Clustering Model Selection

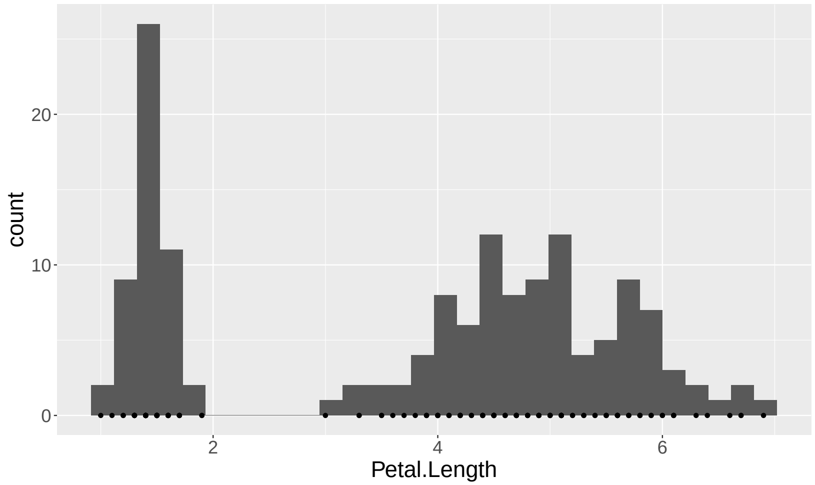
Toby Dylan Hocking

Clustering framework

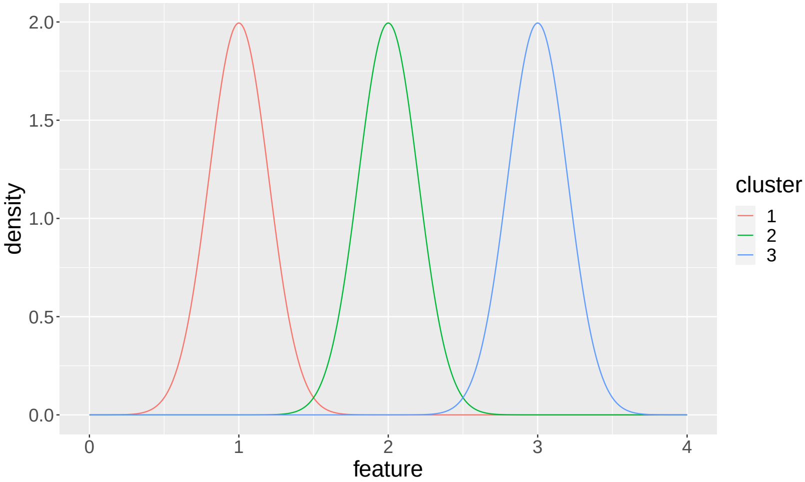
- Let $X = [x_1 \cdots x_n]^\intercal \in \mathbb{R}^{n \times p}$ be the data matrix (input for clustering), where $x_i \in \mathbb{R}^p$ is the input vector for observation i.
- Example iris $n = 150$ observations, $p = 4$ dimensions.
- Consider only one of those columns,

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

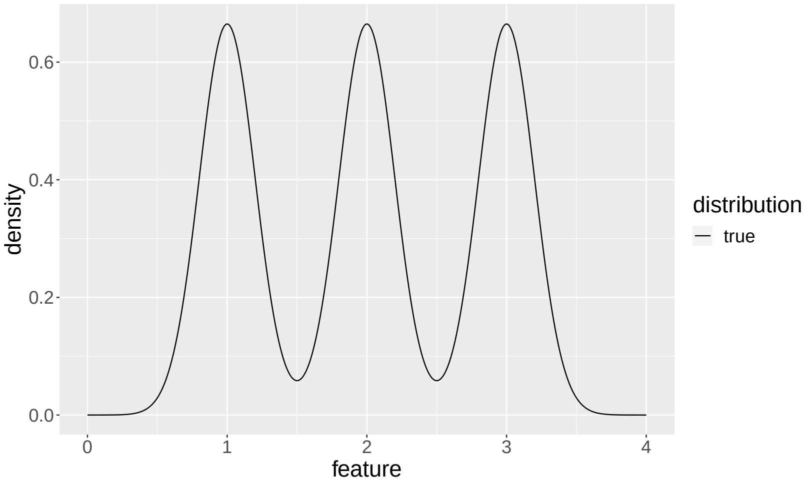
One column can be visualized as a histogram



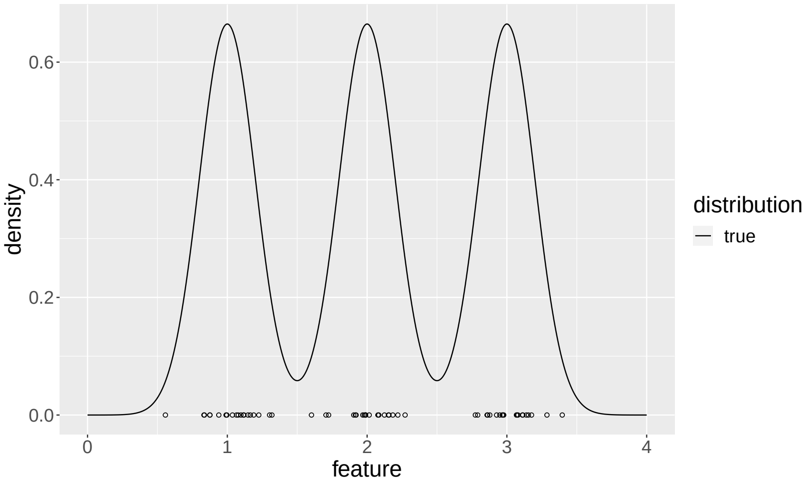
Simulation: three normal densities



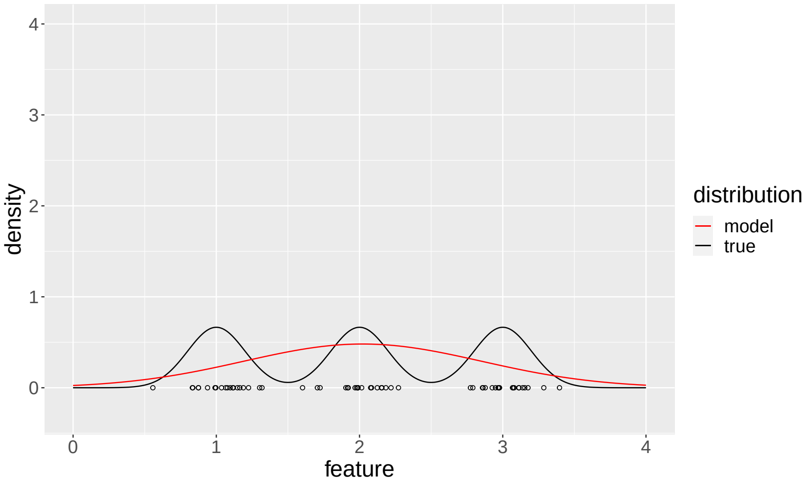
Mixture density



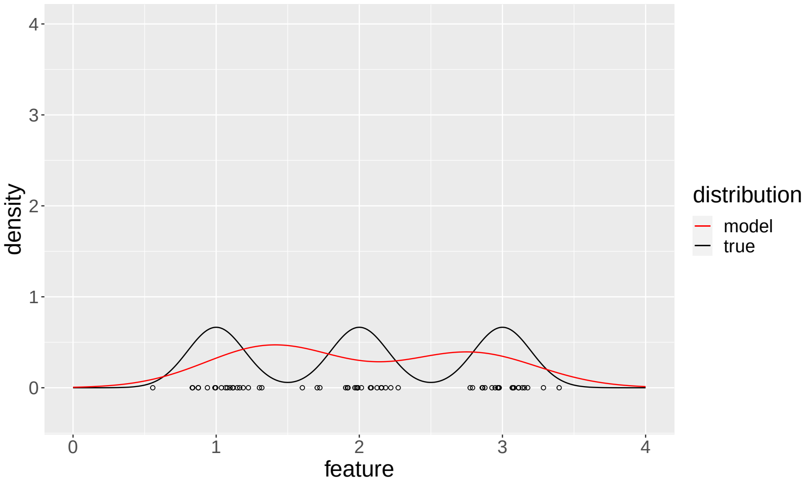
Generate 20 random data from each density



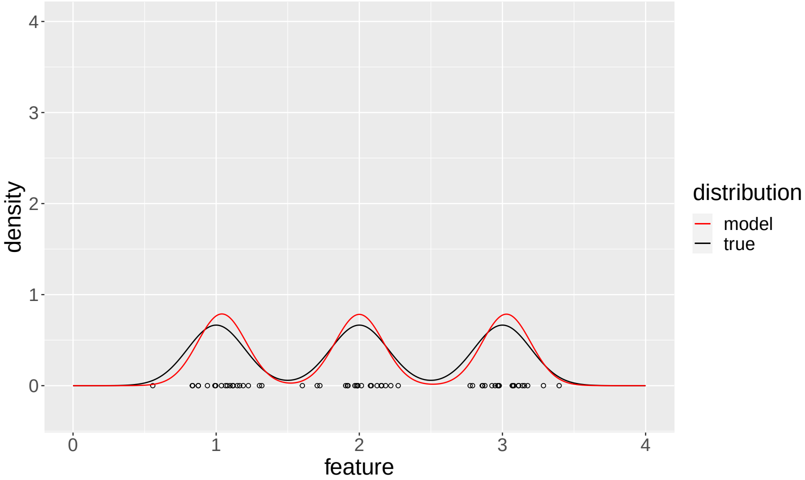
Fit gaussian mixture model 1



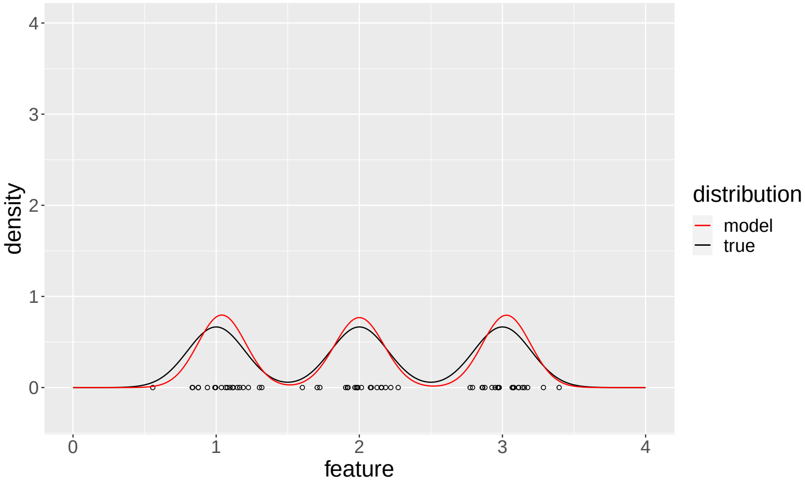
Fit gaussian mixture model 2



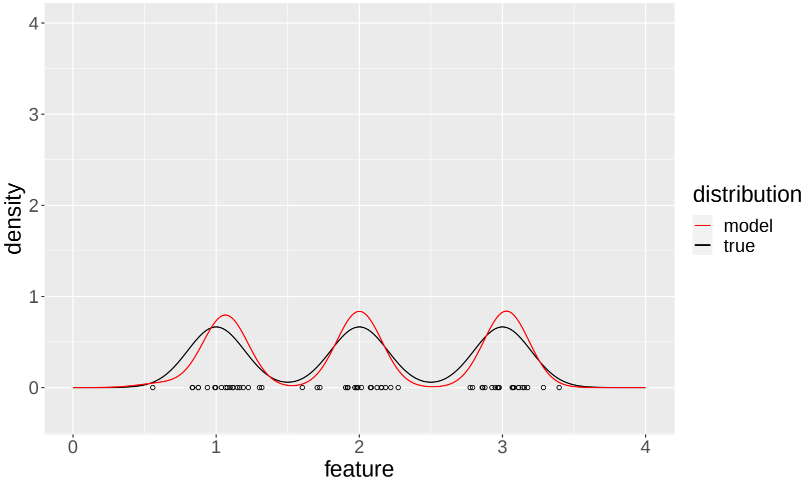
Fit gaussian mixture model 3



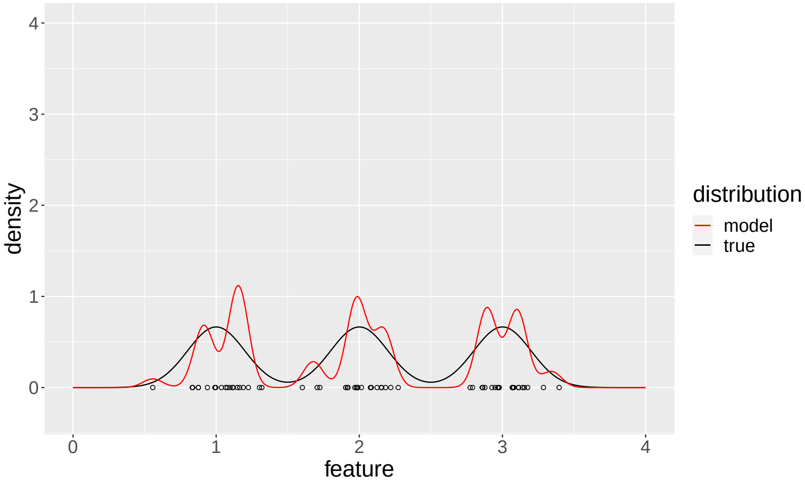
Fit gaussian mixture model 4



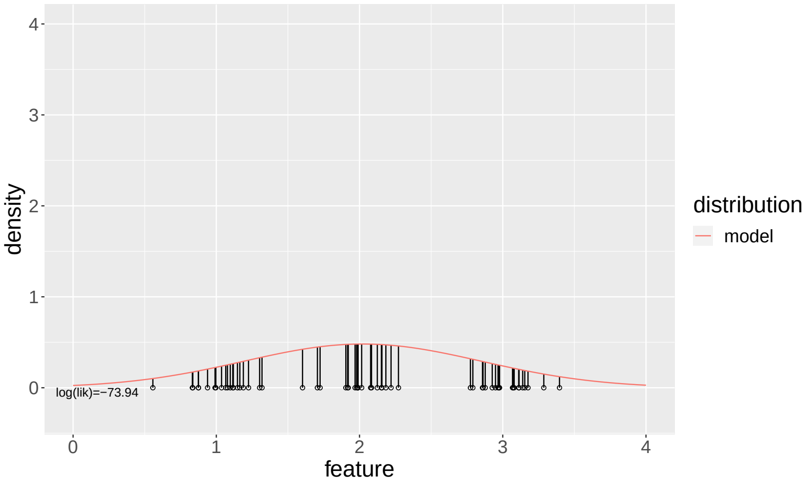
Fit gaussian mixture model 5



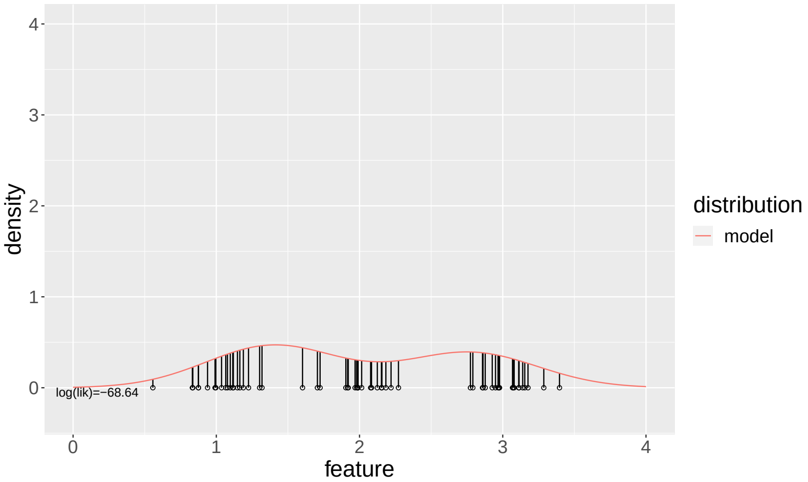
Fit gaussian mixture model 10



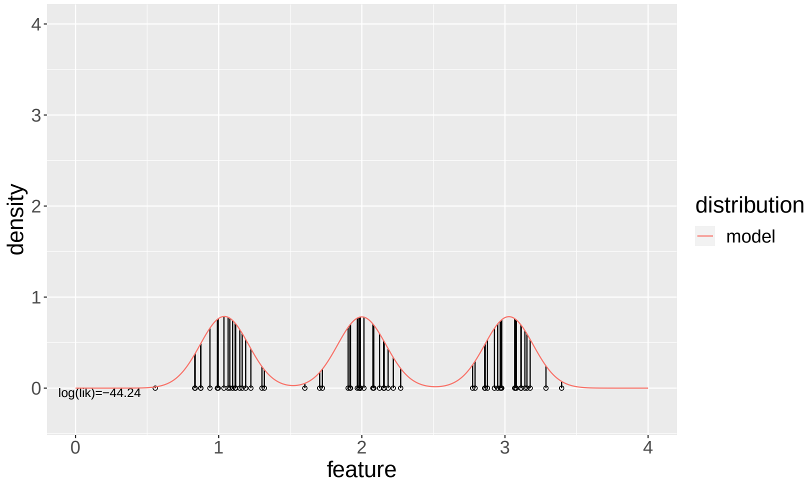
Fit gaussian mixture model 1



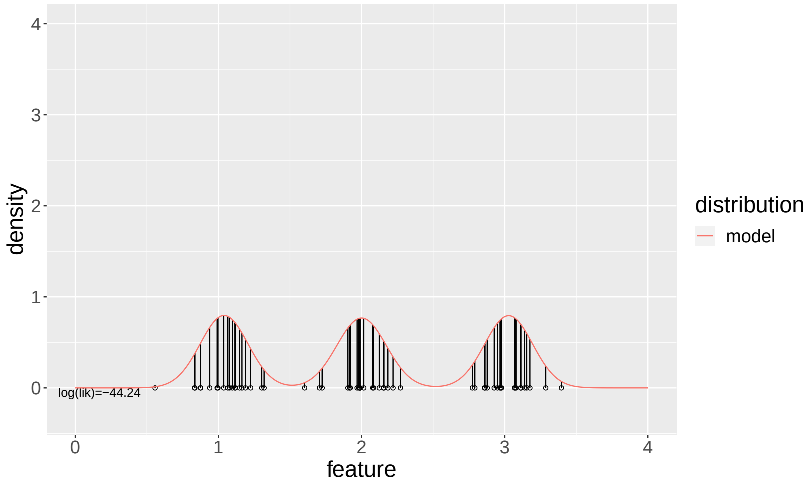
Fit gaussian mixture model 2



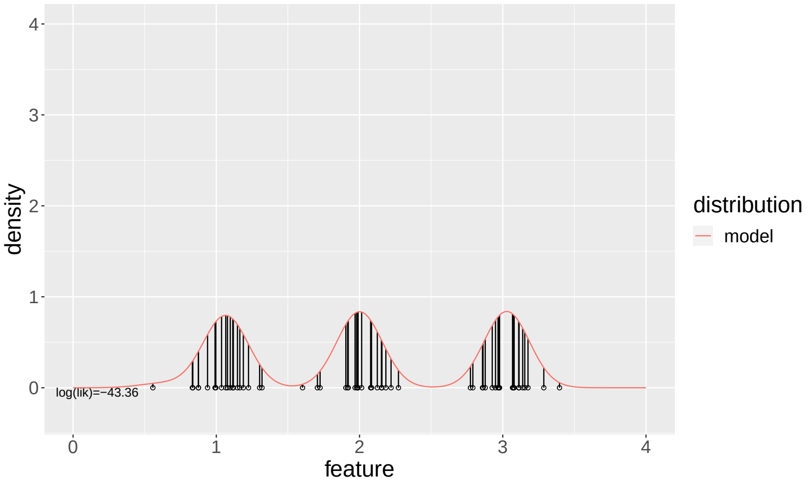
Fit gaussian mixture model 3



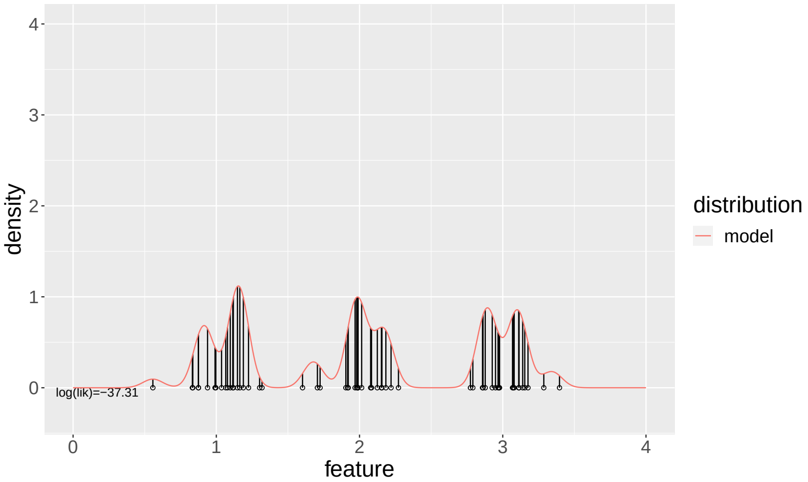
Fit gaussian mixture model 4



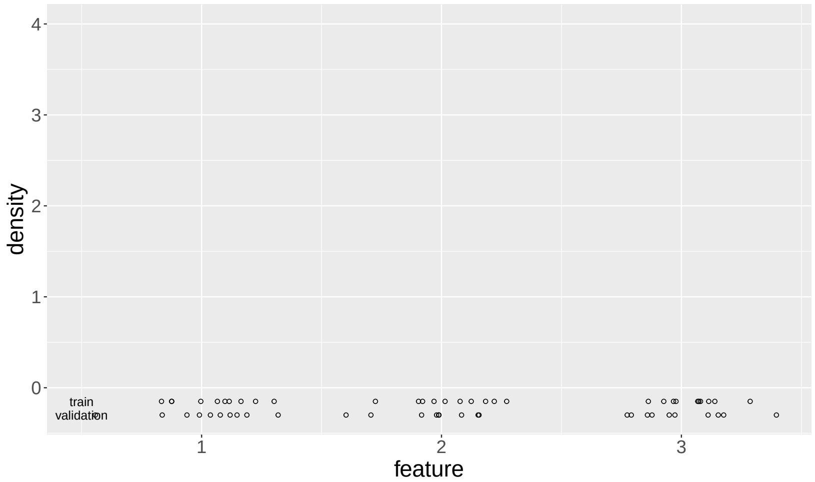
Fit gaussian mixture model 5



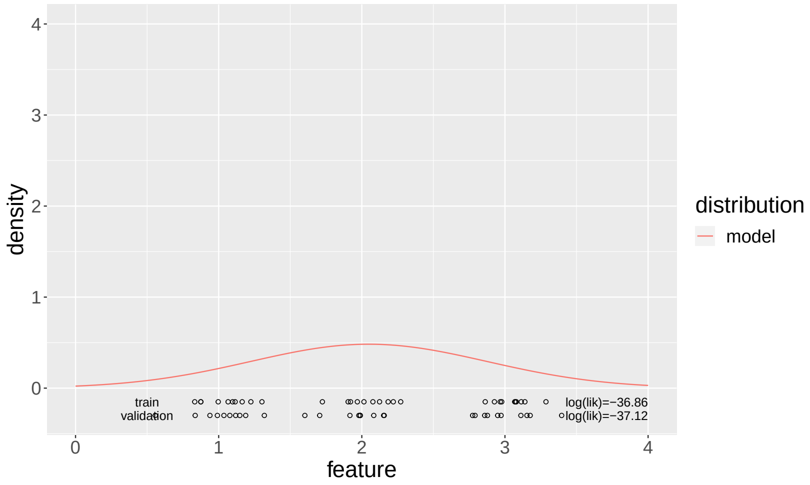
Fit gaussian mixture model 10



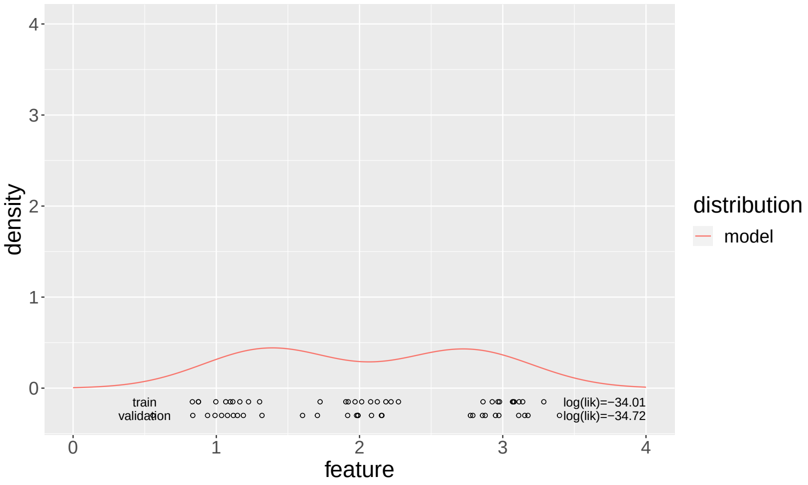
Divide into train and validation



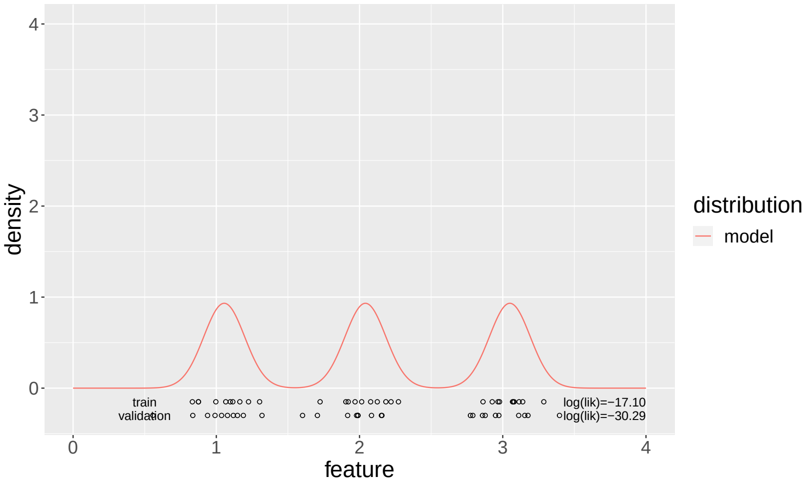
Fit gaussian mixture model 1



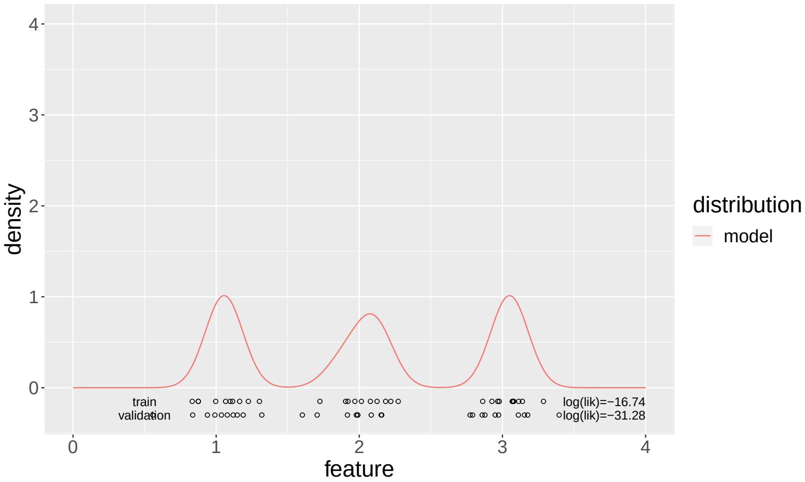
Fit gaussian mixture model 2



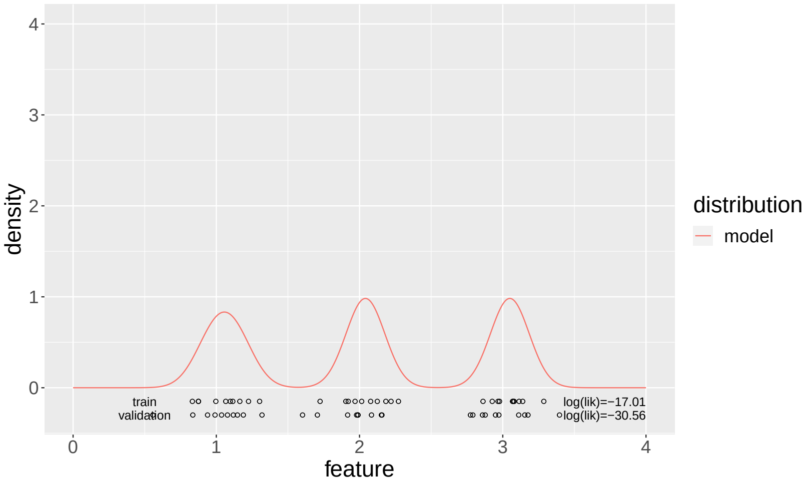
Fit gaussian mixture model 3



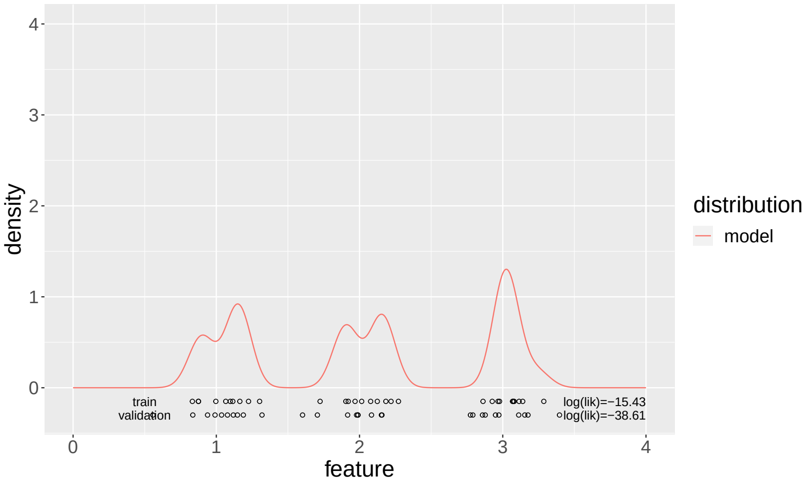
Fit gaussian mixture model 4



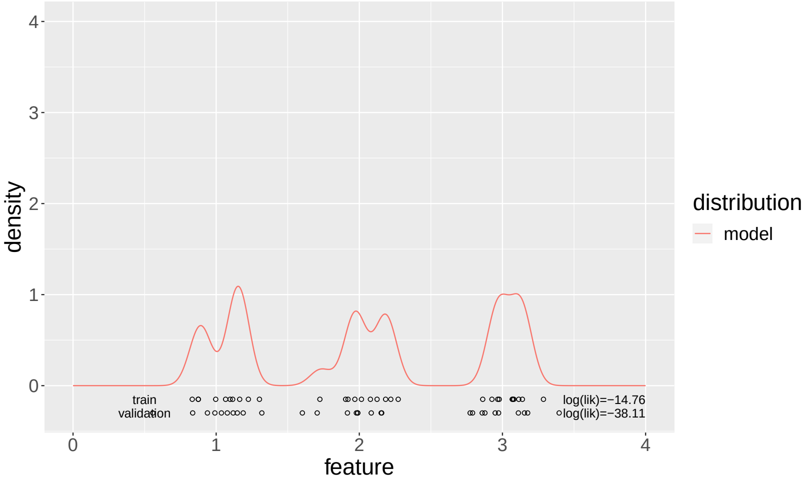
Fit gaussian mixture model 5



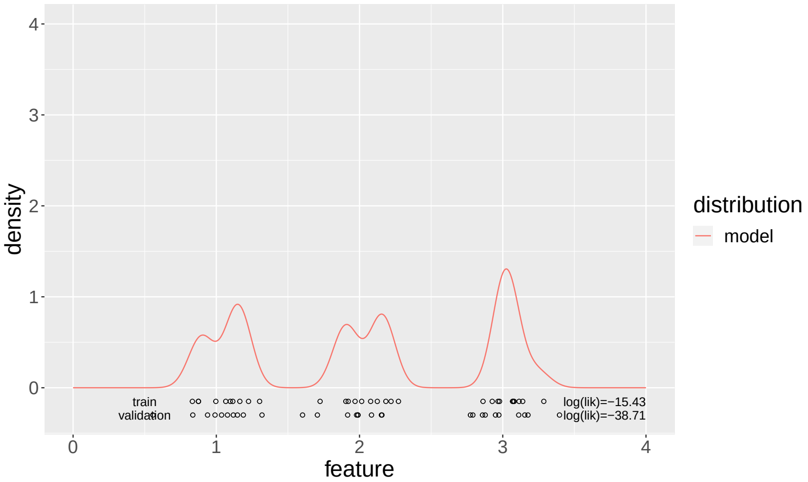
Fit gaussian mixture model 6



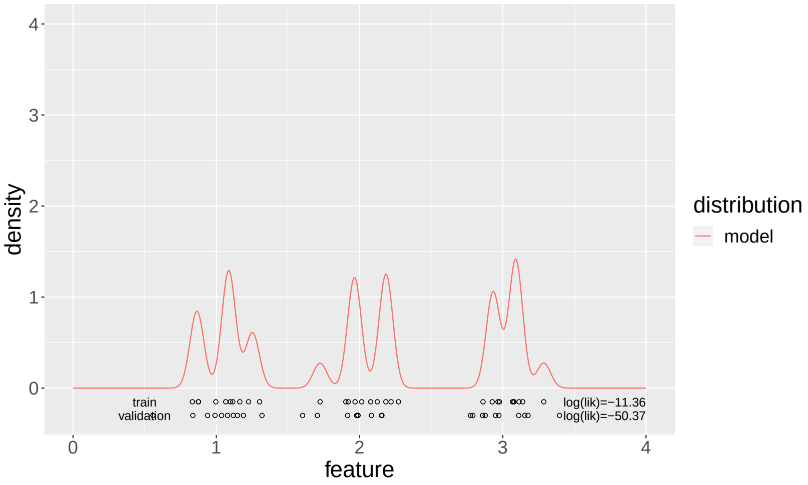
Fit gaussian mixture model 7



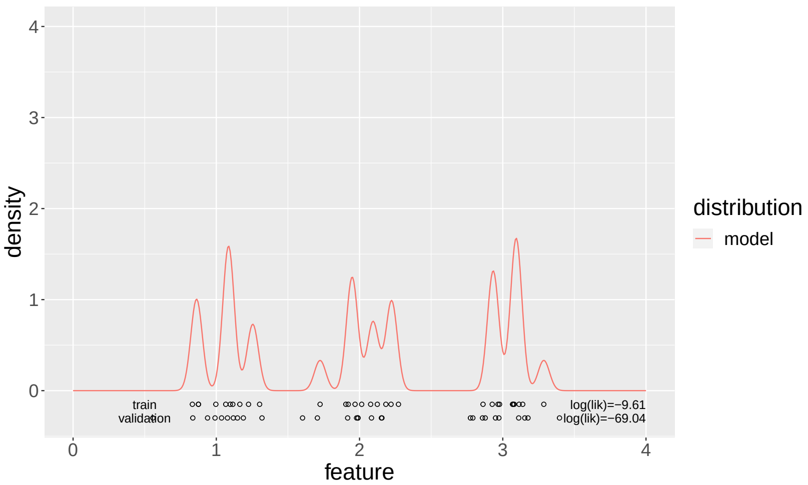
Fit gaussian mixture model 8



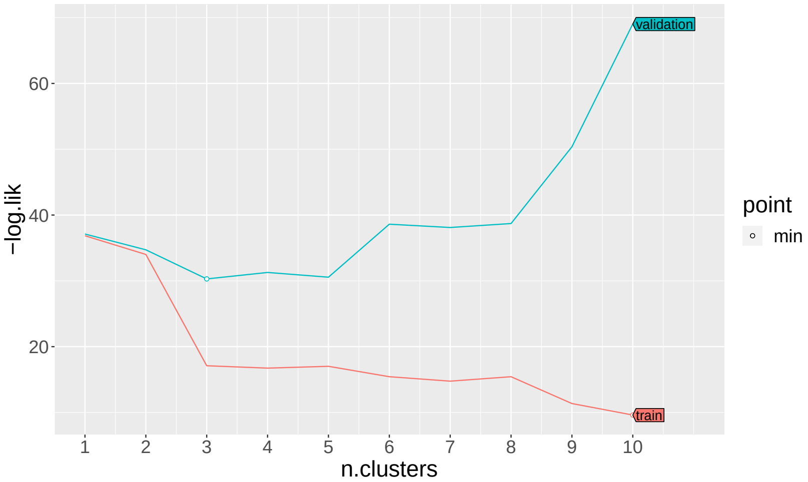
Fit gaussian mixture model 9



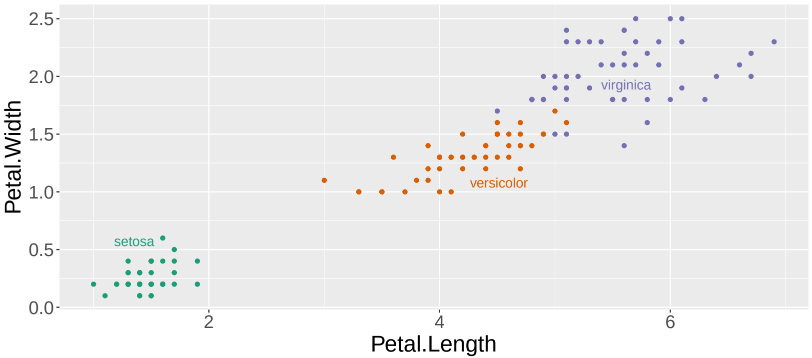
Fit gaussian mixture model 10



Overall log likelihood plot



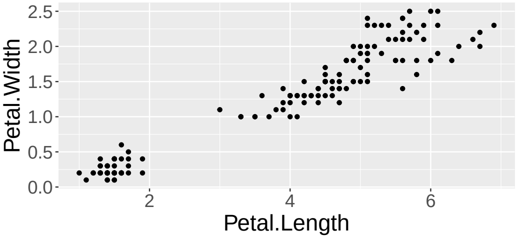
Visualize iris data with labels



Visualize iris data without labels

- 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, $p = 2$ dimensions.

```
##      Petal.Width Petal.Length
## [1,]         0.2          1.4
## [2,]         0.2          1.4
## [3,]         0.2          1.3
## [4,]         0.2          1.5
```



Possible exam questions

- TODO