T-61.5020 Statistical Natural Language Processing

Answers 5 — Markov chains and Hidden Markov Models
Version 1.0

1. a) The Markov chain is drawn in figure 1.

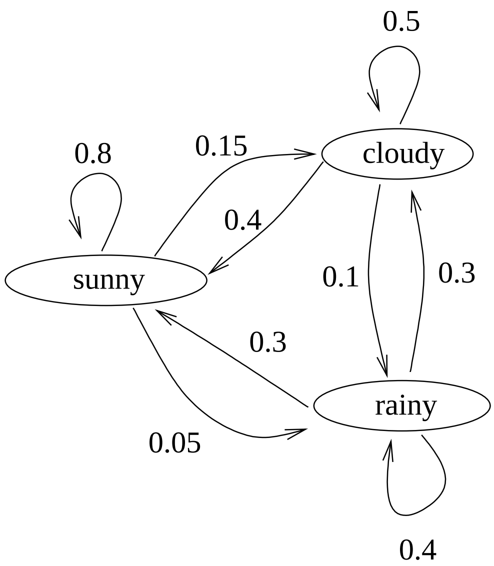


Figure 1: Weather as a Markov chain.

b) Let's calculate the probability for state sequence $\mathbf{S} = (S_3, S_2, S_1, S_1, S_1)$ when we know that we start from state S_2:

$$\begin{aligned}
P(\mathbf{S} \mid q_0 = S_2) &= P(q_1 = S_3, q_2 = S_2, q_3 = S_1, q_4 = S_1, q_5 = S_1 \mid q_0 = S_2) \\
&= P(q_1 = S_3 \mid q_0 = S_2) \cdot P(q_2 = S_2 \mid q_0 = S_2, q_1 = S_3) \\
&\quad \cdot P(q_3 = S_1 \mid q_0 = S_2, q_1 = S_3, q_2 = S_2) \\
&\quad \cdot P(q_4 = S_1 \mid q_0 = S_2, q_1 = S_3, q_2 = S_2, q_3 = S_1) \\
&\quad \cdot P(q_5 = S_1 \mid q_0 = S_2, q_1 = S_3, q_2 = S_2, q_3 = S_1, q_4 = S_1)
\end{aligned}$$

The probability in the first row is split several times using the formula $P(A, B|C) = P(A|C) \cdot P(B|A, C)$.

Let's apply the Markov assumption:

$$\begin{aligned}
P(\mathbf{S} \mid q_0 = S_2) &= P(q_1 = S_3 \mid q_0 = S_2) \cdot P(q_2 = S_2 \mid q_1 = S_3) \\
&\quad \cdot P(q_3 = S_1 \mid q_2 = S_2) \cdot P(q_4 = S_1 \mid q_3 = S_1) \\
&\quad \cdot P(q_5 = S_1 \mid q_4 = S_1)
\end{aligned}$$

This corresponds to the coefficients a_{ij}:

$$\begin{aligned}
P(\mathbf{S} \mid q_0 = S_2) &= a_{23} \cdot a_{32} \cdot a_{21} \cdot a_{11} \cdot a_{11} \\
&= 0.1 \cdot 0.3 \cdot 0.4 \cdot 0.8 \cdot 0.8 \\
&= 0.0077
\end{aligned}$$

c) The expectation value for how long we stay on a single state S_i is

$$\begin{aligned} E(x) &= \int x P(x) dx \\ &= \sum_{n=1}^{\infty} n a_{ii}^n (1 - a_{ii}) \\ &= (1 - a_{ii}) \frac{a_{ii}}{(1 - a_{ii})^2} \\ &= \frac{a_{ii}}{1 - a_{ii}} \end{aligned}$$

In the case of sunny days $a_{11} = 0.8$ so we get $\frac{0.8}{0.2} = 4$ days. This is the expectation for staying on the state, so the answer for the number of successive sunny days on avarage is $4 + 1 = 5$.

2. a) In this problem we want to calculate the probability of state S_i in the third day, when we know that it was sunny on the day of the departure and the temperatures on the following three days were $\mathbf{X} = (x_1, x_2, x_3) = (7°C, 3°C, -8°C)$. Let's denote all the model parameters as $\lambda = \{\mathbf{A}, b_i(x)\}$.

$$\begin{aligned} & P(q_3 = S_i \mid q_0 = S_1, x_1, x_2, x_3, \lambda) \\ = & \frac{P(q_3 = S_i, x_1, x_2, x_3 \mid q_0 = S_1, \lambda)}{P(x_1, x_2, x_3 \mid q_0 = S_1, \lambda)} \\ = & \frac{P(q_3 = S_i, x_1, x_2, x_3 \mid q_0 = S_1, \lambda)}{\sum_{j=1}^{3} P(q_3 = S_j, x_1, x_2, x_3 \mid q_0 = S_1, \lambda)} \end{aligned}$$

Above we used first the formula $P(A|B, C) = \frac{P(A,B|C)}{P(B|C)}$. Then we noted that $P(A) = \sum_B P(A, B)$. Denominator and numerator of the equation have similar terms. Let's lighten the notes using the function $\alpha_t(i)$:

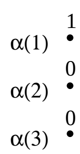
$$P(q_3 = S_i \mid q_0 = S_1, x_1, x_2, x_3, \lambda) = \frac{\alpha_3(i)}{\sum_{j=1}^{3} \alpha_3(j)}$$

Now we will examine how this forward probability $\alpha_t(i)$ can be calculated.

Initial day

We know that it was sunny on the day of the departure, so we can set

$$\begin{aligned} \alpha_0(1) &= 1 \\ \alpha_0(2) &= 0 \\ \alpha_0(3) &= 0 \end{aligned}$$

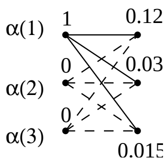


Initialization of the grid for the forward algorithm.

First day

It was sunny on the day before and today it is $7°C$. Let's get the transition probabilities from sunny state and multiply them by the emission probabilities where $x_1 > 5°C$.

$$\begin{aligned}
\alpha_1(1) &= a_{11} \cdot b_1(x \geq 5°C) = 0.8 \cdot 0.15 = 0.120 \\
\alpha_1(2) &= a_{12} \cdot b_2(x \geq 5°C) = 0.15 \cdot 0.2 = 0.030 \\
\alpha_1(3) &= a_{13} \cdot b_3(x \geq 5°C) = 0.05 \cdot 0.3 = 0.015
\end{aligned}$$

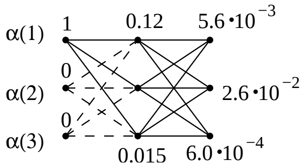


Grid after the first day

Second day

Now we do not know the real weather of the previous day. We sum up all the possibilities weighted by their probabilities:

$$\begin{aligned}
\alpha_2(1) &= \sum_{j=1}^{3} \alpha_1(j) \cdot a_{j1} b_1(-5°C \leq x \leq 5°C) \\
&= (\alpha_1(1) \cdot a_{11} + \alpha_1(2) \cdot a_{21} + \alpha_1(3) \cdot a_{31}) \\
&\quad \cdot b_1(-5°C \leq x \leq 5°C) \\
&= (0.8{\cdot}0.12{+}0.4{\cdot}0.03{+}0.3{\cdot}0.015) \cdot 0.05 \\
&= 5.625 \cdot 10^{-3} \\
\alpha_2(2) &= \sum_{j=1}^{3} \alpha_1(j) \cdot a_{j2} b_2(-5°C \leq x \leq 5°C) \\
&= (0.15{\cdot}0.12{+}0.5{\cdot}0.03{+}0.3{\cdot}0.015) \cdot 0.7 \\
&= 2.625 \cdot 10^{-2} \\
\alpha_2(3) &= \sum_{j=1}^{3} \alpha_1(j) \cdot a_{j3} b_3(-5°C \leq x \leq 5°C) \\
&= (0.05{\cdot}0.12{+}0.1{\cdot}0.03{+}0.4{\cdot}0.015) \cdot 0.4 \\
&= 6.000 \cdot 10^{-4}
\end{aligned}$$

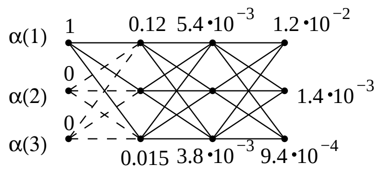


Grid after the second day

Third day

Let's continue as before, now $x_3 < -5°C$.

$$\begin{aligned}
\alpha_3(1) &= \sum_{j=1}^{3} \alpha_2(j) \cdot a_{j1} b_1(-5°C \leq x \leq 5°C) \\
&= (\alpha_2(1) \cdot a_{11} + \alpha_2(2) \cdot a_{21} + \alpha_2(2) \cdot a_{31}) \\
&\quad \cdot b_1(x \leq -5°C) \\
&= (0.8 \cdot 5.625 \cdot 10^{-3} + 0.4 \cdot 2.625 \cdot 10^{-2} \\
&\quad +0.3 \cdot 6.000 \cdot 10^{-4}) \cdot 0.8 \\
&= 1.2144 \cdot 10^{-2} \\
\alpha_3(2) &= \sum_{j=1}^{3} \alpha_2(j) \cdot a_{j2} b_2(-5°C \leq x \leq 5°C) \\
&= (0.15 \cdot 5.625 \cdot 10^{-3} + 0.5 \cdot 2.625 \cdot 10^{-2} \\
&\quad +0.3 \cdot 6.000 \cdot 10^{-4}) \cdot 0.1 \\
&= 1.4149 \cdot 10^{-3} \\
\alpha_3(3) &= \sum_{j=1}^{3} \alpha_2(j) \cdot a_{j3} b_3(-5°C \leq x \leq 5°C) \\
&= (0.05 \cdot 5.625 \cdot 10^{-3} + 0.1 \cdot 2.652 \cdot 10^{-2} \\
&\quad +0.4 \cdot 6.000 \cdot 10^{-4}) \cdot 0.3 \\
&= 9.4387 \cdot 10^{-4}
\end{aligned}$$



The forward grid after the last day

Now we have calculated all the neccessary quantities. Let's insert them to the equation:

$$\begin{aligned}
P(q_3 = S_1 \mid q_0 = S_1, x_1, x_2, x_3, \lambda) &= \frac{\alpha_3(1)}{\sum_{i=j}^{3} \alpha_3(j)} \\
&= \frac{1.2144 \cdot 10^{-2}}{1.2144 \cdot 10^{-2} + 1.4149 \cdot 10^{-3} + 9.4387 \cdot 10^{-4}} \\
&= 0.8874
\end{aligned}$$

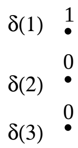
So the probability that it is sunny on the day of the return is 89 %. In a similar manner we can calculate the probability of a cloudy (10 %) and a rainy (1 %) day.

b) This time we try to guess the most probable sequence of weather. In can be done using the Viterbi algorithm, which is very similar to the forward algorithm. The difference is that in each state we now calculate the probability over the best sequence instead of summing up all the sequences.

Viterbi: Initial day

Let's initializate the grid as before.

$$\begin{aligned}\delta_0(1) &= 1\\ \delta_0(2) &= 0\\ \delta_0(3) &= 0\end{aligned}$$

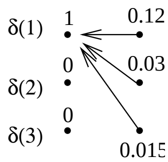


Initialization of the grid for the Viterbi algorithm.

Viterbi: First day

The best path to every state comes from sunny state, as we know the weather, and therefore all the other states have a probability of zero. Let's write down for each state where the most probable path comes from ($\psi_1(i)$). Calculations are still same as in the forward algorithm.

$$\begin{aligned}\delta_1(1) &= a_{11} \cdot b_1(x \geq 5°C) = 0.8 \cdot 0.15 = 0.120\\ \delta_1(2) &= a_{12} \cdot b_2(x \geq 5°C) = 0.15 \cdot 0.2 = 0.030\\ \delta_1(3) &= a_{13} \cdot b_3(x \geq 5°C) = 0.05 \cdot 0.3 = 0.015\\ \psi_1(1) &= 1\\ \psi_1(2) &= 1\\ \psi_1(3) &= 1\end{aligned}$$



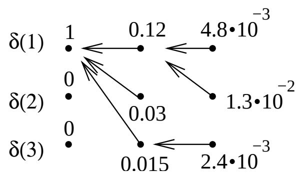
Viterbi search after the first day

Viterbi: Second day

Let's choose the most probable path coming to the each state:

$$\begin{aligned}\delta_2(1) &= \max_j \,(\delta_1(j) \cdot a_{j1} b_1(-5°C \leq x \leq 5°C))\\ &= \max\left(\delta_1(1) \cdot a_{11}, \delta_1(2) \cdot a_{21}, \delta_1(3) \cdot a_{31}\right)\\ &\quad \cdot b_1(-5°C \leq x \leq 5°C)\\ &= \max\left(0.8 \cdot 0.12, 0.4 \cdot 0.03, 0.3 \cdot 0.015\right)\\ &\quad \cdot 0.05\\ &= \max\left(9.6 \cdot 10^{-2}, 1.2 \cdot 10^{-2}, 4.5 \cdot 10^{-3}\right)\\ &\quad \cdot 0.05\\ &= 9.6 \cdot 10^{-2} \cdot 0.05 = 4.8 \cdot 10^{-3}\\ \psi_2(1) &= \underset{j}{\operatorname{argmax}}(\delta_1(j) \cdot a_{j1} b_1(-5°C \leq x \leq 5°C))\\ &= 1\end{aligned}$$

$$
\begin{aligned}
\delta_2(2) &= \max_j \; (\delta_1(j) \cdot a_{j2} b_2(-5°C \le x \le 5°C)) \\
&= \max(0.15 \cdot 0.12, 0.5 \cdot 0.03, 0.3 \cdot 0.015) \\
&\quad \cdot 0.7 \\
&= \max\left(1.8 \cdot 10^{-2}, 1.5 \cdot 10^{-2}, 4.5 \cdot 10^{-3}\right) \\
&\quad \cdot 0.7 \\
&= 1.8 \cdot 10^{-2} \cdot 0.7 = 1.26 \cdot 10^{-2} \\
\psi_2(2) &= \underset{j}{\text{argmax}}(\delta_1(j) \cdot a_{j2} b_2(-5°C \le x \le 5°C)) \\
&= 1 \\
\delta_2(3) &= \max_j \; (\delta_1(j) \cdot a_{j3} b_3(-5°C \le x \le 5°C)) \\
&= \max(0.05 \cdot 0.12, 0.1 \cdot 0.03, 0.4 \cdot 0.015) \\
&\quad \cdot 0.4 \\
&= \max\left(6.0 \cdot 10^{-3}, 3.0 \cdot 10^{-3}, 6.0 \cdot 10^{-3}\right) \\
&\quad \cdot 0.4 \\
&= 6.0 \cdot 10^{-3} \cdot 0.4 = 2.4 \cdot 10^{-3} \\
\psi_2(3) &= \underset{j}{\text{argmax}}(\delta_1(j) \cdot a_{j3} b_2(-5°C \le x \le 5°C)) \\
&= 3
\end{aligned}
$$



Grid after second day

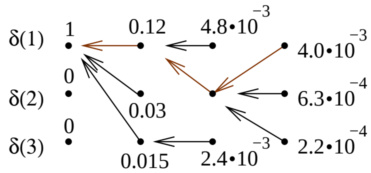
When counting $\psi_2(3)$ we note that the probabilities were same from the states 1 and 3. We can make an arbitrary choice between them. Here the choice was 3.

Viterbi: Third day

Let's choose the most probable path coming to each state:

$$
\begin{aligned}
\delta_3(1) &= \max_j \; (\delta_2(j) \cdot a_{j1} b_1(x \le -5°C)) \\
&= \max(\delta_2(1) \cdot a_{11}, \delta_2(2) \cdot a_{21}, \delta_2(3) \cdot a_{31}) \\
&\quad \cdot b_1(x \le -5°C) \\
&= \max(0.8 \cdot 4.8 \cdot 10^{-3}, 0.4 \cdot 1.26 \cdot 10^{-2}, \\
&\quad 0.3 \cdot 2.4 \cdot 10^{-3}) \cdot 0.8 \\
&= \max\left(3.8 \cdot 10^{-3}, 5.0 \cdot 10^{-3}, 7.2 \cdot 10^{-4}\right) \\
&\quad \cdot 0.8 \\
&= 4.0 \cdot 10^{-3} \\
\psi_3(1) &= \underset{j}{\text{argmax}}(\delta_2(j) \cdot a_{j1} b_1(x \le 5°C)) = 2
\end{aligned}
$$

$$
\begin{aligned}
\delta_3(2) &= \max_j \ (\delta_2(j) \cdot a_{j2} b_2(x \leq -5°C)) \\
&= \max(0.15 \cdot 4.8 \cdot 10^{-3}, 0.5 \cdot 1.26 \cdot 10^{-2}, \\
&\quad 0.3 \cdot 2.4 \cdot 10^{-3}) \cdot 0.1 \\
&= \max \left(7.2 \cdot 10^{-4}, 6.3 \cdot 10^{-3}, 7.2 \cdot 10^{-4}\right) \\
&\quad \cdot 0.1 \\
&= 6.3 \cdot 10^{-3} \cdot 0.1 = 6.3 \cdot 10^{-4} \\
\psi_2(2) &= \operatorname*{argmax}_j(\delta_1(j) \cdot a_{j2} b_2(x \leq -5°C)) = 2 \\
\delta_3(3) &= \max_j \ (\delta_2(j) \cdot a_{j3} b_3(x \leq -5°C)) \\
&= \max(0.05 \cdot 4.8 \cdot 10^{-3}, 0.1 \cdot 1.26 \cdot 10^{-2}, \\
&\quad 0.4 \cdot 2.4 \cdot 10^{-3}) \cdot 0.3 \\
&= \max \left(2.4 \cdot 10^{-4}, 1.3 \cdot 10^{-3}, 7.2 \cdot 10^{-4}\right) \\
&\quad \cdot 0.3 \\
&= 2.16 \cdot 10^{-4} \\
\psi_3(3) &= \operatorname*{argmax}_j(\delta_1(j) \cdot a_{j3} b_2(x \leq -5°C)) = 2
\end{aligned}
$$



Viterbi search after the last day

From the final grid we can get the most probable sequence of states: Let's start from the most probable end state and follow the arrows backwards to the beginning. It seems that it has been sunny, cloudy and again sunny.

Conclusion: The differece between forward and Viterbi algorithms

The forward algorithm gives the correct probability for each state sequence. However, it cannot be used to get the most probable path from the grid.

In Viterbi search, the state probabilities are just approximations. However, it can correctly find the best path.

Computationally both of the algorithms are equally burdensome. Summing in the forward algorithm has just changed to maximization in the Viterbi algorithm.