Hidden Markov Models (HMM)

Devise algorithms for computing the posteriors on a Markov Chain with hidden variables (HMM).
Markov Chains with Discrete Random Variables

Let’s assume we have discrete random variables (e.g., taking 3 discrete values $x_t = \{[1, 0, 0], [0, 1, 0], [0, 0, 1]\})$

Markov Property: $p(x_t | x_1, \ldots, x_{t-1}) = p(x_t | x_{t-1})$

e.g. $p(x_t = [1, 0, 0] | x_{t-1} = [0, 1, 0])$

Stationary, Homogeneous or Time-Invariant if the distribution $p(x_t | x_{t-1})$ does not depend on $t$
Markov Chains with Discrete Random Variables

\[ p(x_1, \ldots, x_T) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1}) \]

What do we need in order to describe the whole procedure?

1. A probability for the first frame/timestamp etc \( p(x_1) \). In order to define the probability we need to define the vector \( \pi = (\pi_1, \pi_2, \ldots, \pi_K) \)

\[ p(x_1 | \pi) = \prod_{c=1}^{K} \pi_c^{x_{1c}} \]

2. A transition probability \( p(x_t | x_{t-1}) \). In order to define it we need a \( K \times K \) transition matrix \( A = [a_{ij}] \)

\[ p(x_t | x_{t-1}, A) = \prod_{j=1}^{K} \prod_{k=1}^{K} a_{jk}^{x_{t-1,j}x_{tk}} \]
Latent Variables

\[ z_t = \begin{bmatrix} z_{t1} \\ z_{t2} \\ z_{t3} \end{bmatrix} \in \{ [1, 0, 0], [0, 1, 0], [0, 0, 1] \} \]

\[ p(X, Z | \theta) = p(x_1, x_2, \ldots, x_T, z_1, z_2, \ldots, z_T | \theta) = \prod_{t=1}^{T} p(x_t | z_t, \theta_x) \prod_{t=1}^{T} p(z_t | \theta_z) \]
Latent Variables in a Markov Chain

\[ p(z_1, \ldots, z_T) = p(z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) \]

\[ p(X, Z|\theta) = p(x_1, x_2, \ldots, x_T, z_1, z_2, \ldots, z_T|\theta) \]

\[ = \prod_{t=1}^{T} p(x_t|z_t, \theta_x) p(z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) \]
Hidden Markov Models

A casino has two dice (our latent variable is probably the dice 😊)

\[ z = \{[1], [0] \} \]

One is fair and one is not

Each die has 6 sides.

\[ x = \{ [0, 1, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0], [0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 1] \} \]

(1) Fair die \( p(x_j = 1|z_1 = 1) = \frac{1}{6} \) for all \( j = \{1, \ldots, 6\} \)

(2) Loaded die \( p(x_j = 1|z_2 = 1) = \frac{1}{10} \) for \( j = \{1, \ldots, 5\} \) and \( p(x_6 = 6|z_2 = 1) = \frac{1}{2} \)

Emission probability
A casino player switches back & forth between fair and loaded die once every 20 turns

(i.e., \( p(z_{t-1} = 1 | z_{t-2} = 1) = p(z_{t} = 1 | z_{t-1} = 1) = 0.05 \))
Hidden Markov Models

Given a string of observations and the above model:
664153216162115234653214356634261655234232315142464156663246

(1) We want to find for a timestamp \( t \) the probabilities of each die given the observations **that far**.

This process is called Filtering: \( p(z_t|x_1, x_2, ..., x_t) \)

(2) We want to find for a timestamp \( t \) the probabilities of each die given the whole string.

This process is called Smoothing: \( p(z_t|x_1, x_2, ..., x_T) \)
(3) Given the observation string find the string of hidden variables that maximize the posterior.

This process is called Decoding (Viterbi).

\[ \arg \max_{y_1 \ldots y_t} p(y_1, y_2, \ldots, y_t | x_1, x_2, \ldots, x_t) \]

(4) Find the probability of the model.

This process is called Evaluation

\[ p(x_1, x_2, \ldots, x_T) \]
Hidden Markov Models

Filtering  Smoothing  Decoding

Taken from Machine Learning: A Probabilistic Perspective by K. Murphy
Hidden Markov Models

(5) Prediction

\[ p(z_{t+\delta} | x_1, x_2, \ldots, x_t) \quad p(x_{t+\delta} | x_1, x_2, \ldots, x_t) \]

(6) Parameter estimation (Baum-Welch algorithm)

\[ A, \pi, \theta \]
Hidden Markov Models

- **Filtering**: $p(z_t|x_1, x_2, ..., x_t)$
- **Viterbi**: $\arg\max_{y_1...y_t} p(y_1, ..., y_t | x_1, ..., x_t)$
- **Prediction**: $p(z_{t+\delta}|x_1, x_2, ..., x_t)$, $p(x_{t+\delta}|x_1, x_2, ..., x_t)$
- **Fixed-lag smoothing**: $p(z_{t-\tau}|x_1, x_2, ..., x_t)$
- **Fixed interval smoothing (offline)**: $p(z_t|x_1, x_2, ..., x_T)$
Hidden Markov Models

\[p(x_1, \ldots, x_T | z_t) = p(x_1, \ldots, x_t | z_t)\]

\[p(x_{t+1}, \ldots, x_T | z_t)\]

\[p(x_1, \ldots, x_{t-1} | x_t, z_t) = p(x_1, \ldots, x_{t-1} | z_t)\]

\[p(x_1, \ldots, x_{t-1} | z_{t-1}, z_t) = p(x_1, \ldots, x_{t-1} | z_{t-1})\]

\[p(x_{t+1}, \ldots, x_T | z_t, z_{t+1}) = p(x_{t+1}, \ldots, x_T | z_{t+1})\]

\[p(x_{t+2}, \ldots, x_T | z_{t+1}, x_{t+1}) = p(x_{t+2}, \ldots, x_T | z_{t+1})\]

\[p(x_1, \ldots, x_T | z_{t-1}, z_t) = p(x_1, \ldots, x_{t-1} | z_{t-1})p(x_t | z_t)\]

\[p(x_{t+1}, \ldots, x_T | z_t)\]

\[p(z_{T+1} | z_T, x_1, \ldots, x_T ) = p(z_{T+1} | z_T )\]
Filtering and smoothing

Filtering: \( p(z_t|x_1, x_2, \ldots, x_t) \)  
Smoothing: \( p(z_t|x_1, x_2, \ldots, x_T) \)

\[
p(z_t|x_1, x_2, \ldots, x_T) = \frac{p(x_1, x_2, \ldots, x_T|z_t)p(z_t)}{p(x_1, x_2, \ldots, x_T)}
\]

\[
= \frac{p(x_1, x_2, \ldots, x_t|z_t)p(x_{t+1}, \ldots, x_T|z_t)p(z_t)}{p(x_1, x_2, \ldots, x_T)}
\]

\[
= \frac{p(x_1, x_2, \ldots, x_t, z_t)p(x_{t+1}, \ldots, x_T|z_t)}{p(x_1, x_2, \ldots, x_T)}
\]

\[
= \frac{\alpha(z_t)\beta(z_t)}{p(x_1, x_2, \ldots, x_T)}
\]
Forward Probabilities

\[ \alpha(z_t) = p(x_1, \ldots, x_t, z_t) \]
\[ = p(x_1, \ldots, x_t | z_t)p(z_t) \]
\[ = p(x_t | z_t)p(x_1, \ldots, x_{t-1} | z_t)p(z_t) \]
\[ = p(x_t | z_t)p(x_1, \ldots, x_{t-1}, z_t) \]
\[ = p(x_t | z_t) \sum_{z_{t-1}} p(x_1, \ldots, x_{t-1}, z_t, z_{t-1}) \]
\[ = p(x_t | z_t) \sum_{z_{t-1}} p(x_1, \ldots, x_{t-1}, z_t | z_{t-1})p(z_t | z_{t-1})p(z_{t-1}) \]
\[ = p(x_t | z_t) \sum_{z_{t-1}} p(x_1, \ldots, x_{t-1} | z_{t-1}, z_t)p(z_t | z_{t-1})p(z_{t-1}) \]
Forward Probabilities

\[
\begin{align*}
= p(x_t | z_t) \sum_{z_{t-1}} p(x_1, \ldots, x_{t-1} | z_{t-1}) p(z_{t-1}) p(z_t | z_{t-1}) \\
= p(x_t | z_t) \sum_{z_{t-1}} p(x_1, \ldots, x_{t-1}, z_{t-1}) p(z_t | z_{t-1}) \\
= p(x_t | z_t) \sum_{z_{t-1}} \alpha(z_{t-1}) p(z_t | z_{t-1})
\end{align*}
\]

Recursive formula and initial condition \( \alpha(z_1) \)

\[
\alpha(z_1) = p(x_1, z_1) = p(z_1) p(x_1 | z_1) = \prod_{k=1}^{2} \left\{ \pi_k p(x_1 | z_{1k}) \right\}^{z_{1k}}
\]
Forward Probabilities

\[ \alpha(z_{t-1\ 1}) \quad \alpha(z_{t\ 1}) \]

\[ \begin{array}{c}
\circ \ 1 \\
\downarrow \quad a_{11} \\
\circ \ 2 \\
\downarrow \quad a_{21} \quad \alpha(z_{t1}) = p(x_t|z_{t1} = 1)(a_{11}\alpha(z_{t-1\ 1}) + a_{21}\alpha(z_{t-1\ 2})) \\
\end{array} \]

\[ \begin{array}{c}
\circ \ 1 \\
\downarrow \quad a_{12} \\
\circ \ 2 \\
\downarrow \quad a_{22} \quad \alpha(z_{t\ 2}) = p(x_t|z_{t\ 2} = 1)(a_{12}\alpha(z_{t-1\ 1}) + a_{22}\alpha(z_{t-1\ 2})) \\
\end{array} \]

In total \( O(2^2) \) computation (in general \( O(K^2) \))
Filtering

\[ p(x_1, ..., x_t) = \sum_{z_t} p(x_1, ..., x_t, z_t) = \sum_{z_t} \alpha(z_t) = \alpha(z_{t1}) + \alpha(z_{t2}) \]

Filtering \[ p(z_t | x_1, ..., x_t) = \frac{p(x_1, ..., x_t, z_t)}{p(x_1, ..., x_t)} = \frac{\alpha(z_t)}{\sum_{z_t} \alpha(z_t)} = \tilde{a}(z_t) \]

Evaluation: How can we compute \[ p(x_1, ..., x_T) \]?

\[ p(x_1, ..., x_T) = \sum_{z_T} p(x_1, ..., x_T, z_T) = \sum_{z_T} \alpha(z_T) \]
Backward probabilities

\[
\beta(z_t) = p(x_{t+1}, ..., x_T | z_t) \\
= \sum_{z_{t+1}} p(x_{t+1}, ..., x_T, z_{t+1} | z_t) \\
= \sum_{z_{t+1}} p(x_{t+1}, ..., x_T | z_t, z_{t+1}) p(z_{t+1} | z_t) \\
= \sum_{z_{t+1}} p(x_{t+1}, ..., x_T | z_{t+1}) p(z_{t+1} | z_t) \\
= \sum_{z_{t+1}} p(x_{t+2}, ..., x_T | z_{t+1}) p(x_{t+1} | z_{t+1}) p(z_{t+1} | z_t) \\
= \sum_{z_{t+1}} \beta(z_{t+1}) p(x_{t+1} | z_{t+1}) p(z_{t+1} | z_t)
\]
Computing Backward probabilities

\[
\beta(z_{t+1}) = \beta(z_{t+1}) a_{11} p(x_t | z_{t+1} = 1) + \beta(z_{t+1}) a_{12} p(x_t | z_{t+1} = 1)
\]
Computing Backward probabilities

\[ \begin{align*}
\beta(z_{t,1}) & \quad \beta(z_{t+1,1}) \\
1 & \quad 1 \\
\xrightarrow{a_{21}} & \\
2 & \quad 2 \\
\xleftarrow{a_{22}} & \\
\beta(z_{t,2}) & \quad \beta(z_{t+1,2})
\end{align*} \]

\[ p(x_t|z_{t+1,1} = 1) = \beta(z_{t,1})a_{21}p(x_t|z_{t+1,1} = 1) + \beta(z_{t+1,2})a_{22}p(x_t|z_{t+1,2} = 1) \]

\[ p(z_T|x_1, \ldots, x_T) = \frac{p(x_1, \ldots, x_T, z_T)}{p(x_1, \ldots, x_T)} = \frac{\alpha(z_T) \ast 1}{p(x_1, \ldots, x_T)} \Rightarrow \beta(z_T) = 1 \]


Prediction

\[ p(z_{t+2}|x_1, \ldots, x_t) = \sum_{z_{t+1}} \sum_{z_t} p(z_{t+2}, z_{t+1}, z_t|x_1, \ldots, x_t) \]

\[ = \sum_{z_{t+1}} \sum_{z_t} p(z_{t+2}, z_{t+1}|x_1, \ldots, x_t, z_t) p(z_t|x_1, \ldots, x_t) \]

\[ = \sum_{z_{t+1}} \sum_{z_t} p(z_{t+2}, z_{t+1}|z_t) p(z_t|x_1, \ldots, x_t) \dagger(a(z_t)) \]

\[ = \sum_{z_{t+1}} \sum_{z_t} p(z_{t+2}|z_{t+1}) p(z_{t+1}|z_t)a(z_t) \]

\[ \Rightarrow A^T A^T \tilde{a} \]
Prediction

and

\[ p(x_{t+2}|x_1, ..., x_t) = \sum_{z_{t+2}} p(x_{t+2}|z_{t+2}) \ p(z_{t+2}|x_1, ..., x_t) \]
Smoothed transition

\[ \xi(z_{t-1}, z_t) = p(z_{t-1}, z_t|x_1, ..., x_T) \]

\[
= \frac{p(x_1, ..., x_T|z_{t-1}, z_t)p(z_{t-1}, z_t)}{p(x_1, ..., x_T)} \\
= \frac{p(x_1, ..., x_{t-1}|z_{t-1})p(x_t|z_t)p(x_{t+1}, ..., x_T|z_t)p(z_t|z_{t-1})p(z_{t-1})}{p(x)} \\
= \frac{\alpha(z_{t-1})p(x_t|z_t)p(z_t|z_{t-1})\beta(z_t)}{p(x_1, ..., x_T)}
\]
Parameter Estimation (Baum-Welch algorithm)

We need to define an EM algorithm

and we have all the necessary ingredients 😊

\[ p(z_t | x_1, x_2, ..., x_T) \quad p(z_{t-1}, z_t | x_1, ..., x_T) \]
Summary

We saw how to perform
(a) Filtering
(b) Smoothing
(c) Evaluation
(d) Prediction

Next we will see how to perform (e) EM and (f) Decoding