Feldmann, Mews April 21, 2023

Hidden Markov Models - Practical Session 2

Exercise 1 - Independent mixture models

a) We are going to use the data set of eruptions of the Old Faithful (`?faithful`). The data is included in R. Read in the data set using the following line of code:
`data <- faithful`

b) Plot a histogram of the variable `eruptions`. Would you use a mixture of two normal distributions to analyze the eruption times? Why/why not?

c) The function `mllk()` (from lecture slide 43) returns the negative log likelihood of some input data given certain parameter values for $\mu_1, \mu_2, \sigma_1, \sigma_2$ and π_1 for a mixture model of two normal distributions.

```
mllk <- function(theta, x) {
  mu <- theta[1:2]
  sigma <- theta[3:4]
  pi <- theta[5]
  logl <- sum(log(pi * dnorm(x, mu[1], sigma[1]) +
                  (1 - pi) * dnorm(x, mu[2], sigma[2])))
  return(-logl)
}
```

Use this function to calculate the log-likelihood for the durations of the `eruptions` given the parameter set $\mu_1 = 2, \mu_2 = 4.5, \sigma_1 = 1, \sigma_2 = 1$ and $\pi_1 = 0.5$.

d) Also given on lecture slide 43 is the code for using an optimiser to find the parameter values that maximize the likelihood. Use the parameter values from c) as starting values and the `eruptions` data as input to find the maximum likelihood estimates.

e) Add the estimated normal densities to the histogram from b) in two different colours. To weight them accordingly, you can simply multiply the density values with the estimated π_1, π_2. Then add the overall mixture density using a third colour. In your opinion, does the model fit the data well?
Hint: With the argument `probability = TRUE`, the area of the histogram will sum to one which facilitates comparison to density curves.

f) Using the estimated parameters from d), simulate $n = 272$ new eruption durations.
Hint: You could first simulate a sequence with values $1, 2$ determining which mixture component is active for which of the 272 simulated eruption durations using `sample(1:2, ...)` and then use `rnorm()` to simulate eruption durations from both mixture components.

g) Plot a histogram of the simulated data from f) and compare it to the histogram of the actual data.
Hint: With `par(mfrow=c(nrows, ncols))` (default is `par(mfrow=c(1,1))`) you can decide how/whether you want to display multiple plots at once.

Exercise 2 - Markov chains

a) We extend the example from the lecture (a 2-state Markov chain to describe the weather) to a 3-state Markov chain. Possible states are sunny (state 1), cloudy (2), and rainy (3) days. Create a the following transition probability matrix in R:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

b) Suppose today is a sunny day. What are the probabilities that tomorrow is a sunny, cloudy, or rainy day? How about the day after tomorrow or a week from now?

c) Use the code from lecture slide 57 to calculate the stationary distribution of this Markov chain and compare it to the probabilities calculated in b).

d) Using the transition probability matrix above and a for-loop, simulate a Markov chain of length $T = 1000$, starting with a sunny day (initial distribution $\delta_1 = (1, 0, 0)$). Calculate how often which state appears in your simulated state sequence and compare these proportions to the stationary distribution.