Workshop 9

COMP90051 Machine Learning

Semester 2, 2018

Learning Outcomes

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

1. apply cross validation/information theoretic approaches to choose the optimal number of clusters for a GMM
2. generate data from a GMM
3. fit GMMs in scikit-learn

Slides

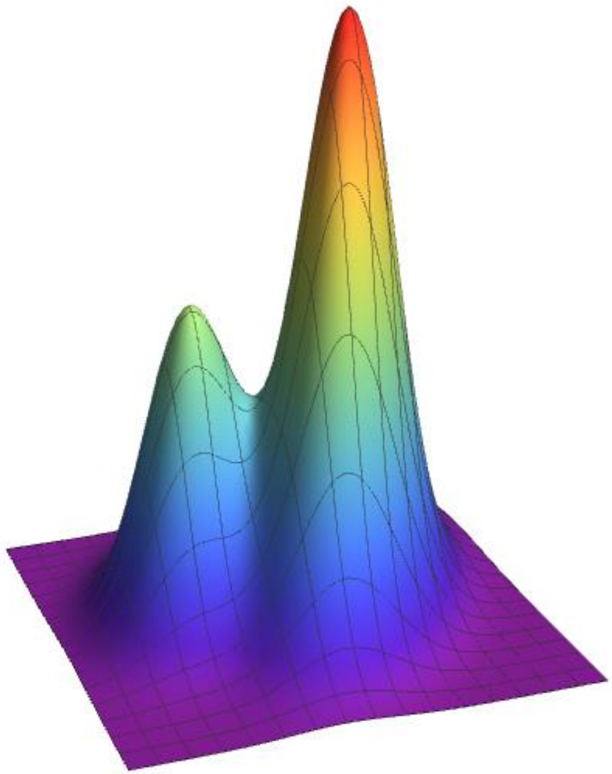
Worksheet 9

Gaussian mixture model

- Data set $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ without labels
- GMM assumes each $\mathbf{x}_i \in \mathbb{R}^m$ is drawn i.i.d. from

$$\sum_{c=1}^{k} w_c \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

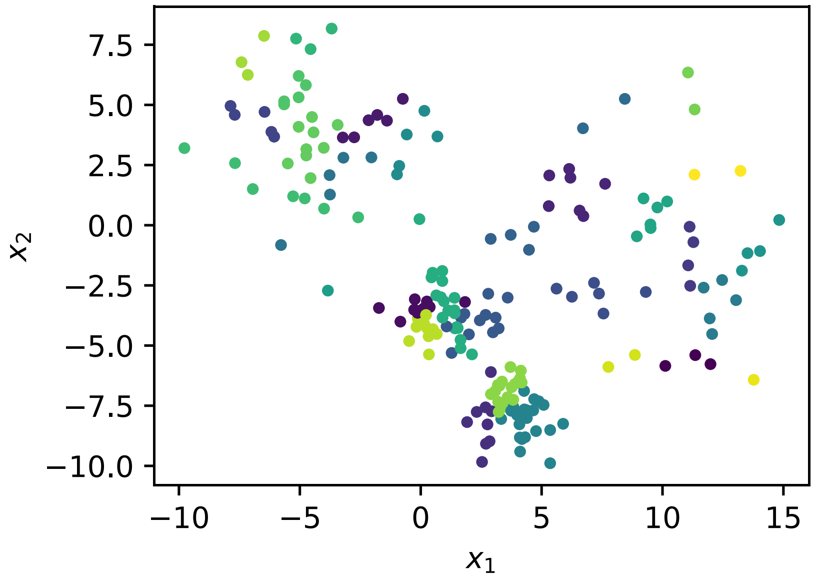
- EM algorithm allows us to find $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$, w_c $\forall c$ that maximise the likelihood

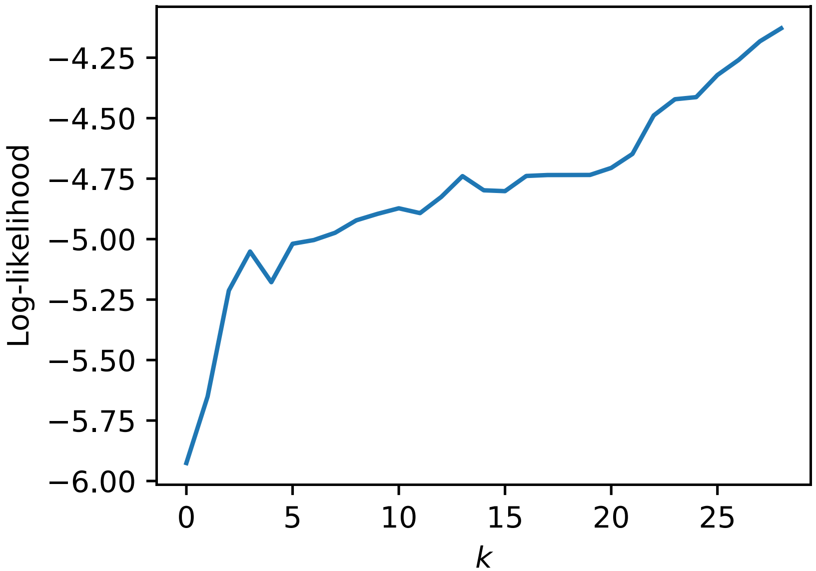


Assumption: k is known

Selecting k

- Why not treat k as a parameter to be optimised?
- No, not a good idea
- Larger k ⇒ more flexible model ⇒ overfitting





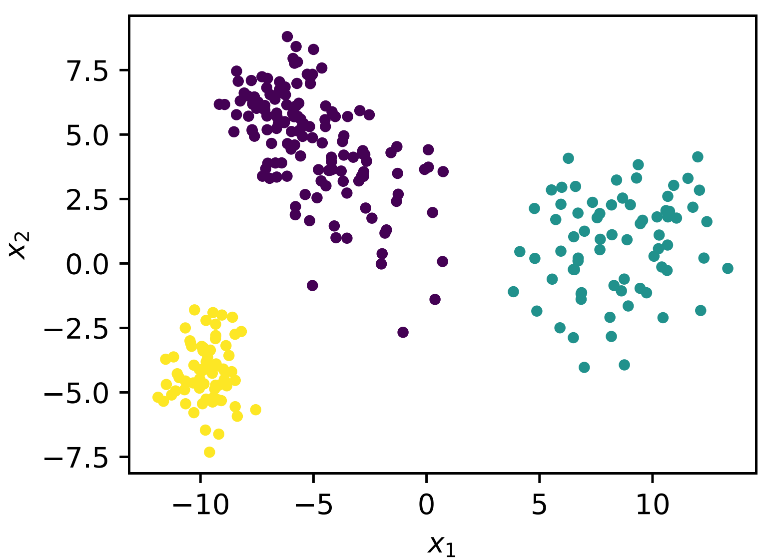
Approaches for selecting k

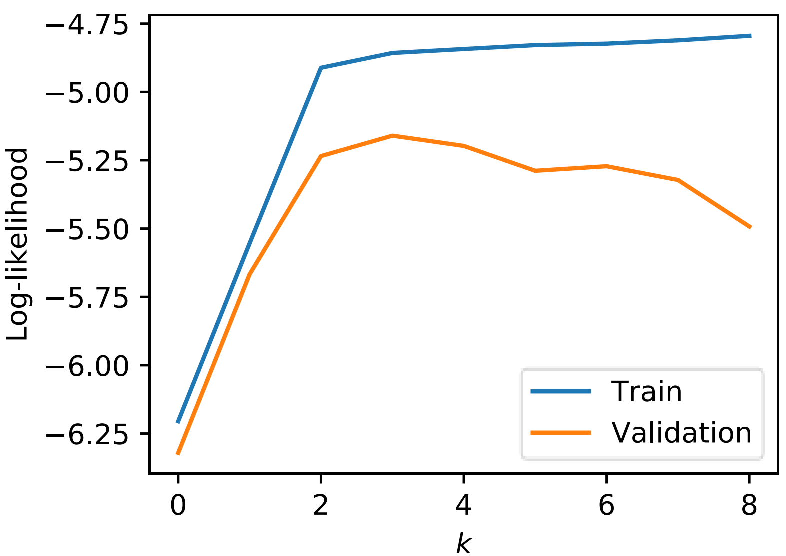
- Sometimes k is known from context
  - e.g. clustering genetic profiles from cells with a known number of cell types
- Less principled:
  - Subjective choice based on visualisation (may need dimensionality reduction)
  - Try plausible values and check whether the results vary
- More principled:
  - Cross-validation
  - Information theory

  Covered today
  - Kink method
  - Non-parametric models

Cross-validation

- Evaluate goodness of fit using the log-likelihood
- Fit a GMM on the training set for a range of k, then compute the log-likelihood on training/validation sets
- Expect to see validation log-likelihood plateau, then drop, beyond the "optimal" k





Akaike information criterion (AIC)

Let N_{par} be the number of independent parameters in a model and L^* be the maximum value of the likelihood function. The Akaike information criterion is defined as

$$\text{AIC} = 2N_{\text{par}} - 2\ln L^*$$

- Used generally for model selection: smaller is better
- Information theoretic interpretation: estimates (relative) information lost in approximating the true model by proposed model.
- Trade-off between model complexity (first term) and goodness of fit (second term)

Akaike information criterion (AIC)

- AIC estimator is only valid asymptotically—when the number of instances n is large.
- For small n, should use a corrected AIC (correction depends on the model). For univariate linear models:

$$\mathrm{AICc} = 2N_{\mathrm{par}} + \ln L^* + \frac{2N_{\mathrm{par}}\left(N_{\mathrm{par}} + 1\right)}{n - N_{\mathrm{par}} - 1}$$

Bayesian information criterion

Let N_{par} be the number of independent parameters in a model and L^* be the maximum value of the likelihood function evaluated on a sample of size n. The Bayesian information criterion is defined as

$$\text{BIC} = N_{\text{par}} \ln n - 2 \ln L^*$$

- Similar to AIC, but can be motivated by a Bayesian argument
- Approximately maximises $p(\text{model}|\text{data})$, independent of prior over models
- In practice, BIC tends to underfit, whereas AIC tends to overfit

Applying AIC and BIC to GMMs

- Fit a GMM on the data for a range of k
- Compute AIC/BIC (depends on maximum likelihood for optimal parameters)
- Choose the model with the smallest AIC/BIC