Project 2 EM Algorithm

Team K

March 24, 2023

Problem 6: Hidden Markov Models

According to slides 18, the HMM is parameterized by:

$$\begin{aligned} &\text{Initial state probabilities:} && I_k = P(Z_1 = k) \\ &\text{Transition probabilities:} && T_{kl} = P(Z_n = l | Z_{n-1} = k) \\ &\text{Emission probabilities:} && E_{kx} = P(X_n = x | Z_n = k) \end{aligned}$$

(a)

I_k should sum up to 1 $\longrightarrow$ degree of freedom $= K - 1$

In the $K \times K$ matrix T, each row needs to sum up to 1 $\longrightarrow$ degree of freedom $= K - 1$ for each row

Similarly, in the $K \times M$ matrix X, each row needs to sum up to 1 $\longrightarrow$ degree of freedom $= M - 1$ for each row

Hence, the maximum number of free parameters is $(K-1) + K \times (K-1) + K \times (M-1)$

(b)

Computing the stationary distribution is equivalent to solving $\pi^t = \pi^t T$.

Let $\pi^t = \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix}$), we need to solve:

$$\begin{aligned} \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix} &= \begin{bmatrix}\pi_1 & \pi_2\end{bmatrix} \begin{bmatrix}0.2 & 0.8 \\ 0.6 & 0.4\end{bmatrix} \\ &= \begin{bmatrix}0.2\pi_1 + 0.6\pi_2 & 0.8\pi_1 + 0.4\pi_2\end{bmatrix} \end{aligned}$$

Given that $\pi_1 = 1 - \pi_2$e can solve the equations:

$$\begin{cases} \pi_1 = 0.2\pi_1 + 0.6\pi_2 \\ \pi_2 = 0.8\pi_1 + 0.4\pi_2 \\ \pi_1 = 1 - \pi_2 \end{cases} \Rightarrow \begin{cases} \pi_1 = \dfrac{3}{7} \\ \pi_2 = \dfrac{4}{7} \end{cases}$$

Alternatively we can solve by solving eigens (which is porbably the more proper way to do it because solving the equations above by hand for high dimensional transition matrices is not a very smart idea). Since $\pi^t = \pi^t T$, we can re-write it into $\pi = T^t \pi$, which means π is just the corresponding eigenvector of eigenvalue $\lambda = 1$.

Note that the eigenvector should be normalized to satisfy the constraint that $\pi_1 + \pi_2 = 1$.

```
# initialize the transition matrix
T <- matrix(c(0.2, 0.8, 0.6, 0.4), 2, 2, byrow=TRUE)

# solve for the eigenvalues and eigenvectors for transpose of T
eigens <- eigen(t(T))

# get the index of lambda = 1
index <- which(eigens$values == 1)[1]
ev <- eigens$vectors[, index]

# normalize so that the probs sum to 1
pi <- ev / sum(ev)
pi
```

```
## [1] 0.4285714 0.5714286
```

```
# check that the two computation confirms
abs(pi[1] - 3/7) < 1e-5
```

```
## [1] TRUE
```

```
abs(pi[2] - 4/7) < 1e-5
```

```
## [1] TRUE
```

Problem 7: Predictinig protein secondary structure using HMMs

(a)

Read `proteins train.tsv`, `proteins test.tsv` and `proteins new.tsv` into the memory and store each in a `data.frame`

```
train <- read.csv("data/proteins_train.tsv", sep = "\t", header = FALSE)
test  <- read.csv("data/proteins_test.tsv",  sep = "\t", header = FALSE)
new   <- read.csv("data/proteins_new.tsv",   sep = "\t", header = FALSE)
```

```
header <- c("identifier",
             "sequence",
             "structure")
colnames(train) <- header
colnames(test)  <- header
colnames(new)   <- header[1:2]
```

```
head(train)
```

```
##   identifier
## 1    >101M:A
## 2    >102L:A
## 3    >102M:A
## 4    >103L:A
## 5    >103M:A
## 6    >104L:B
##
## 1                MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEA
## 2   MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVR
## 3                MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKAGVTVLTALGAILKKKGHHEA
## 4 MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNSLDAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVR
## 5                MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKAGVTVLTALGAILKKKGHHEA
## 6  MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSAAELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVR
##
## 1                CCCCHHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHCGGGGGGCTTTTTCCSHHHHHHCHHHHHHHHHHHHHHHHHHHHTTTTCCHH
## 2   CCHHHHHHHHHCCEEEEEEECTTSCEEEETTEEEESSSCTTTHHHHHHHHHHHTSCCTTBCCHHHHHHHHHHHHHHHHHHHHHHHHCTTHHHHHHHSCHHH
## 3                CCCCHHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHCGGGGGGCTTTTTCCSHHHHHHCHHHHHHHHHHHHHHHHHHHHTTTTCCHH
## 4 CCHHHHHHHHHCCEEEEEEECTTSCEEEETTEECCCCCCCCCHHHHHHHHHHHHTSCCTTBCCHHHHHHHHHHHHHHHHHHHHHHHHCTTTHHHHHHSCHHH
## 5                CCCCHHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHCGGGGGGCTTTTTCCSHHHHHHCHHHHHHHHHHHHHHHHHHHHHTTTCCHH
## 6  CCHHHHHHHHHCCCSBCEECTTSCEECTTTCCCCCCSSCCHHHHHHHHHHHHSCSCCTTBCCHHHHHHHHHHHHHHHHHHHHHHHTCTTHHHHHHHSCSSH
```

(b)

> Estimate the vector of initial state probabilities I, the matrix of transition probabilities T and the matrix for emission probabilities E by maximum likelihood.

First get all the possible amino acid and secondary structure states.

```r
# get all the possible amino acid states
aa_states <- train$sequence %>%
  strsplit("") %>%
  unlist() %>%
  unique() %>%
  sort()
aa_states
```

```
##  [1] "A" "C" "D" "E" "F" "G" "H" "I" "K" "L" "M" "N" "P" "Q" "R" "S" "T" "U" "V"
## [20] "W" "X" "Y"
```

```r
# get all the possible secondary structure states, and sort them
ss_states <- train$structure %>%
  strsplit("") %>%
  unlist() %>%
  unique() %>%
  sort()
ss_states
```

```
## [1] "B" "C" "E" "G" "H" "I" "S" "T"
```

```r
MLE <- function(data, aa_states, ss_states) {
  k <- length(ss_states) # num of latent state
  m <- length(aa_states) # num of observed state
  # initialize the parameters
  I <- rep(0.0, k)
  names(I) <- ss_states

  E <- matrix(0.0, nrow=k, ncol=m)
  dimnames(E) <- list(ss_states, aa_states)

  Tr <- matrix(0.0, nrow=k, ncol=k)
  dimnames(Tr) <- list(ss_states, ss_states)

  N <- nrow(data) # num of data points

  # iterate over each row of the data
  for (i in 1:N) {
    seq <- data$sequence[i]
    struct <- data$structure[i]

    ss_1st <- struct %>% substr(1, 1)
    I[ss_1st] <- I[ss_1st] + 1.0

    for (j in 1:nchar(seq)) {
      aa <- seq %>% substr(j, j)
      ss <- struct %>% substr(j, j)

      E[ss, aa] <- E[ss, aa] + 1.0

      if (j < nchar(seq)) {
        ss_next <- struct %>% substr(j+1, j+1)

        Tr[ss, ss_next] <- Tr[ss, ss_next] + 1.0
      }
    }
  }

  # convert thee counts to log probs
  I <- I / sum(I)
  E <- E / rowSums(E)
  Tr <- Tr / rowSums(Tr)

  return(list(I=I, E=E, Tr=Tr))
}
```

```
res <- MLE(train, aa_states, ss_states)
```

```
res$I
```

```
## B C E G H I S T
## 0 1 0 0 0 0 0 0
```

```
res$E
```

```
##            A          C          D          E          F          G          H
## B 0.04524422 0.03239075 0.06580977 0.02467866 0.04215938 0.05192802 0.03239075
## C 0.06450873 0.01585481 0.07227845 0.04888328 0.03388859 0.08271453 0.02583216
## E 0.05804938 0.02348868 0.02798043 0.04002518 0.05247046 0.04809315 0.02752267
## G 0.10814116 0.01677149 0.09696017 0.08333333 0.03511530 0.06289308 0.03022362
## H 0.12500000 0.01249647 0.05261579 0.08629271 0.03646569 0.03503601 0.02188647
## I 0.23529412 0.00000000 0.17647059 0.00000000 0.00000000 0.00000000 0.00000000
## S 0.05445111 0.01536492 0.08605701 0.05215985 0.02331694 0.14219287 0.02304738
## T 0.06342143 0.01168713 0.07736021 0.06015118 0.02348148 0.19911006 0.02122983
##            I          K          L          M          N          P          Q
## B 0.06221080 0.04575835 0.07506427 0.01645244 0.05449871 0.04215938 0.02982005
## C 0.04323518 0.06118295 0.07018550 0.02319447 0.05708306 0.08277187 0.03145159
## E 0.09306783 0.05767745 0.10511258 0.02374617 0.02400366 0.01782393 0.03135639
## G 0.03162124 0.05957372 0.05835080 0.01502446 0.05241090 0.06446541 0.03546471
## H 0.05491034 0.07296668 0.11285654 0.02974089 0.03641274 0.01929187 0.04131954
## I 0.05882353 0.00000000 0.23529412 0.05882353 0.00000000 0.00000000 0.00000000
## S 0.02749511 0.07244423 0.05067727 0.01489319 0.05492284 0.05761844 0.03093200
## T 0.02466091 0.07918297 0.04975071 0.01302740 0.06733501 0.06792473 0.03683054
##            R          S          T           U          V           W
## B 0.04627249 0.04627249 0.08020566 0.000000000 0.13110540 0.008740360
## C 0.04733507 0.07981880 0.07184839 0.000000000 0.05106224 0.008171106
## E 0.03833720 0.05198409 0.07779018 0.000000000 0.12651275 0.019740795
## G 0.03913347 0.06970650 0.04297694 0.000174703 0.02428372 0.024633124
## H 0.06412384 0.04048997 0.04356114 0.000000000 0.07183705 0.013061282
## I 0.11764706 0.00000000 0.00000000 0.000000000 0.05882353 0.058823529
## S 0.05364243 0.08807871 0.07224206 0.000000000 0.03935575 0.016106207
## T 0.04122661 0.06438643 0.04444325 0.000000000 0.02021123 0.010132418
##              X          Y
## B 0.0010282776 0.06580977
## C 0.0009461280 0.02775309
## E 0.0000000000 0.05521701
## G 0.0000000000 0.04874214
## H 0.0000000000 0.02963499
## I 0.0000000000 0.00000000
## S 0.0002695599 0.02473212
## T 0.0001072214 0.02433925
```

```
res$Tr
```

```
##              B          C            E           G          H            I
## B 0.0185089974 0.60874036 0.0303341902 0.020565553 0.02365039 0.000000e+00
## C 0.0288740867 0.51503859 0.1113882450 0.025176795 0.08471493 0.000000e+00
## E 0.0039195491 0.10814522 0.8126054988 0.004520356 0.00557892 0.000000e+00
## G 0.0078616352 0.11198463 0.0186932215 0.697938505 0.03336827 0.000000e+00
## H 0.0004412595 0.01782689 0.0003000565 0.002859362 0.91049492 1.765038e-05
## I 0.0000000000 0.00000000 0.0000000000 0.000000000 0.05882353 8.235294e-01
## S 0.0258777546 0.38250556 0.0857874520 0.018262686 0.06738999 6.738999e-05
## T 0.0179059669 0.22709484 0.0695866617 0.012866563 0.04026162 5.361068e-05
##            S          T
## B 0.15372751 0.14447301
## C 0.13873647 0.09607089
## E 0.02932509 0.03590536
## G 0.05765199 0.07250175
## H 0.01623835 0.05182152
## I 0.00000000 0.11764706
## S 0.35656042 0.06354876
## T 0.12040959 0.51182115
```

(c)

Estimate the stationary distribution π of the Markov chain by solving the eigenvalue problem and by using a brute-force approach.

Eigenvalue method

Brute-force

(d)

Having estimated the parameters, i.e., the emission and transition matrices E, T and the vector of initial state probabilities I, you can predict the latent state sequence Z of a protein's amino acid sequence X using the Viterbi algorithm. Use the Viterbi algorithm provided in `viterbi.r` (carefully read the parameter description!) and iterate over each `data.frame` of `proteinstest.tsv` and `proteins new.tsv` row by row and use the amino acid sequence to predict its secondary structure, which you add to the `data.frame` as a new column. Save the extended `data.frame` of `proteins new.tsv` including the predicted secondary structure as a tsv file and hand it in together with your pdf.

(e)

Estimate confidence intervals for each parameter in I, E and T with bootstrapping. In a single bootstrap run i estimate the probabilities for I_i, E_i and T_i the same as before, but not on the original data set `proteins train.tsv`, but on the resampled data set. i.e., sample with replacement as many rows from `proteins train.tsv` as the original data set has. Run a thousand bootstraps and compute the empirical 95% confidence intervals for each single parameter in $\{I_i\}_i$, $\{E_i\}_i$ and $\{T_i\}_i$.

(f)

Use the following measure to compute the accuracy of the predicted secondary structure $P = (p_i)$ for the `data.frame` of `proteins test.tsv` given the real secondary structure $S = (s_i)$:

$$a(P, S) = \frac{1}{L} \sum_i \begin{cases} 1 & \text{if } p_i = s_i \\ 0 & \text{if } p_i \neq s_i \end{cases}$$

with sequence length L. Compute the accuracy for every protein in your `data.frame` and store the accuracies in a vector. What is the accuracy of the Viterbi algorithm over all sequences (i.e. call `summary` on the vector of accuracies)?

(g)

Instead of using the Viterbi algorithm, now randomly guess secondary structures for all sequences. Compare the global accuracies of the Viterbi and the random approach and plot all accuracy distributions using boxplots.