Clustering Model Selection

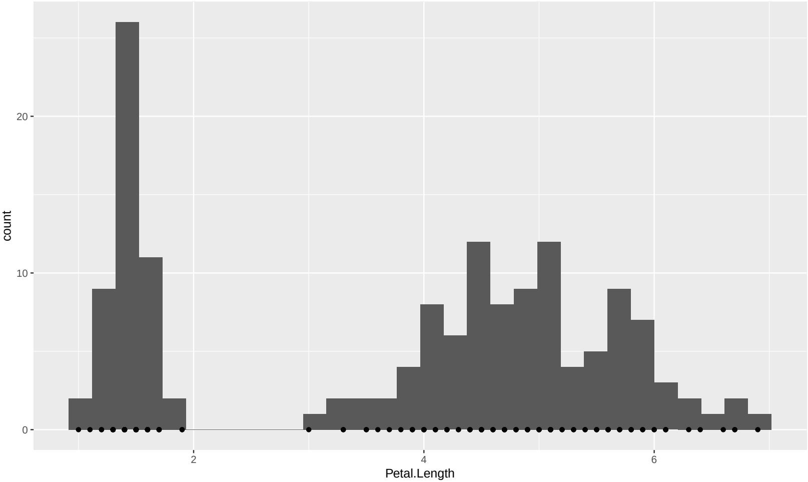
Toby Dylan Hocking

Clustering framework

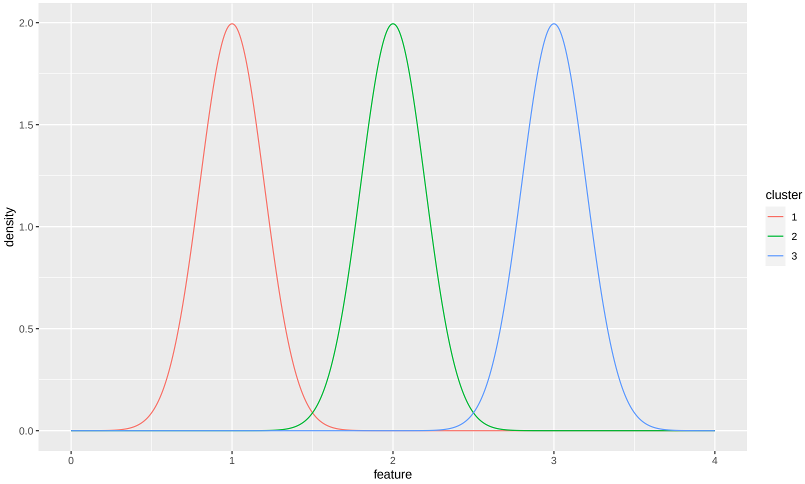
- Let $X = [x_1 \cdots x_n]^\intercal \in \mathbb{R}^{n \times p}$ be the data matrix (input for clustering), where $x_i \in \mathbb{R}^p$ is the input vector for observation i.
- Example iris $n = 150$ observations, $p = 4$ dimensions.
- Consider only one of those columns,

```
##      Petal.Length
## [1,]          1.4
## [2,]          1.4
## [3,]          1.3
## [4,]          1.5
## [5,]          1.4
## [6,]          1.7
```

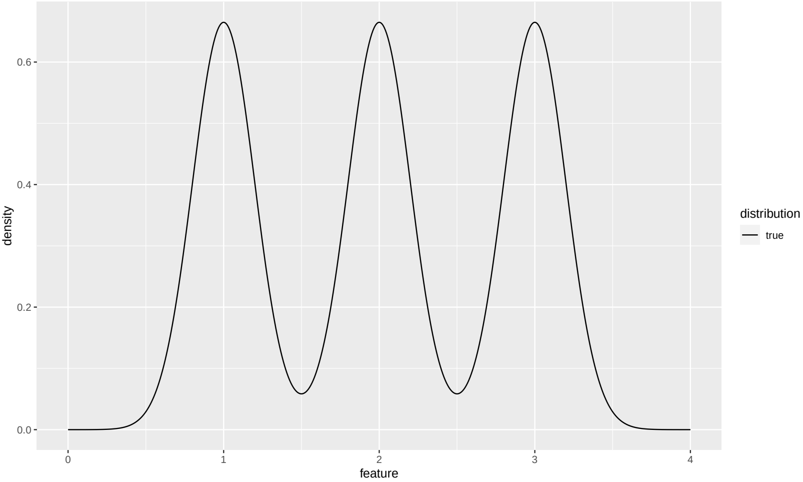
One column can be visualized as a histogram



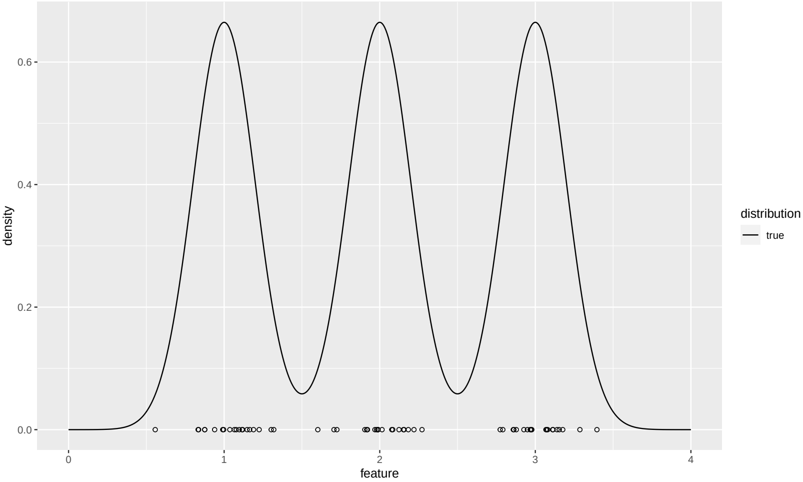
Simulation: three normal densities



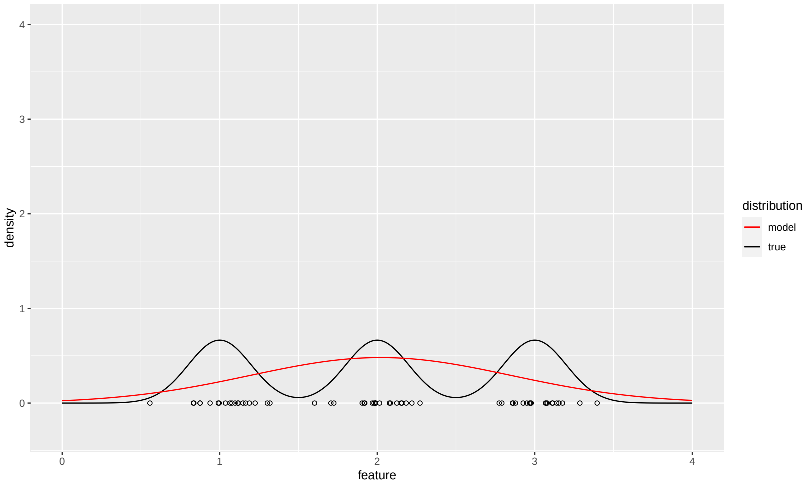
Mixture density



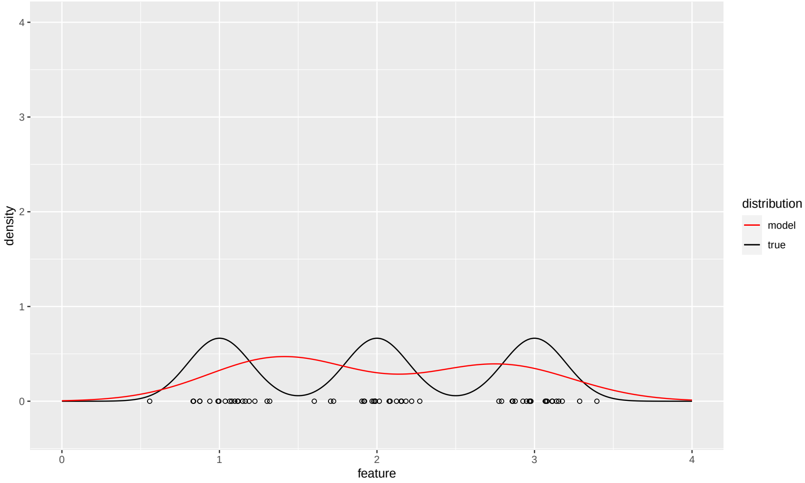
Generate 20 random data from each density



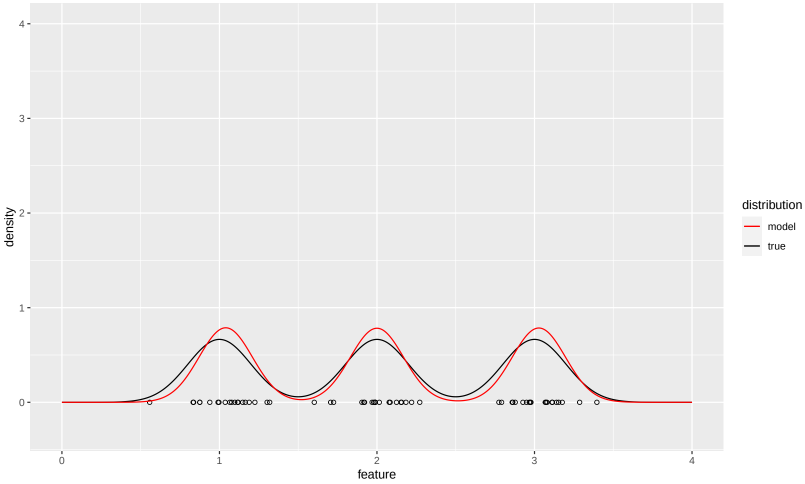
Fit gaussian mixture model 1



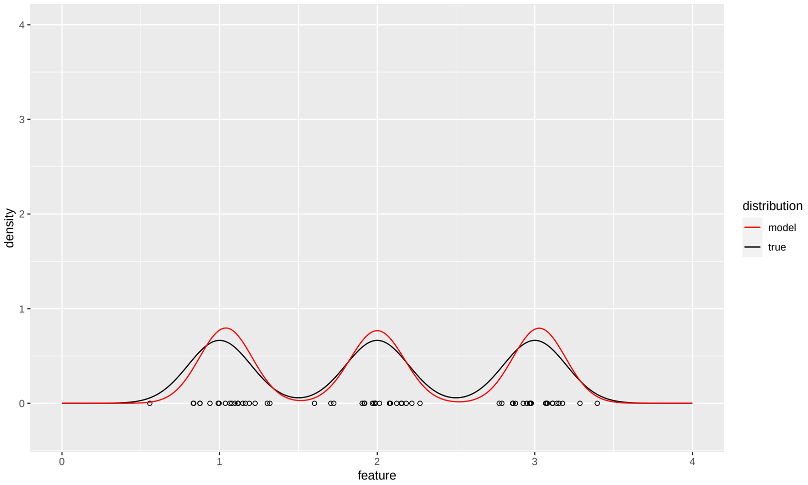
Fit gaussian mixture model 2



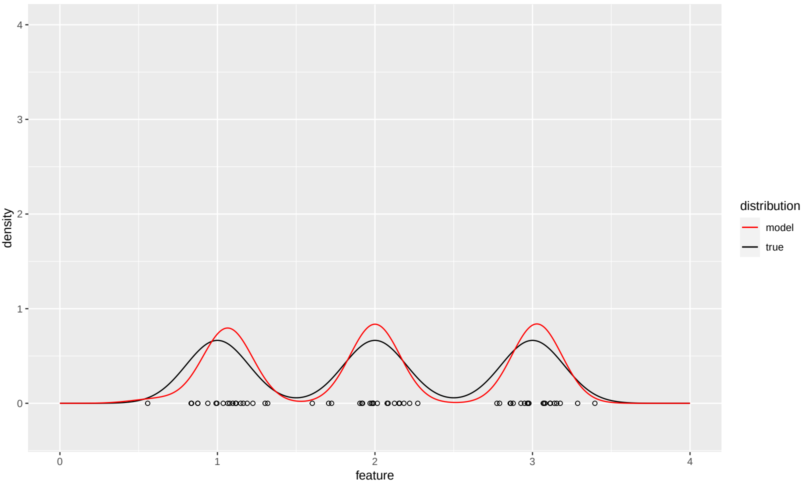
Fit gaussian mixture model 3



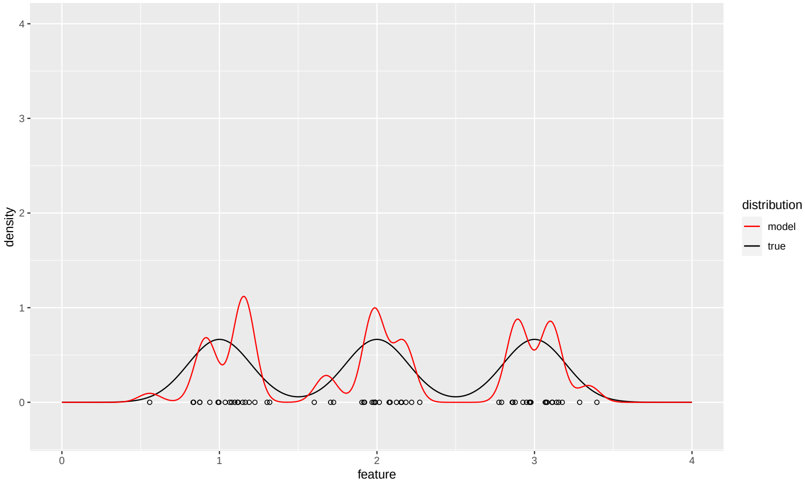
Fit gaussian mixture model 4



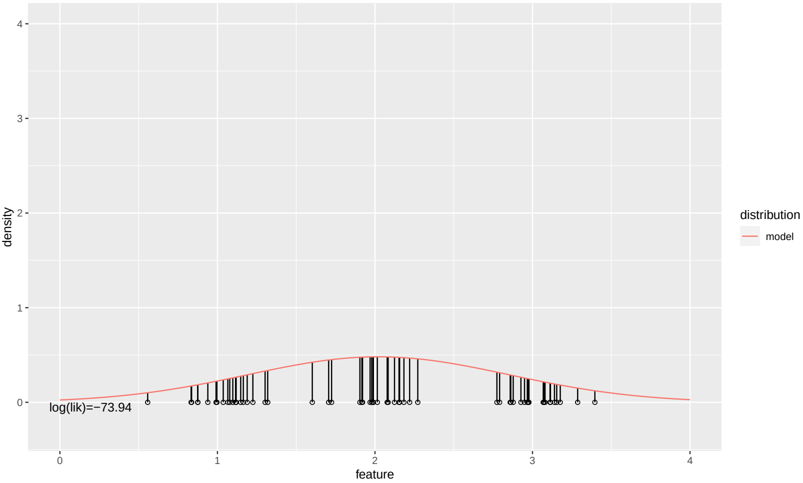
Fit gaussian mixture model 5



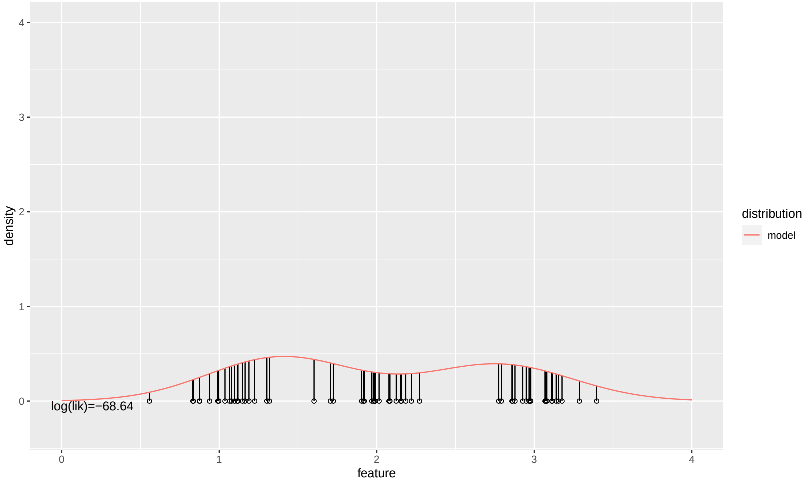
Fit gaussian mixture model 10



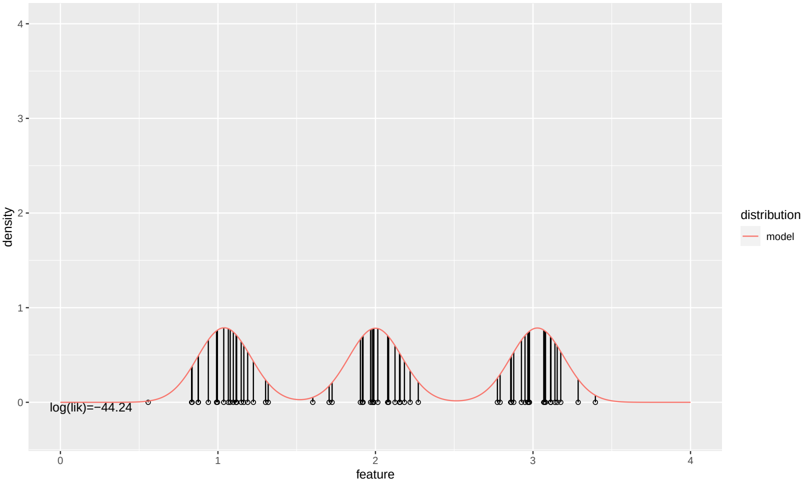
Fit gaussian mixture model 1



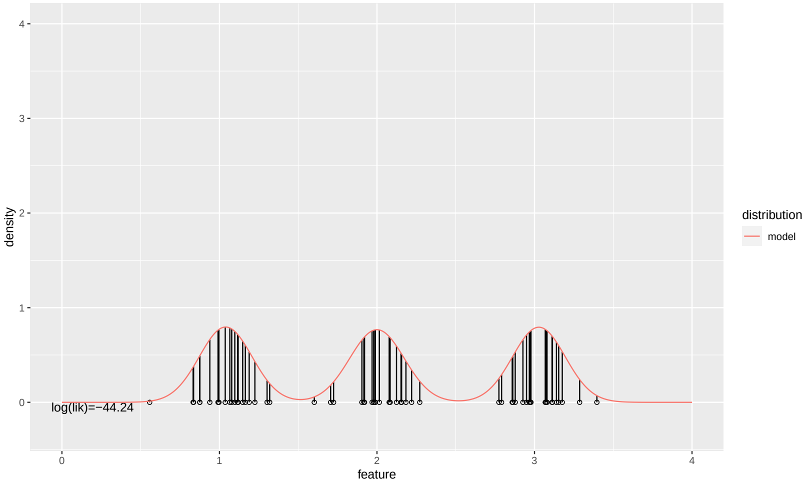
Fit gaussian mixture model 2



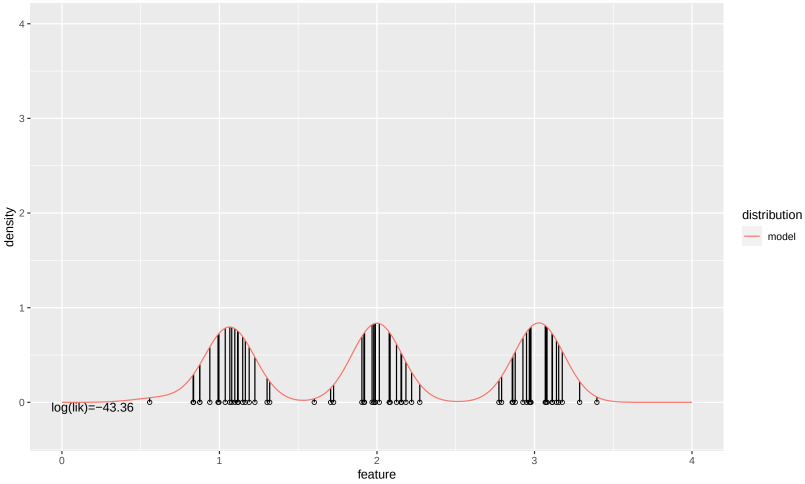
Fit gaussian mixture model 3



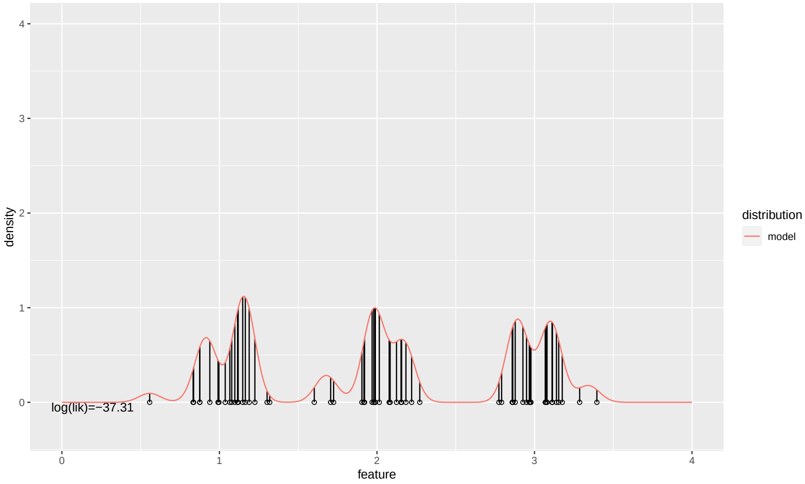
Fit gaussian mixture model 4



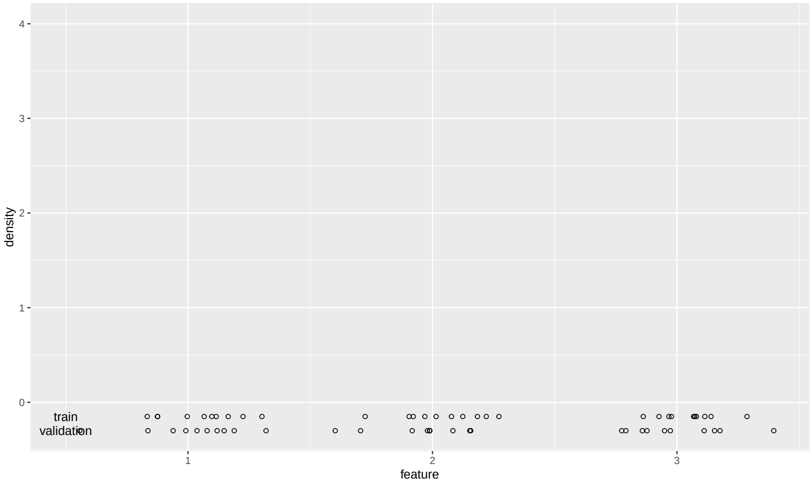
Fit gaussian mixture model 5



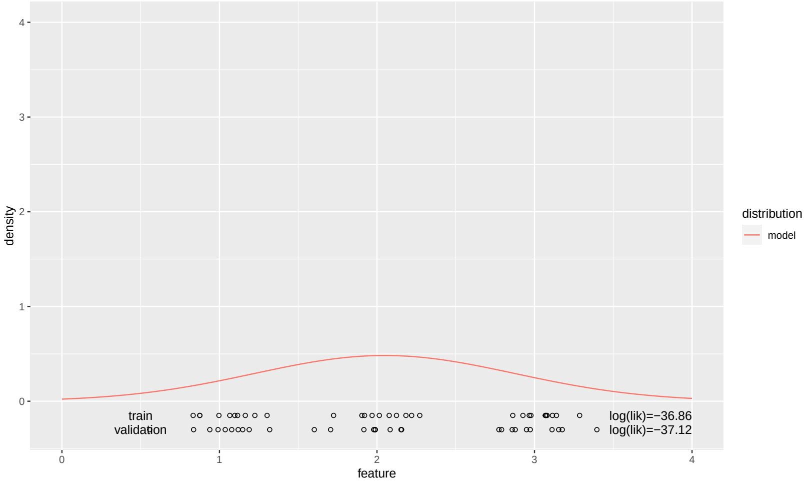
Fit gaussian mixture model 10



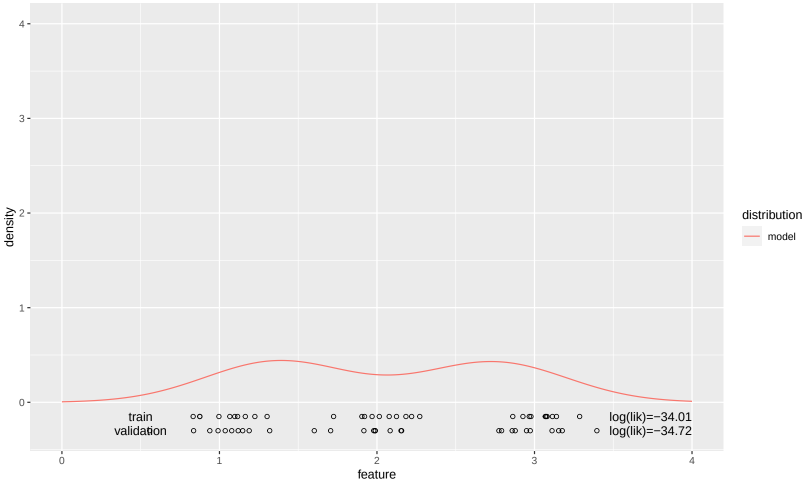
Divide into train and validation



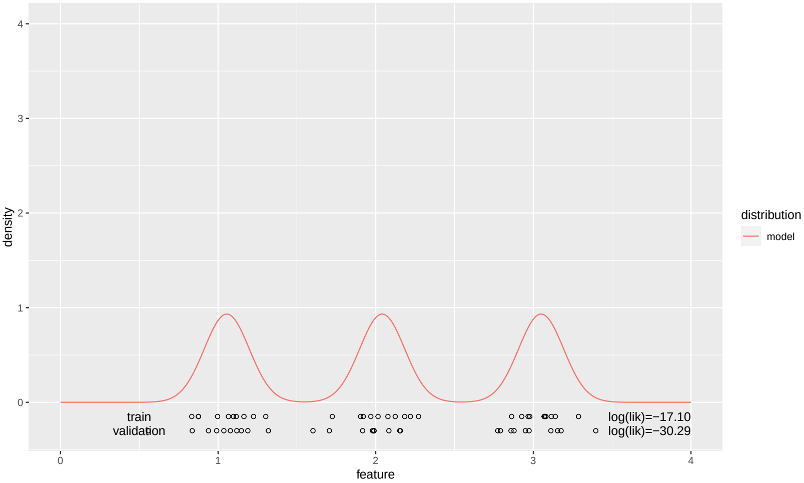
Fit gaussian mixture model 1



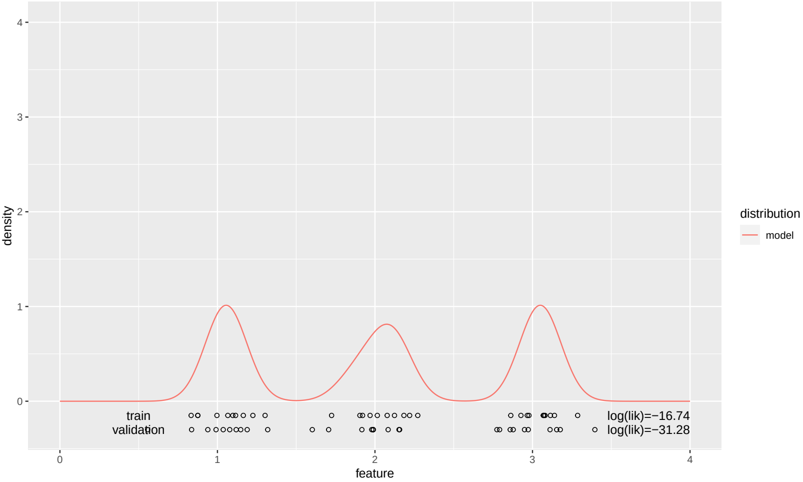
Fit gaussian mixture model 2



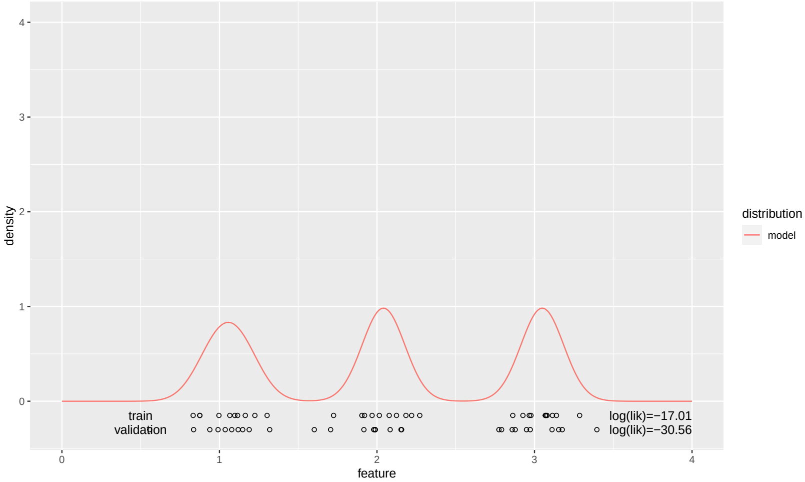
Fit gaussian mixture model 3



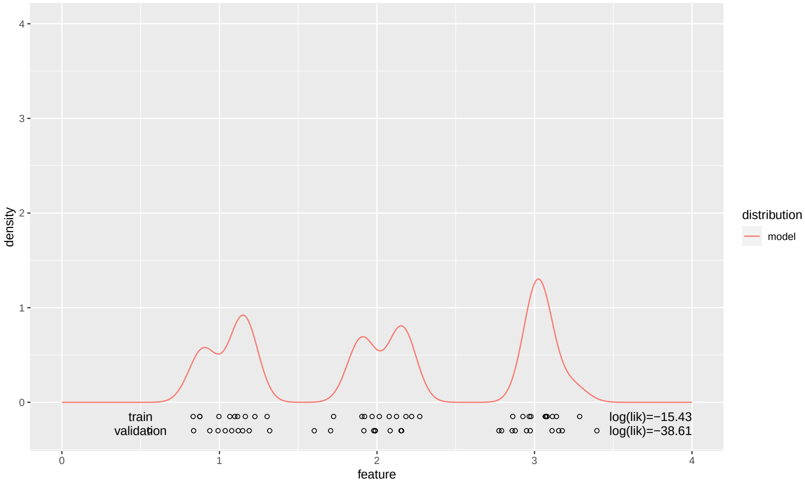
Fit gaussian mixture model 4



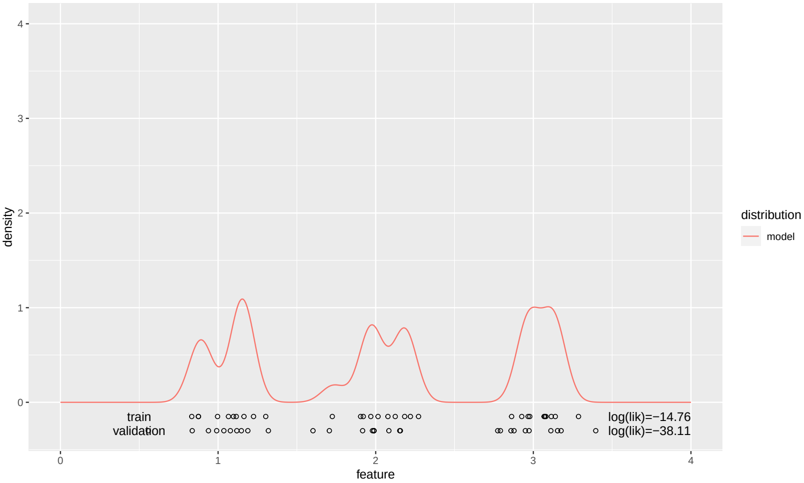
Fit gaussian mixture model 5



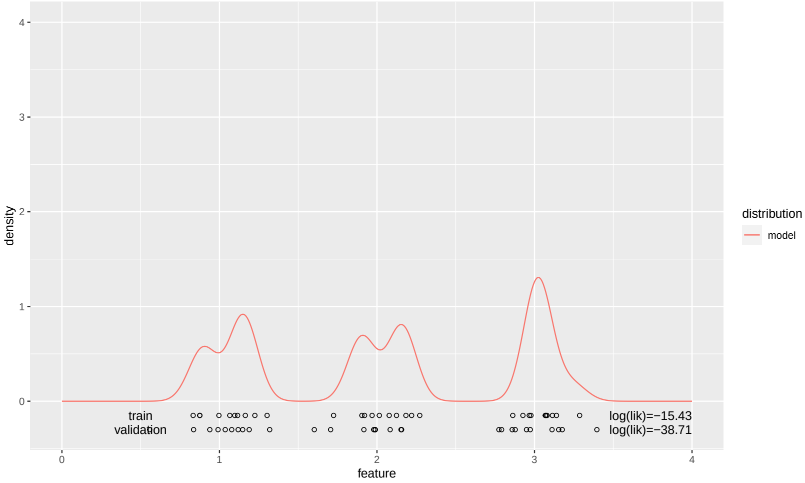
Fit gaussian mixture model 6



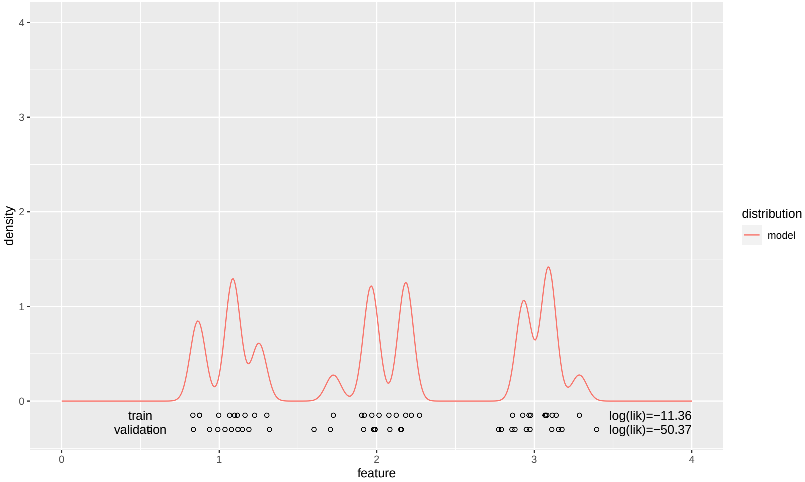
Fit gaussian mixture model 7



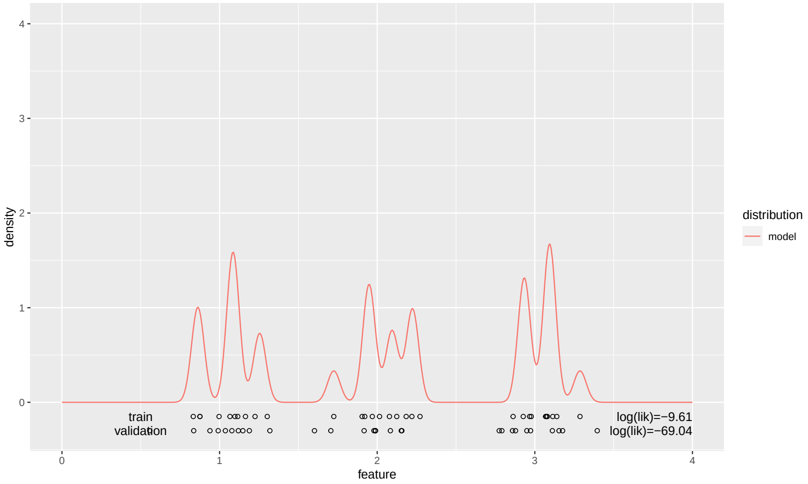
Fit gaussian mixture model 8



Fit gaussian mixture model 9



Fit gaussian mixture model 10



Overall negative log likelihood plot

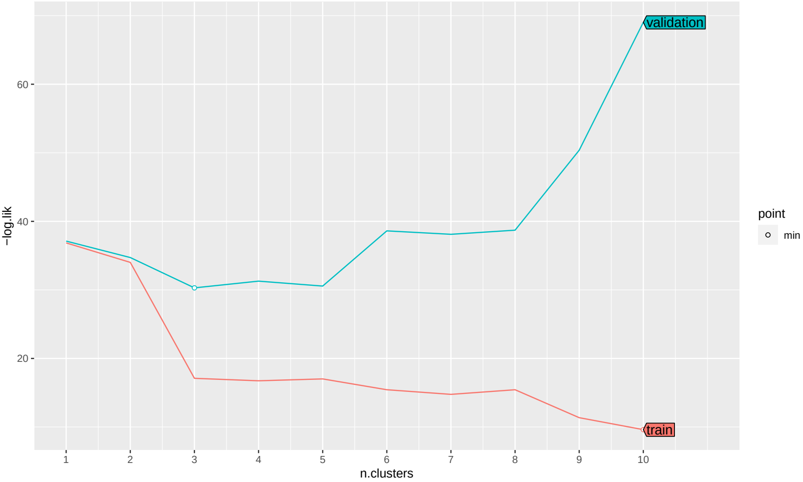


Diagram of 3-fold cross-validation

- K-fold cross-validation randomly assigns a fold ID number from 1 to K to each row.
- There are K splits; for each split data with that fold ID are validation, and all others are train.
- For each hyper-parameter (e.g., number of clusters), we compute the mean log likelihood over all validation sets/splits.
- Select model with largest mean validation log likelihood.

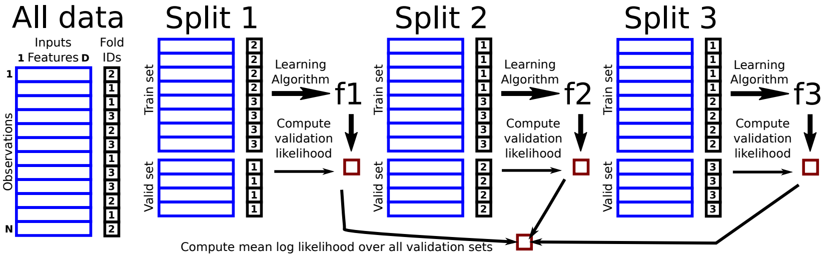


Figure 1: Cross-validation for unsupervised learning

Possible exam questions

- What kinds of clustering hyper-parameter values result in underfitting, and why should that be avoided?
- What kinds of clustering hyper-parameter values result in overfitting, and why should that be avoided?
- Using cross-validation with a single split, how should the number of clusters be chosen in Gaussian mixture models?
- Using K-fold cross-validation, how should the number of clusters be chosen in Gaussian mixture models?
- Describe/draw typical (negative) log likelihood curves, as a function of the number of clusters. Explain/draw where over/under-fitting occur, and which model size should be selected.