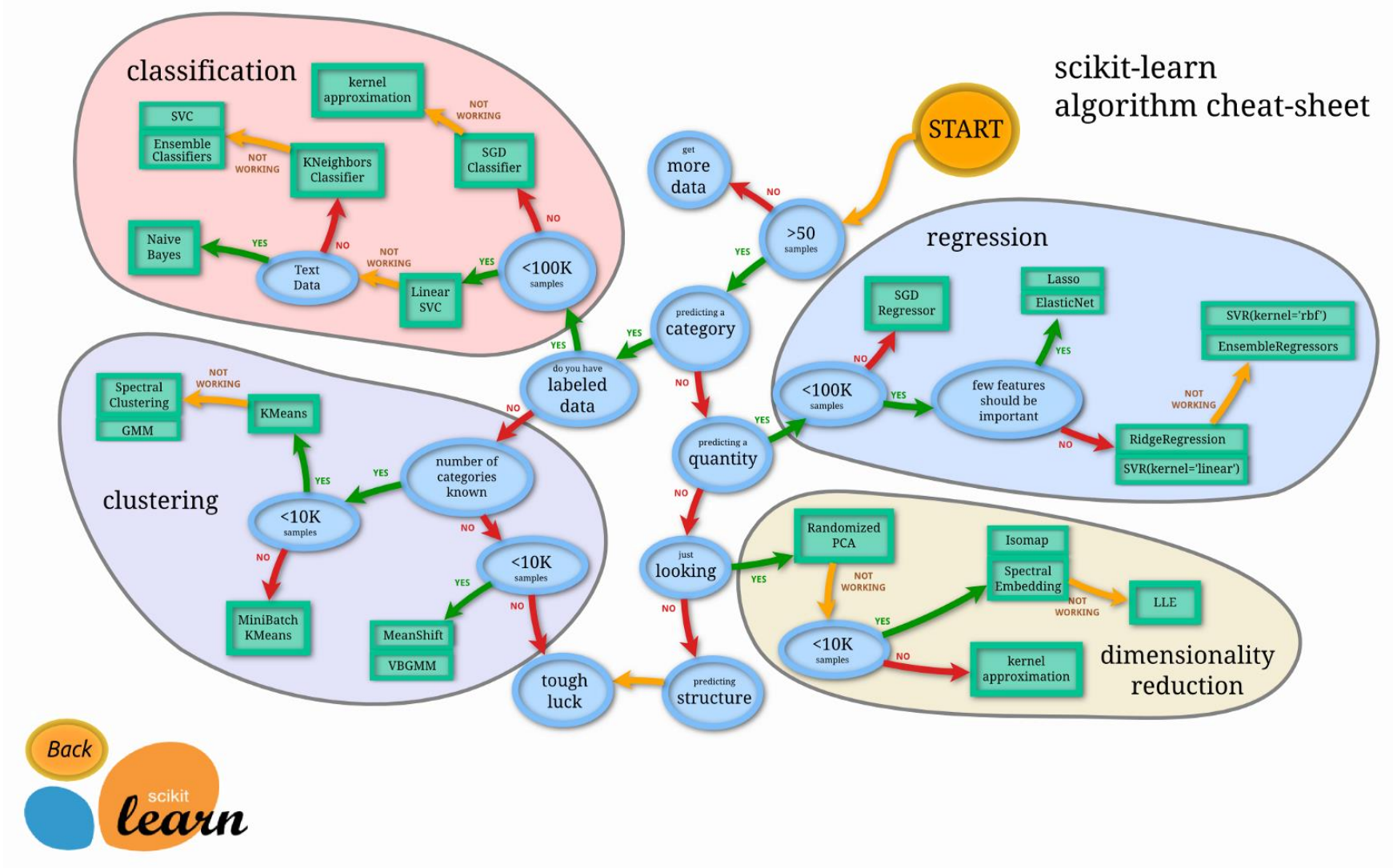
Unsupervised Learning 2

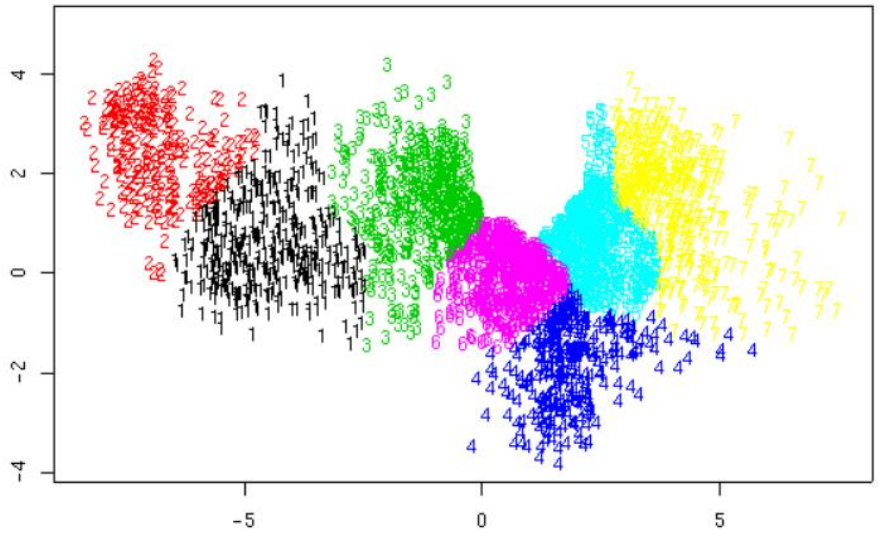
Clustering – Kmeans og Gaussian Mixture Models

Hvor er vi ? Clustering..



Clustering

- Gruppér data i en række “clusters”
- Ofte vælges antal clusters/komponenter
- Bygger (ofte implicit) på afstandsmål
- Evaluering af clusters kan være vanskelig

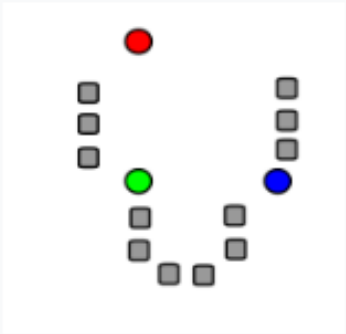


Anvendelser

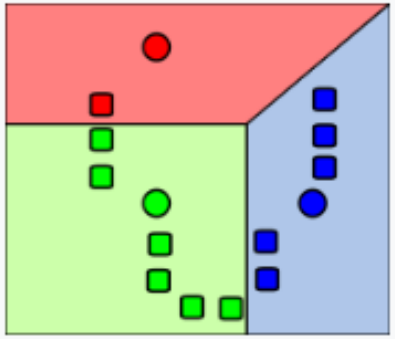
- Data mining / exploration
- Topic discovery
- Novelty/outlier detection
- Codebooks / data diskretisering
- Visualisering

K-means

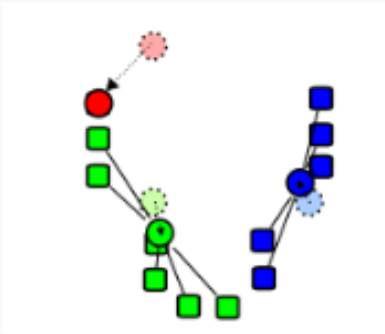
Demonstration of the standard algorithm



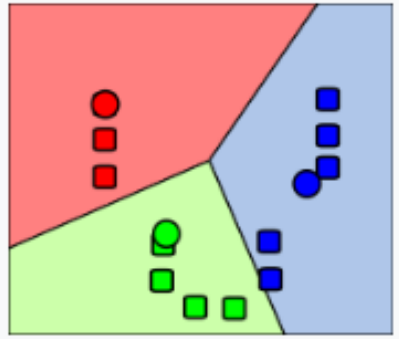
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



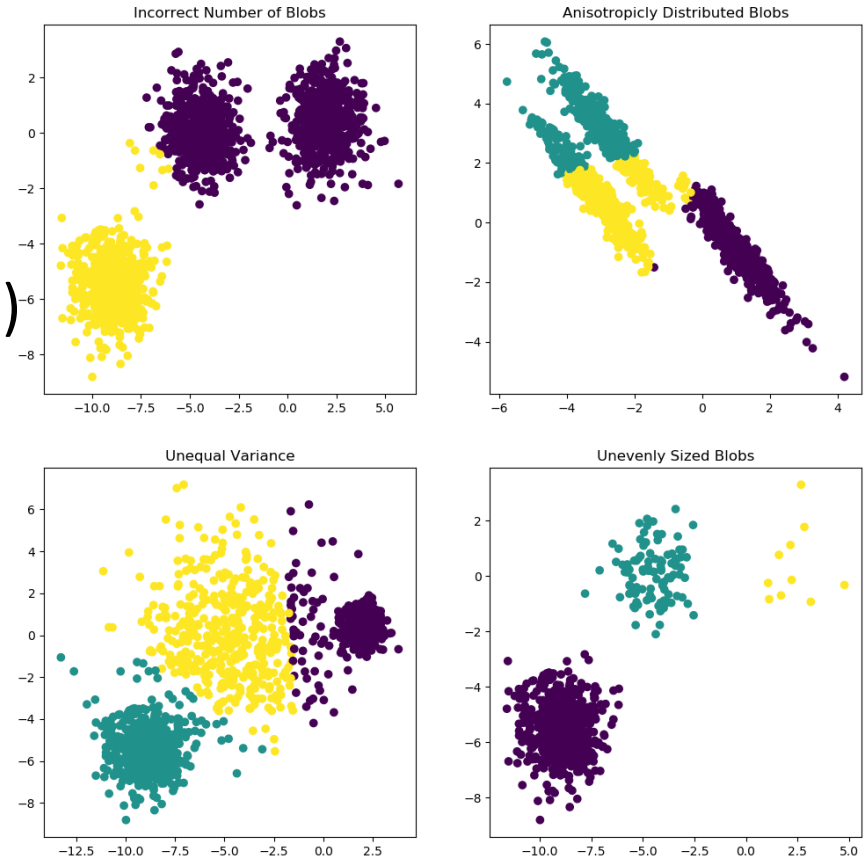
3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

K-means – problemer

- Indirekte antagelser og karakteristika :
  - K kendt
  - Spherical covariance matricer
  - 0/1 cluster relationship (hard assignment)



Gaussian Mixture Models (GMM)

- Probabilistisk alternativ til Kmeans til clustering
- Giver sandsynlighed for tilhørsforhold til cluster
- Modellerer clusters med fulde kovarians matricer
- Men – stadigvæk antages antal clusters kendt (kan dog benytte log-likelihood på test sæt til bestemmelse af optimal K)

GMM

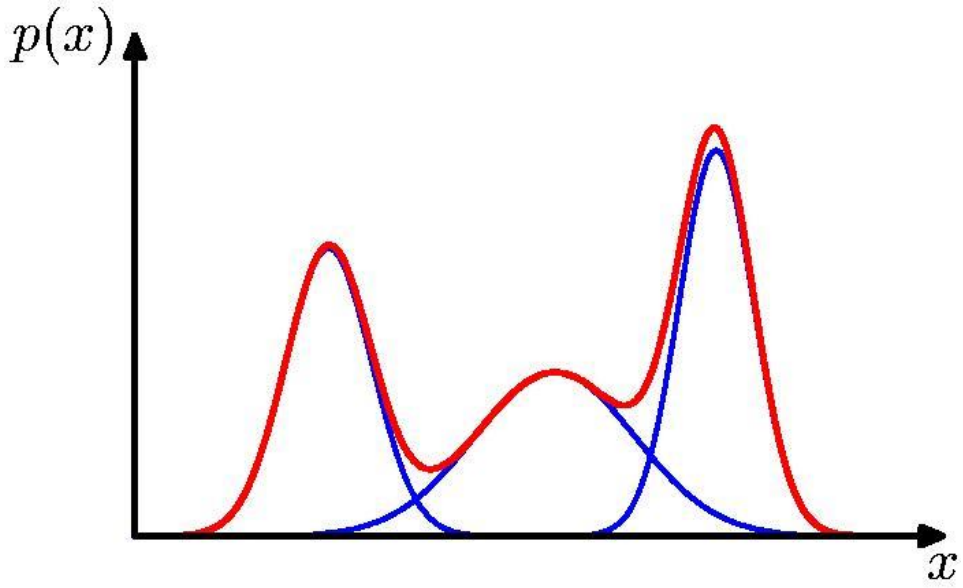
Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

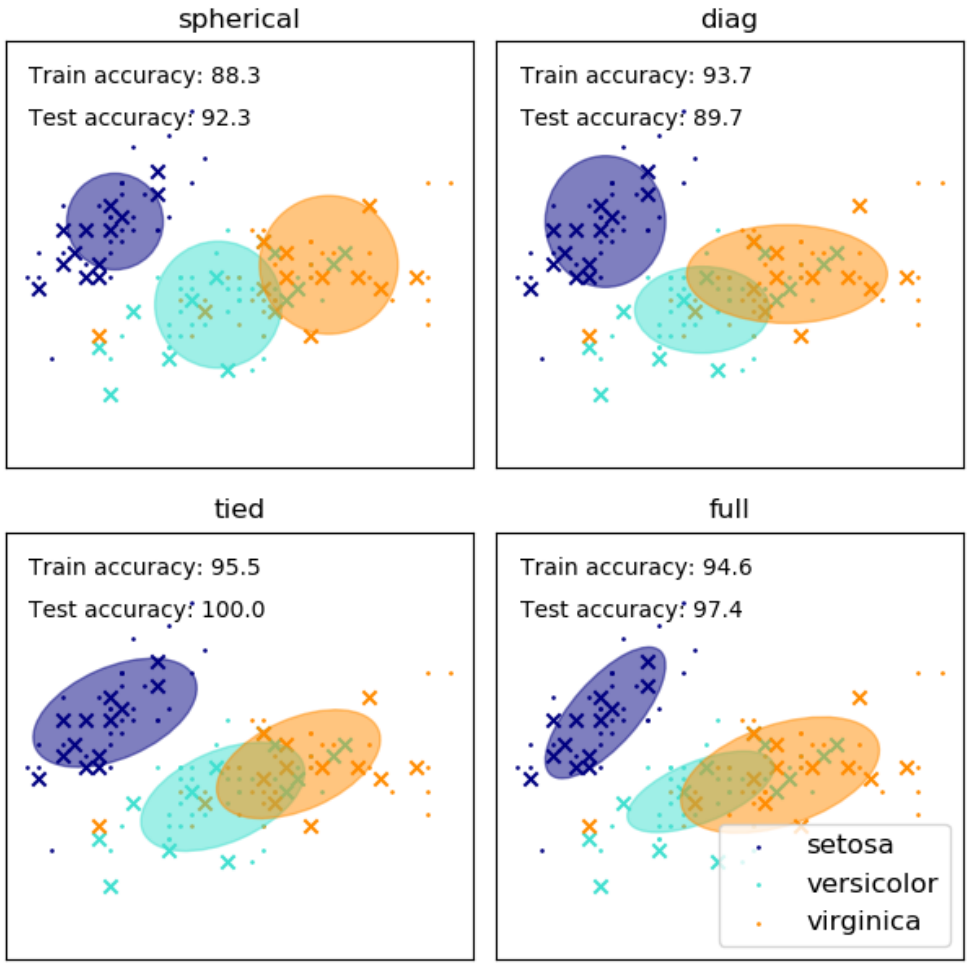
Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

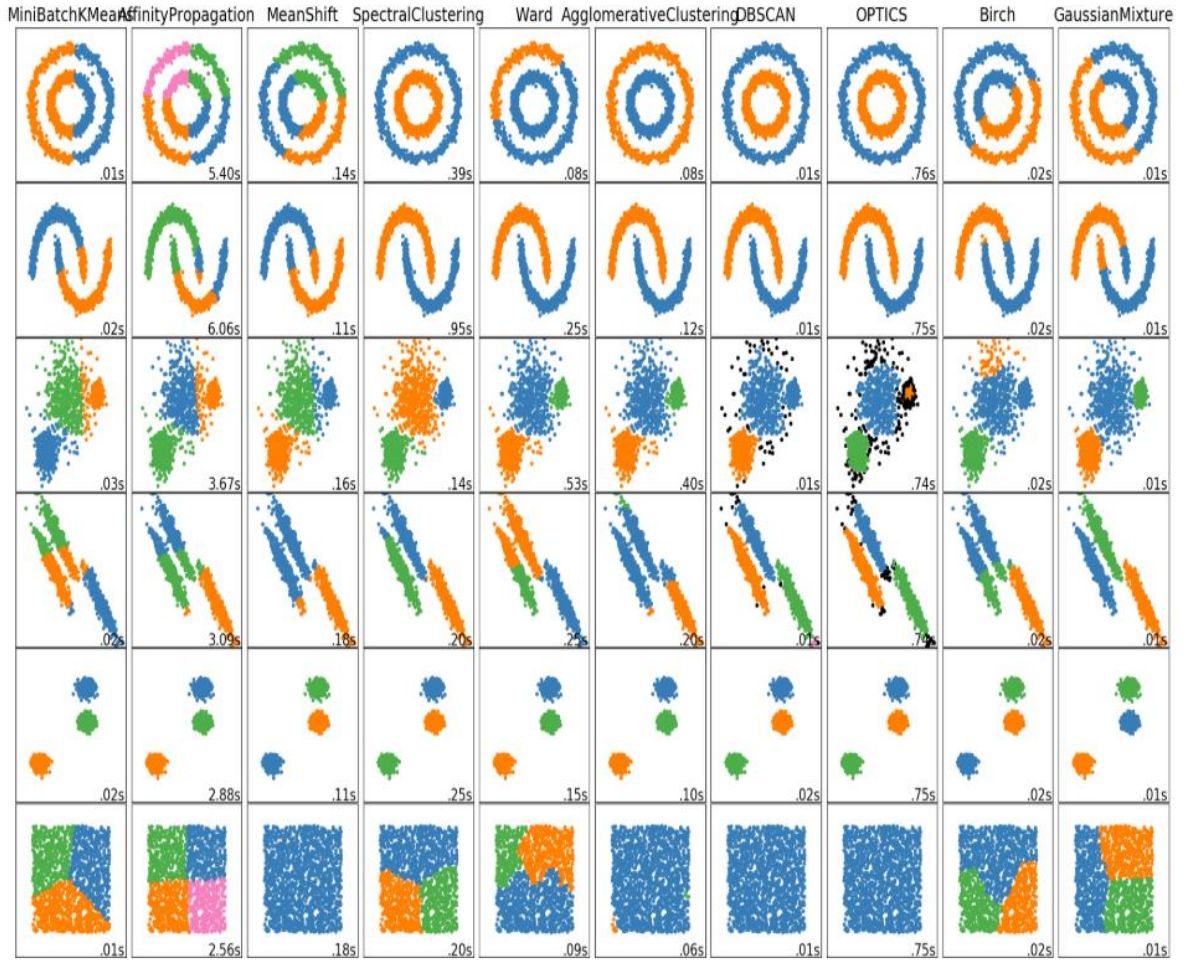


GMM covariance matrices



Many others..

- Hierarchical clustering
- Spectral clustering
- …



A comparison of the clustering algorithms in scikit-learn