Workshop 12

COMP90051 Machine Learning

Semester 2, 2020

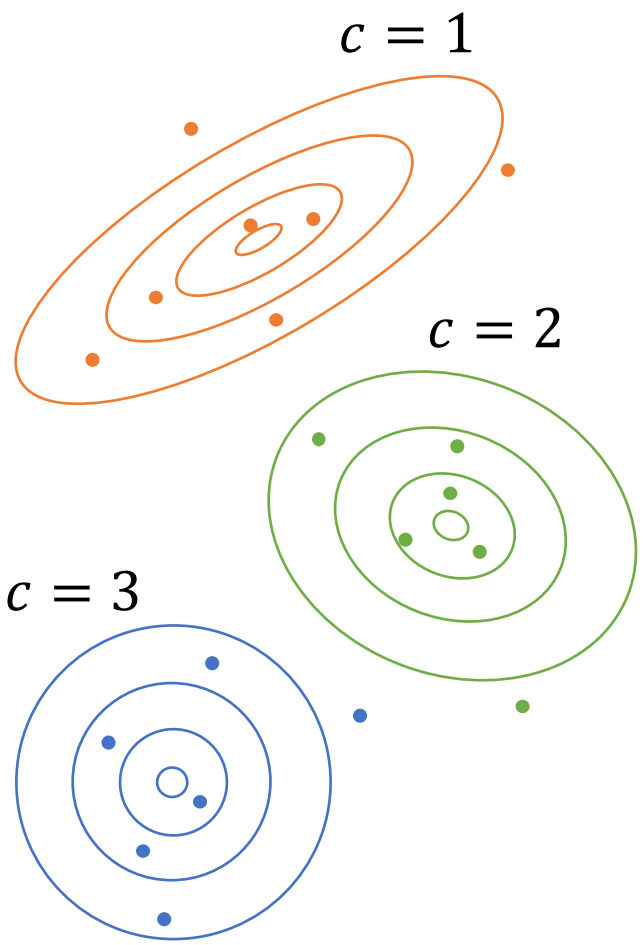
Learning Outcomes

By the end of this workshop you should be able to:

1. generate data from a GMM
2. fit GMMs using scikit-learn
3. select an appropriate value for the number of mixture components using a model selection criterion

Gaussian Mixture Model

- A probabilistic model for clustering data in $\mathbb{R}^m$
- Associated with each cluster $c \in \{1, \ldots, k\}$ is a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$
- Assume each data point $\mathbf{x}_i$ is generated by:
  1. Assigning to a cluster:
     $$z_i \sim \text{Categorical}(\mathbf{w})$$
  2. Drawing from the Gaussian distribution associated with cluster z_i:
     $$\mathbf{x}_i | z_i, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} \sim \text{Normal}\left(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)$$



Inference

- Observe $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, but don't observe $\mathbf{Z} = (z_1, \ldots, z_n)$ or parameters $\theta = (w_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots)$
- MLE estimate for θ determined by maximising the marginal likelihood:

$$L(\theta; \mathbf{X}) = p_\theta(\mathbf{X}) = \int p_\theta(\mathbf{X}, \mathbf{Z}) \, \mathrm{d}\mathbf{Z}$$

- Typically solve optimisation problem using the *expectation-maximisation (EM) algorithm*
- No guarantee of convergence to global optimum

Worksheet 12