

# ITMAL Øvelser – Kmeans og GMM

## Øvelse 1 :

I denne øvelse skal I eksperimentere med K-means algoritmen. I skal arbejde med datasættet "Handwritten digits" fra Scikit Learn, som er scannede billeder af håndskrevne tal imellem 0 og 9 (dvs. 10 kategorier). I skal forestille jer at I ikke kender de sande værdier (dvs. at I ikke ved hvilke cifre der er på billederne – altså, et unsupervised learning problem).

- Kør eksemplet i linket nedenunder.
- Modificer eksemplet, så I plotter hvilke både sande værdi (cifre) og cluster-nummer. Hvor godt stemmer de overens ?
- Prøv at plotte et par eksempler, som er kommet i en "forkert" cluster – skyldes forvekslingen at cifrene ligner hinanden ?
- Prøv med flere clusters – giver det en bedre inddeling ?
- Optional : Prøv at anvende K-means algoritmen efter PCA til en højere dimension end 2 (fx 10 eller 100). Måske får I en bedre clustering her, hvor I ikke har smidt så meget information væk ?

Tips :

Demonstration af kmeans på handwritten digits - [https://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py](https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py)

## Øvelse 2 :

I skal nu anvende Gaussian Mixture Model på samme data som i øvelse 1 – dvs. PCA-projicerede handwritten digits i 2D. Find gerne inspiration i nedenstående link.

- Fit GMM modellen til jeres data.

I skal nu afprøve om metoden kan benyttes til at finde outliers / anomalies. Dette gøres ved at sætte en nedre grænse for sandsynligheden (vi vil ikke acceptere usandsynlige værdier), hvorunder samples regnes for at være outliers / anomalies.

- Sæt en threshold-værdi på jeres sandsynlighed (Gaussian\_mixture.score\_samples som udregner  $\log(P(x))$  for givet punkt  $x$ ) og undersøg om de fundne outliers reelt også er outliers (fx kig på 2-5 outliers)

Tips :

Eksempel med density estimation med GMM - [https://scikit-learn.org/stable/auto\\_examples/mixture/plot\\_gmm\\_pdf.html#sphx-glr-auto-examples-mixture-plot-gmm-pdf-py](https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_pdf.html#sphx-glr-auto-examples-mixture-plot-gmm-pdf-py)

### **Øvelse 3 :**

Afprøv Kmeans og GMM på jeres egne data.