Tutorial Sections 17-18

1. Consider you have a close friend who lives far away from you, and you two communicate through regular phone conversations. You are aware that your friend exhibits mood patterns that are significantly influenced by the weather. Specifically, they tend to feel happy on sunny days and sad on rainy days. Based on your interactions, you have observed the following:

   - When it is sunny, there is an 80% chance that your friend will report feeling happy during a phone call.
   - Conversely, during rainy days, your friend indicates feeling sad 60% of the time when you inquire about their mood.

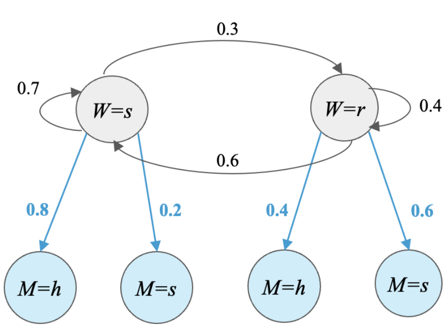
   After monitoring the weather patterns in your friend's location over a period, you have deduced that 75% of the days are sunny, the likelihood of experiencing another sunny day following a sunny day is 70%, and the likelihood of encountering consecutive rainy days is 40%. For simplicity, assume that the weather alternates only between sunny and rainy conditions, and your friend's mood responses are exclusively 'happy' or 'sad'.

   (a) Create a diagram representing the hidden Markov model for this scenario. Your diagram should include both the transition probabilities between weather states and the emission probabilities reflecting your friend's mood responses given the weather conditions.

   (b) Imagine that you called your friend daily over a week, eliciting the following mood responses: happy (Monday), happy (Tuesday), sad (Wednesday), sad (Thursday), sad (Friday), happy (Saturday). Apply the Viterbi algorithm to deduce the most probable sequence of weather conditions (sunny or rainy) for each day, based on the observed sequence of your friend's mood.

Solution: This is a contextualised problem involving the Viterbi algorithm for determining the optimal sequence of hidden states in a hidden Markov model. In this case, the observation is the friend's mood, which can be denoted by the random variable M, and whose outcomes are $m \in \Omega_M = \{h, s\}$ representing 'happy' and 'sad', respectively. The hidden state is the weather, which can be denoted by the RV W, whose outcomes are $w \in \Omega_W = \{s, r\}$ representing 'sunny' and 'rainy', respectively.

(a) Considering the text, we can say that the observed probability of a sunny day is $P(W = s) = 0.75$, and since there are only two possibilities for the weather condition, $P(W = r) = 0.25$. In the absence of a marginal probability for the initial state, we can use the observed probability as the marginal initial state probability. Similarly, we have from the text that $P(W = s|W = s) = 0.7$, and since there are only two possibilities for the transition from $W = s$, the remaining transition probability is $P(W = r|W = s) = 0.3$. Similarly, the text says that $P(W = r|W = r) = 0.4$, which implies that $P(W = s|W = r) = 0.6$.
For the emission probabilities, the problem states that $P(M = h|W = s) = 0.8$, and since there are only two possibilities for the friend's mood, the above probability implies that $P(M = s|W = s) = 0.2$. Similarly, since $P(M = s|W = r) = 0.6$ from the text above, we have $P(M = h|W = r) = 0.4$. Putting all this info in a diagram representing the HMM, we have:



(b) Considering the Viterbi algorithm, the first step is to compute the initial optimal probability function based on the first observation ($M = h$):

$$p_0^\star(W = s) = P(M_0 = h|W_0 = s)P(W_0 = s) = 0.8 \times 0.75 = 0.6$$
$$p_0^\star(W = r) = P(M_0 = h|W_0 = r)P(W_0 = r) = 0.4 \times 0.25 = 0.1.$$

From these initial optimal probabilities, we can use Bellman recursion to find the optimal probability function at $t = 1$, which will be given by:

$$p_1^\star(w) = \max_{w' \in \Omega_W} p_0^\star(w')P(W_1 = w|W_0 = w')P(M_1 = h|W_1 = w)$$

Since $w, w' \in \Omega_W = \{s, r\}$, we can expand the above equation to:

$$\begin{aligned} p_1^\star(W = s) = \max\{&p_0^\star(w = s)P(W_1 = s|W_0 = s)P(M_1 = h|W_1 = s), \\ &p_0^\star(w = r)P(W_1 = s|W_0 = r)P(M_t = h|W_1 = s)\} = \\ &\max\{0.6 \times 0.7 \times 0.8, 0.1 \times 0.6 \times 0.8\} = \max\{0.336, 0.048\} = 0.336 \end{aligned}$$

$$\begin{aligned} p_1^\star(W = r) = \max\{&p_0^\star(w = s)P(W_1 = r|W_0 = s)P(M_1 = h|W_1 = r), \\ &p_0^\star(w = r)P(W_1 = r|W_0 = r)P(M_t = h|W_1 = r)\} = \\ &\max\{0.6 \times 0.3 \times 0.4, 0.1 \times 0.4 \times 0.4\} = \max\{0.072, 0.016\} = 0.072, \end{aligned}$$

so that the optimal state at time $t = 1$ given w is:

$$\begin{aligned} Y_1^\star(W = s) &= \arg\max\{p_0^\star(w = s)P(W_1 = s|W_0 = s), p_0^\star(w = r)P(W_1 = s|W_0 = r)\} \\ &= \arg\max\{0.6 \times 0.7, 0.1 \times 0.6\} = \arg\max\{0.42, 0.06\} = s \end{aligned}$$

$$\begin{aligned} Y_1^\star(W = r) &= \arg\max\{p_0^\star(w = s)P(W_1 = r|W_0 = s), p_0^\star(w = r)P(W_1 = r|W_0 = r)\} \\ &= \arg\max\{0.6 \times 0.3, 0.1 \times 0.4\} = \arg\max\{0.18, 0.04\} = s. \end{aligned}$$

Using the same idea for the following day, in which $M_2 = s$, we have:

$$\begin{aligned} p_2^\star(W = s) = \max\{&p_1^\star(w = s)P(W_2 = s|W_1 = s)P(M_2 = s|W_2 = s), \\ &p_1^\star(w = r)P(W_2 = s|W_1 = r)P(M_2 = s|W_2 = s)\} = \\ &\max\{0.336 \times 0.7 \times 0.2, 0.072 \times 0.6 \times 0.2\} = \max\{0.04704, 0.00864\} = 0.04704 \end{aligned}$$

$$\begin{aligned} p_2^\star(W = r) = \max\{&p_1^\star(w = s)P(W_2 = r|W_1 = s)P(M_2 = s|W_2 = r), \\ &p_1^\star(w = r)P(W_2 = r|W_1 = r)P(M_2 = s|W_2 = r)\} = \\ &\max\{0.336 \times 0.3 \times 0.6, 0.072 \times 0.4 \times 0.6\} = \max\{0.06048, 0.01728\} = 0.06048, \end{aligned}$$

which gives the optimal state $Y_2^\star$,

$$\begin{aligned} Y_2^\star(W = s) = &\arg\max\{p_1^\star(w = s)P(W_2 = s|W_1 = s), p_1^\star(w = r)P(W_2 = s|W_1 = r)\} = \\ &\arg\max\{0.336 \times 0.7, 0.072 \times 0.6\} = \arg\max\{0.2352, 0.0432\} = s \end{aligned}$$

$$\begin{aligned} Y_2^\star(W = r) = &\arg\max\{p_1^\star(w = s)P(W_2 = r|W_1 = s), p_1^\star(w = r)P(W_2 = r|W_1 = r)\} = \\ &\arg\max\{0.336 \times 0.3, 0.072 \times 0.4\} = \arg\max\{0.1008, 0.0288\} = s. \end{aligned}$$

For $M_3 = s$,

$$p_3^\star(W = s) = \max\{0.04704 \times 0.7 \times 0.2, 0.06048 \times 0.6 \times 0.2\} = \max\{0.0065856, 0.0072576\} = 0.0072576$$

$$p_3^\star(W = r) = \max\{0.04704 \times 0.3 \times 0.6, 0.06048 \times 0.4 \times 0.6\} = \max\{0.0084672, 0.0145152\} = 0.0145152,$$

so that $Y_3^\star$ is given by

$$Y_3^\star(W = s) = \arg\max\{0.04704 \times 0.7, 0.06048 \times 0.6\} = \arg\max\{0.032928, 0.036288\} = r$$

$$Y_3^\star(W = r) = \arg\max\{0.04704 \times 0.3, 0.06048 \times 0.4\} = \arg\max\{0.014112, 0.024192\} = r.$$

For $M_4 = s$,

$$p_4^\star(W = s) = \max\{0.0072576 \times 0.7 \times 0.2, 0.0145152 \times 0.6 \times 0.2\} = \max\{0.00101606, 0.001741824\} = 0.001741824$$

$$p_4^\star(W = r) = \max\{0.0072576 \times 0.3 \times 0.6, 0.0145152 \times 0.4 \times 0.6\} = \max\{0.001306368, 0.003483648\} = 0.003483648,$$

so that $Y_4^\star$:

$$Y_4^\star(W = s) = \arg\max\{0.0072576 \times 0.7, 0.0145152 \times 0.6\} = \arg\max\{0.00508032, 0.00870912\} = r$$

$$Y_4^\star(W = r) = \arg\max\{0.0072576 \times 0.3, 0.0145152 \times 0.4\} = \arg\max\{0.00217728, 0.00580608\} = r.$$

For $M_5 = h$,

$$p_5^\star(W = s) = \max\{0.001741824 \times 0.7 \times 0.8, 0.003483648 \times 0.6 \times 0.4\} = \max\{0.00097542144, 0.00083607552\} = 0.00097542144$$

$$p_5^\star(W = r) = \max\{0.001741824 \times 0.3 \times 0.8, 0.003483648 \times 0.4 \times 0.4\} = \max\{0.00041803776, 0.00055738368\} = 0.00055738368,$$

$$Y_5^\star(W = s) = \arg\max\{0.001741824 \times 0.7, 0.003483648 \times 0.6\} = \arg\max\{0.0012192768, 0.0020901888\} = r$$

$$Y_5^\star(W = r) = \arg\max\{0.001741824 \times 0.3, 0.003483648 \times 0.4\} = \max\{0.0005225472, 0.0013934592\} = r.$$

Backtracking, we can reconstruct the best sequence of hidden states as:

$$\begin{aligned} y_5^\star &= \arg\max_{w \in \Omega_W} p_5^\star = s, \\ y_4^\star &= Y_5^\star(y_5^\star) = r \\ y_3^\star &= r, y_2^\star = r, y_1^\star = s, y_0^\star = s \end{aligned}$$

which yields the optimal sequence $[s, s, r, r, r, s]$.

2. One of the applications of RNNs is natural language processing. A common application is sentiment analysis, where the goal is to determine the sentiment (positive, negative, neutral) expressed in text data. In this simplified example, the text data will be sequences of words represented as one-hot encoded vectors (i.e., each word is represented as a vector where one element is 1 (representing the word) and the rest are 0).

   Given a sentence "Good movie", represented as one-hot encoded vectors for "Good" and "movie", calculate the hidden states and output of the RNN for this sequence.

   To solve the problem, let's assume we have a very limited vocabulary based on these two words only, so that "Good" can be represented as $[1, 0]$ and "movie" as $[0, 1]$. For the network, assume the weights for input to the hidden state are given by

$$W_{xh} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix},$$

   from hidden to the next hidden state is $u_{hh} = 0.2$, and from the hidden state to output is $v_{hy} = 0.15$. Also, consider the recurrence following recurrence relation to the problem:

$$\begin{aligned} z_t &= \max\left(0, W^T x_t + u^T z_{t-1}\right) \\ y_t &= \max\left(0, v^T z_t\right) \end{aligned}$$

   Use the following thresholds to classify each sentiment:

   - $\max(y_t) > 0.02$ for positive;
   - $0.01 \leq \max(y_t) \leq 0.02$ for neutral.
   - $0 \leq \max(y_t) < 0.01$ for negative;

Solution: This problem introduces the basic workings of an RNN, including the handling of sequential data, the concept of hidden states, and the production of an output at each step.
To solve the problem using an RNN, let's assume we have a very limited vocabulary where "Good" = [1, 0] and "movie" = [0,1]. Let's also assume the initial hidden state is defined as $z_0 = 0$ in our vocabulary.
For $x_1 = [1, 0]^T$ (one-hot vector for "Good"), we would have:

$$\begin{aligned} z_1 = \max\left(0, W_{xh}^T x_1 + u_{hh} z_0\right) &= \max\left(0, \begin{bmatrix} 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0.2 \cdot 0\right) \\ &= \max(0, 0.1 \cdot 1 + 0.2 \cdot 0 + 0) = \max(0, 0.1) = 0.1 \end{aligned}$$

We don't need to make any prediction after the first word using the RNN, once the goal is to evaluate the whole sentence. However, if we wish to know the intermediate prediction of the RNN output after "Good", it would then be:

$$y_1 = \max(0, v_{hy} \cdot z_1) = \max(0, 0.15 \cdot 0.1) = 0.015,$$

which would fall into the neutral range.
For input "movie", we have $x_2 = [0, 1]^T$ and:

$$\begin{aligned} z_2 = \max\left(0, W_{xh}^T x_2 + u_{hh} z_1\right) &= \max\left(0, \begin{bmatrix} 0.1 & 0.2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 0.2 \cdot 0.1\right) \\ &= \max(0, 0.1 \cdot 0 + 0.2 \cdot 1 + 0.02) = \max(0, 0.22) = 0.22, \end{aligned}$$

which results in the output:

$$y_2 = \max(0, v_{hy} \cdot z_2) = \max(0, 0.15 \cdot 0.22) = 0.033,$$

which provides a sentiment classification of the class "positive".