Evaluación MLII: Ejercicio 2

Aprendizaje Supervisado Secuencial

Inmaculada Perea Fernández

junio 2017

Como sabemos, en la composición del ADN intervienen 4 bases nitrogenadas: adenina (A), guanina (G), timina (T) y citosina (C). Supongamos que la generación de estas bases está regulada por una variable no observable *S*, con dos posibles estados *S1* y *S2*. En la generación consecutiva de dos bases, la variable *S* permanece en el mismo estado en el 80% de los casos (ambos estados son igualmente probables inicialmente). Por otra parte, las probabilidades de generar las distintas bases en función del estado de la variable *S* aparecen recogidas en la siguiente tabla:

	A	G	T	C
S_1	0.4	0.3	0.2	0.1
S_2	0.1	0.1	0.3	0.5

Figure 1: Tabla enunciado Ejercicio 2 (Evaluación MLII)

Carga de librerías

```
if (!require('HMM')) install.packages('HMM'); library('HMM')
```

Establecimiento de la semilla

```
set.seed(123456789)
```

Apartado 1

Construya el modelo *HMM* correspondiente usando la librería *HMM*

1.1 Definición de parámetros

```
# Vector con los nombres de los estados ocultos
Estados = c("S1", "S2")

# Vector con los nombres de los símbolos observados
Observado = c("A", "G", "T", "C")

# Vector con las probabilidades iniciales de los estados
ProbIni = c(0.5, 0.5)

# Matriz con las probabilidades de transición entre estados
ProbTrans = matrix(c(0.8, 0.2, 0.2, 0.8),nrow=2,byrow=T)
```

```r
# Matriz con las probabilidades de emisión de los estados
ProbEmis = matrix(c(0.4, 0.3, 0.2, 0.1, 0.1, 0.1, 0.3, 0.5),nrow=2,byrow=T)
```

1.2 Construcción del modelo

```r
modHMM = initHMM(Estados, Observado, ProbIni, ProbTrans, ProbEmis)
```

```r
print(modHMM)
```

```
## $States
## [1] "S1" "S2"
##
## $Symbols
## [1] "A" "G" "T" "C"
##
## $startProbs
##  S1  S2
## 0.5 0.5
##
## $transProbs
##     to
## from  S1  S2
##   S1 0.8 0.2
##   S2 0.2 0.8
##
## $emissionProbs
##       symbols
## states   A   G   T   C
##     S1 0.4 0.3 0.2 0.1
##     S2 0.1 0.1 0.3 0.5
```

Apartado 2

Calcule la probabilidad de obtener la siguiente secuencia: *CGTCAGATA*

2.1 Secuencia observada

```r
SecObs = c("C", "G", "T", "C", "A", "G", "A", "T", "A")
```

2.2 Cálculo de las probabilidades forward

```r
(Pforward = exp(forward(modHMM, SecObs)))
```

```
##       index
## states    1     2       3        4          5           6           7
##     S1 0.05 0.027 0.00516 0.000546 0.00042912 0.000118908 4.04352e-05
##     S2 0.25 0.021 0.00666 0.003180 0.00026532 0.000029808 4.76280e-06
##       index
```

```
## states            8            9
##     S1 6.660144e-06 2.416781e-06
##     S2 3.569184e-06 4.187376e-07
```

2.3 Cálculo de las probabilidades backward

```
(Pbackward = exp(backward(modHMM, SecObs)))
```

```
##       index
## states            1           2           3          4         5       6
##     S1 1.302221e-05 4.81680e-05 0.000199296 0.00175280 0.0053504 0.02152
##     S2 8.737632e-06 7.30944e-05 0.000271344 0.00059072 0.0020336 0.00928
##       index
## states     7    8 9
##     S1 0.064 0.34 1
##     S2 0.052 0.16 1
```

2.4 Probabilidad de obtener la secuencia observada

```
for (t in 1:9)
  print(sum(Pforward[,t] * Pbackward[,t]))
```

```
## [1] 2.835518e-06
## [1] 2.835518e-06
## [1] 2.835518e-06
## [1] 2.835518e-06
## [1] 2.835518e-06
## [1] 2.835518e-06
## [1] 2.835518e-06
## [1] 2.835518e-06
## [1] 2.835518e-06
```

Apartado 3

Dado que se ha observado la secuencia anterior, calcule la probabilidad de que la variable S se haya encontrado en cada uno de los dos estados posibles a lo largo de la generación de la secuencia (probabilidades a posteriori)

3.1 Cálculo de las probabilidades a posteriori

```
(Pposterior = posterior(modHMM, SecObs))
```

```
##       index
## states         1        2         3         4         5         6
##     S1 0.2296266 0.458659 0.3626735 0.3375146 0.8097157 0.90244527
##     S2 0.7703734 0.541341 0.6373265 0.6624854 0.1902843 0.09755473
##       index
## states          7         8         9
##     S1 0.91265597 0.7986014 0.8523241
```

```
##      S2 0.08734403 0.2013986 0.1476759
```

3.2 Probabilidad a posteriori para cada estado

A continuación calcularemos la probabilidad de cada estado de la variable S condicionado a la secuencia observada $CGTCAGATA$

```
cat("P(Y9=S1 | X=CGTCAGATA): ",round(Pposterior["S1",9], 3), "\n")
```

```
## P(Y9=S1 | X=CGTCAGATA):  0.852
```

```
cat("P(Y9=S2 | X=CGTCAGATA): ",round(Pposterior["S2",9], 3), "\n")
```

```
## P(Y9=S2 | X=CGTCAGATA):  0.148
```

Apartado 4

A partir de la secuencia observada, determine la secuencia de estados más probable para la variable *S*

```
(Estados = viterbi(modHMM, SecObs))
```

```
## [1] "S2" "S2" "S2" "S2" "S1" "S1" "S1" "S1" "S1"
```

Apartado 5

Genere una secuencia de longitud 100 mediante simulación

```
(sim = simHMM(modHMM, 100))
```

```
## $states
##   [1] "S1" "S1" "S1" "S1" "S2" "S1" "S1" "S2" "S2" "S2" "S2" "S1" "S1" "S1"
##  [15] "S1" "S1" "S1" "S1" "S2" "S2" "S2" "S2" "S1" "S1" "S1" "S2" "S2" "S2"
##  [29] "S2" "S2" "S2" "S2" "S2" "S2" "S1" "S1" "S2" "S2" "S2" "S2" "S2" "S2"
##  [43] "S1" "S2" "S2" "S2" "S2" "S2" "S2" "S2" "S2" "S2" "S1" "S1" "S1" "S1"
##  [57] "S1" "S1" "S1" "S1" "S1" "S2" "S2" "S2" "S2" "S2" "S1" "S2" "S2" "S2"
##  [71] "S2" "S2" "S2" "S1" "S1" "S1" "S1" "S1" "S1" "S1" "S1" "S1" "S1" "S1"
##  [85] "S1" "S1" "S1" "S1" "S2" "S2" "S2" "S1" "S1" "S1" "S1" "S1" "S1" "S1"
##  [99] "S1" "S1"
##
## $observation
##   [1] "G" "T" "G" "G" "A" "A" "A" "C" "T" "C" "C" "A" "A" "G" "C" "A" "A"
##  [18] "A" "A" "C" "C" "C" "G" "T" "G" "A" "C" "C" "C" "T" "T" "C" "G" "C"
##  [35] "C" "A" "G" "T" "C" "G" "C" "G" "G" "C" "G" "C" "T" "C" "C" "C" "C"
##  [52] "C" "A" "G" "G" "G" "A" "G" "G" "A" "T" "C" "T" "T" "A" "C" "G" "G"
##  [69] "G" "C" "G" "C" "C" "A" "G" "A" "T" "G" "G" "G" "T" "A" "G" "A" "A"
##  [86] "A" "T" "A" "C" "C" "T" "A" "T" "A" "C" "A" "A" "A" "G" "G"
```