Gaussian Mixture Model

Leonie Rumi

January 24, 2023

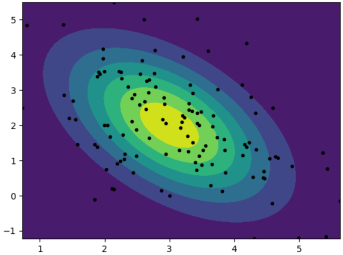
GMM Recap

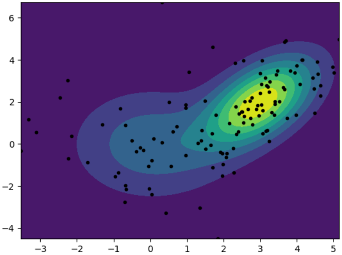
- Probabilistic Clustering
- Goal: Assign a probability to each data point for belonging to each cluster
- Assumes data is generated by a multivariate mixture model

Multivariate Mixture Model

Multivariate normal distribution
$X \sim N(\mu, \Sigma)$

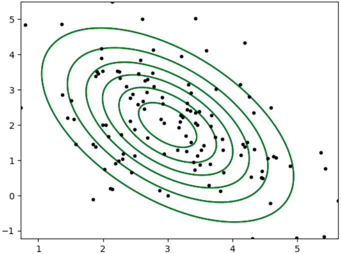
Multivariate mixture model
$X \sim \sum_i \pi_i \cdot N_i(\mu_i, \Sigma_i)$



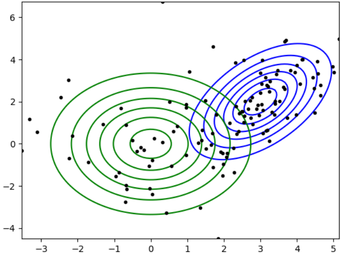


Multivariate Mixture Model

Multivariate normal distribution
$X \sim N(\mu, \Sigma)$



Multivariate mixture model
$X \sim \sum_i \pi_i \cdot N_i(\mu_i, \Sigma_i)$



GMM - Algorithm

Initialize parameters

GMM - Algorithm

Initialize parameters

Repeat:

Estimate posterior probabilities $p_{u,i}$ with given μ_i and Σ_i

$$p_{u,i} = \frac{\pi_i \cdot \Phi(x_u|\mu_i, \Sigma_i)}{\sum_k \pi_k \cdot \Phi(x_u|\mu_k, \Sigma_k}$$

GMM - Algorithm

Initialize parameters

Repeat:

Estimate posterior probabilities $p_{u,i}$ with given μ_i and Σ_i

$$p_{u,i} = \frac{\pi_i \cdot \Phi(x_u|\mu_i, \Sigma_i)}{\sum_k \pi_k \cdot \Phi(x_u|\mu_k, \Sigma_k}$$

Maximize parameters with given probabilities $p_{u,i}$

$$\pi_i = \frac{\sum_u p_{u,i}}{n}$$

$$\mu_i = \frac{\sum_u p_{u,i} \cdot x_u}{\sum_u p_{u,i}}$$

$$\Sigma_i = \frac{\sum_u p_{u,i} \cdot (x_u - \mu_i)(x_u - \mu_i)^T}{\sum_u p_{u,i}}$$

Project

Idea: Cluster days in the year 2022 according to weather in Florence
$\rightarrow$ Discover patterns throughout the year

1. Collect data
2. Preprocess
3. Perform clustering using GMM
4. Validate

Data

Data set *Visual Crossing*

- Global weather data over more than 50 years
- Accumulated data per day for a location
- Variables like minimum / maximum temperature, precipitation, humidity

Choose variable to use

Data: Weather in Florence in 2022

Choose variables to use:

- Minimum temperature
- Maximum temperature
- Humidity
- Precipitation
- Precipitation cover
- Cloud cover
- Solar radiation

Results

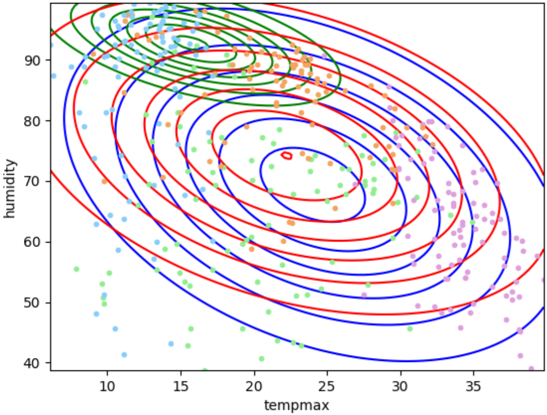
- Learned multivariate distribution for k clusters that cover the data
- In general: not a clear separability in data

Results

- Learned multivariate distribution for k clusters that cover the data
- In general: not a clear separability in data

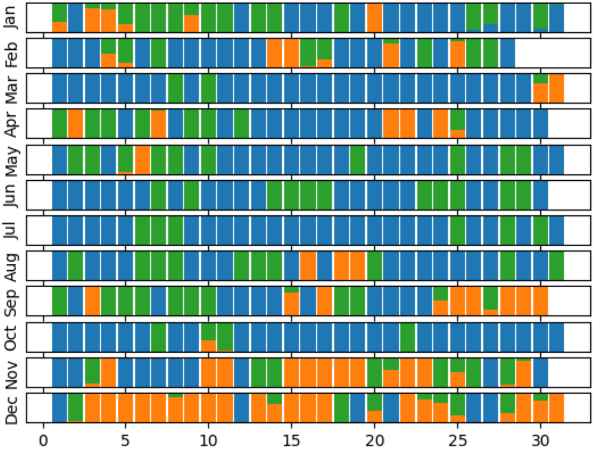
E.g. $k = 3$ distributions on variables maximum temperature and humidity

Displaying seasons in colors: **spring**, **summer**, **fall**, **winter**



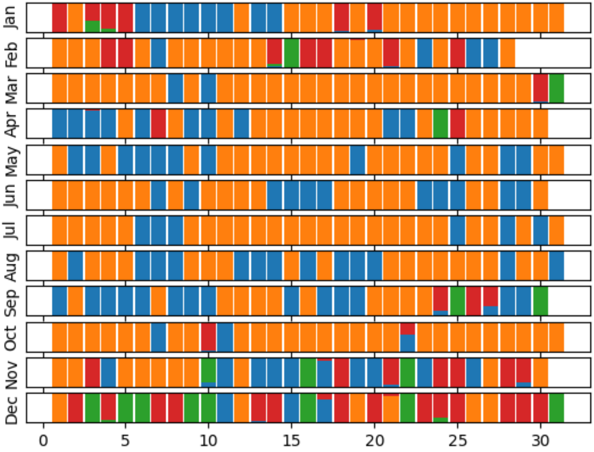
Results cont'd

Probabilities that each day belongs to each of the $k = 3$ clusters



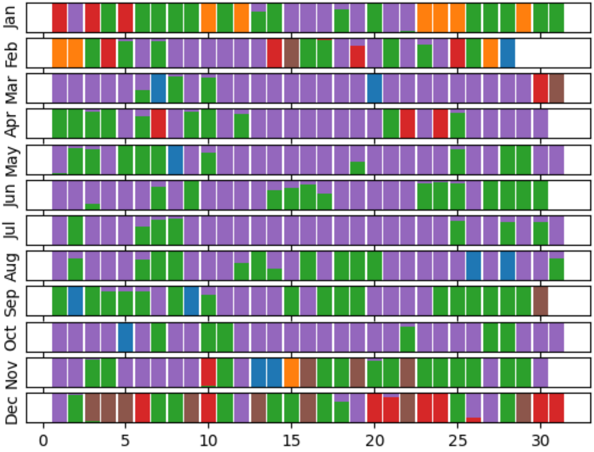
Using different k

Probabilities that each day belongs to each of the $k = 4$ clusters



Using different k

Probabilities that each day belongs to each of the $k = 6$ clusters



Variation of the algorithm

Using the result of 15 iterations K-Means to initialize parameters Results:

- Reduced amount of iterations, faster convergence
- Mostly only slight changes in clusters

Next steps

- Findings:
  - Similar weather conditions in summer
  - More variation in november, december and january
- Use PCA on input data to reduce run time
- Compare clustering of different years to discover development of weather across the years

Thank you for your attention!

Sources

- Data set: https://www.visualcrossing.com/
- Lecture slides Clustering