Advanced Algorithms for Data Science
Homework 3

Krikun Gosha

1 Dynamic programming

Question *Propose an algorithm to compute the optimal shipping schedule, that is an optimal sequence of n choices between A and B generating the minimal cost.*

This exercise could be reduced to Interval Scheduling Problem. All we need is in correct way describe optimal solution function:

$$OPT(i) = min \begin{cases} OPT(i-1) + r \cdot s_i \\ OPT(i-4) + c \cdot 4, \text{if } i \geq 4 \end{cases}$$

and $OPT(0) = 0$. First case represent calculation cost by one week (based on previous), second case "replace" optimal results of 4 previous weeks if it have less cost. Second case we can calculate only in comparison with first one, so for first three steps we have to conduct costs to compare with.

And for getting the final answer we should make backward induction, from the end, to determine a sequence of optimal actions.

2 Protein sequence alignment

Question *Compute a maximal-score alignment of protein sequences under the scoring matrix BLOSUM62 and the indel penalty d = 8.*

Sequences:

```
HEAGAWGHEE
PAWHEAE
```

If I right get the question - we should calculate by hands (or by implementation in some programming language). I think I will faster calculate by hands.

		H		E		A		G		A		W		G		H		E		E	
	0	-8		-16		-24		-32		-40		-48		-56		-64		-72		-80	
P	-8	-2	-16	-9	-24	-17	-32	-26	-40	-33	-48	-44	-56	-50	-64	-58	-72	-65	-80	-73	-88
		-16	-2	-10	-9	-17	-17	-25	-25	-33	-33	-41	-41	-49	-49	-57	-57	-65	-65	-73	-73
A	-16	-10	-10	-3	-17	-5	-25	-17	-33	-21	-41	-36	-49	-41	-57	-51	-65	-58	-73	-66	-81
		-24	-10	-18	-3	-11	-5	-13	-13	-21	-21	-29	-29	-37	-37	-45	-45	-53	-53	-61	-61
W	-24	-18	-18	-13	-11	-6	-13	-7	-21	-16	-29	-10	-37	-31	-45	-39	-53	-48	-61	-56	-69
		-32	-18	-26	-11	-19	-6	-18	-7	-15	-15	-25	-10	-18	-18	-26	-26	-34	-34	-42	-42
H	-32	-16	-26	-18	-19	-13	-14	-8	-15	-9	-23	-17	-18	-12	-26	-10	-34	-26	-42	-34	-50
		-40	-16	-24	-18	-26	-13	-21	-8	-16	-9	-17	-17	-25	-12	-20	-10	-18	-18	-26	-26
E	-40	-32	-24	-11	-26	-19	-21	-15	-16	-9	-17	-14	-25	-19	-20	-12	-18	-5	-26	-13	-32
		-48	-24	-32	-11	-19	-19	-27	-15	-23	-9	-17	-14	-22	-19	-27	-12	-20	-5	-13	-13
A	-48	-42	-32	-25	-19	-7	-27	-19	-23	-11	-17	-12	-22	-14	-27	-21	-20	-13	-13	-6	-21
		-56	-32	-40	-19	-27	-7	-15	-15	-23	-11	-19	-12	-20	-14	-22	-20	-28	-13	-21	-6
E	-56	-48	-40	-27	-27	-20	-15	-9	-23	-16	-19	-14	-20	-14	-22	-14	-28	-15	-21	-8	-14
		-64	-40	-48	-27	-35	-15	-23	-9	-17	-16	-24	-14	-22	-14	-22	-14	-22	-15	-23	-8

Each cell in table consists of 4 values (this will help us do backward induction):

$Score_{(i-1,j-1)} + s(S_i, T_i)$ $\quad$ $Score_{(i,j-1)} - d$
$Score_{(i-1,j)} - d$ $\quad$ maximal between all three

I done backward induction with coloring cells (different ways in different colors) and stressed forks with red color.
So this sequences could be align with same score by 6 different ways:

```
H  E  A  G  A  W  G  H  E  E
      |        |           |
   P  A        W  H  E  A  E
```

Score: $-8 + -1 + 4 + -8 + -8 + 11 + -2 + 0 + -1 + 5 = -8$

```
H  E  A  G  A  W  G  H  E     E
      |        |     |  |     |
   P  A        W     H  E  A  E
```

Score: $-8 + -1 + 4 + -8 + -8 + 11 + -8 + 8 + 5 + -8 + 5 = -8$

```
H  E  A  G  A  W  G  H  E  E
            |  |        |
   P        A  W  H  E  A  E
```

Score: $-8 + -1 + -8 + -8 + 4 + 11 + -2 + 0 + -1 + 5 = -8$

```
H  E  A  G  A  W  G  H  E     E
            |  |     |  |     |
   P        A  W     H  E  A  E
```

Score: $-8 + -1 + -8 + -8 + 4 + 11 + -8 + 8 + 5 + -8 + 5 = -8$

```
H  E  A  G  A  W  G  H  E  E
            |  |        |
      P     A  W  H  E  A  E
```

Score: $-8 + -8 + -1 + -8 + 4 + 11 + -2 + 0 + -1 + 5 = -8$

```
H  E  A  G  A  W  G  H  E     E
            |  |     |  |     |
      P     A  W     H  E  A  E
```

Score: $-8 + -8 + -1 + -8 + 4 + 11 + -8 + 8 + 5 + -8 + 5 = -8$

3 Hidden Markov Models

Question *Finish the example of the Forward-Backward algorithm that we didn't finish in class.*

Two hidden states: raining, not-raining
Probabilities to stay in the same state is 0.7, to change 0.3
Probabilities modelling the person's behaviour:

	umbrella	no umbrella
raining	0.9	0.1
not raining	0.2	0.8

Observation sequence: (umbrella, umbrella, no umbrella)

What is the probability that it was not raining on day 2
(this is Forward-Backward Problem)

So we need calculate $P(\pi_2 = \text{not raining}|\text{umbrella})$

From lecture:

$$P(\pi_i = k|x) = \frac{P(x, \pi_i = k)}{P(x)} = \frac{f_k(i) \cdot b_k(i)}{P(x)}$$

Transition matrix $A = a_{kl}$:

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$$

Emitting matrix $E = e_k(b)$:

$$E = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

And initial state probability matrix (suppose all states equiprobable) $I = p_0(k)$:

$$I = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

Forward probability $f_k(i) = P(x_1 \dots x_i, \pi_i = k)$:

$$f_k(i) = e_k(x_i) \cdot \sum_{l \in Q} f_l(i-1) \cdot a_{lk}, \text{ and base case: } f_k(1) = p_o(k) \cdot e_k(x_1)$$

Where Q - set of hidden states (raining, not raining).

Backward probability $b_k(i) = P(\pi_i = k, x_{i+1} \dots x_n)$:

$$b_k(i) = \sum_{l \in Q} e_l(x_{i+1}) \cdot b_l(i+1) \cdot a_{kl}, \text{ and base case: } b_k(n-1) = \sum_{l \in Q} a_{kl} \cdot e_l(x_n)$$

Lets compute forward and backward probabilities:

$f_{not\ raining}(2) = 0.2 \cdot (0.5 \cdot 0.2 \cdot 0.7 + 0.5 \cdot 0.9 \cdot 0.3) = 0.041$
$b_{not\ raining}(2) = (0.7 \cdot 0.8 + 0.3 \cdot 0.1) = 0.59$

$f_{raining}(2) = 0.9 \cdot (0.5 \cdot 0.9 \cdot 0.7 + 0.5 \cdot 0.2 \cdot 0.3) = 0.405$
$b_{raining}(2) = (0.7 \cdot 0.1 + 0.3 \cdot 0.8) = 0.31$

And lets recover $P(x)$ from $\sum_k P(\pi_i = k|x) = 1$ then: $\frac{0.041 \cdot 0.59}{P(x)} + \frac{0.405 \cdot 0.31}{P(x)} = 1$
then: $P(x) = 0.041 \cdot 0.59 + 0.405 \cdot 0.31 = 0.14974$

Thus $P(\pi_2 = \text{not raining}|\text{umbrella}) = 0.161546681$