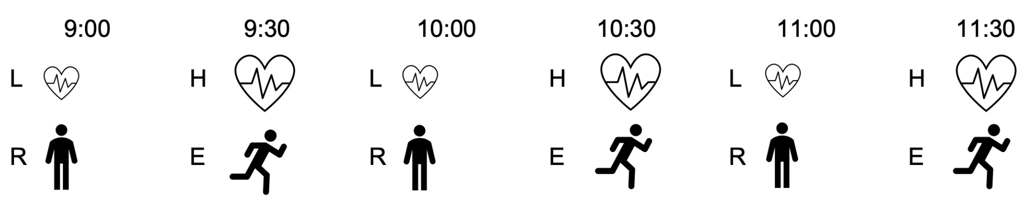
Week 11 - Sequence modelling, Hidden Markov models

- Sequence modelling (data in which ordering matters)

#Code

Sequence modelling: intuition

- **Problem**: Smartwatch-based Activity Monitoring System
- **Measured observation (X_t):** heart rate (high vs low)
  - $X_t \in \Omega_X = \{h, l\}$
- **Inferred observation (Y_t):** activity (rest vs. exercise)
  - $Y_t \in \Omega_Y = \{r, e\}$
  - It is a **hidden state** (not directly observable)

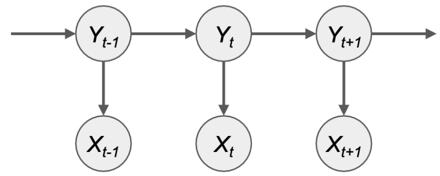


- Assumption:
  - we aren't randomly on rest or exercise;
  - If we are at rest at a given time, it's likely we will continue at rest
- Hidden Markov models （HMMs)
  - Sequences through a series of (discrete) hidden states that are not directly observable, but follows a certain probability distribution

#Code

Sequence modelling: Hidden Markov Models (HMMs)

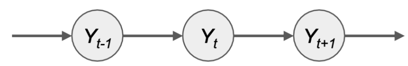
- The **hidden Markov model** (HMM) captures time-dependent RVs which are not directly measured
- Each **hidden states** $Y_t \in \Omega_Y$ with K distinct values, depends only upon the one before it in time, Y_{t-1} for all $t = 0,1,\ldots, T$
- The measured **observations** X_t depend only upon the associated hidden state, Y_t



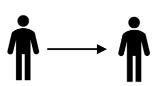
#Code

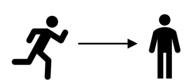
Sequence modelling: model fitting

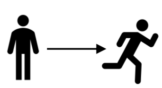
- given observed data for $X_0, X_1, \ldots, X_T$, estimate the distribution functions $P(X_t|Y_t),\ P(Y_t|Y_{t-1})$

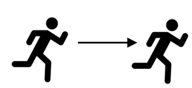


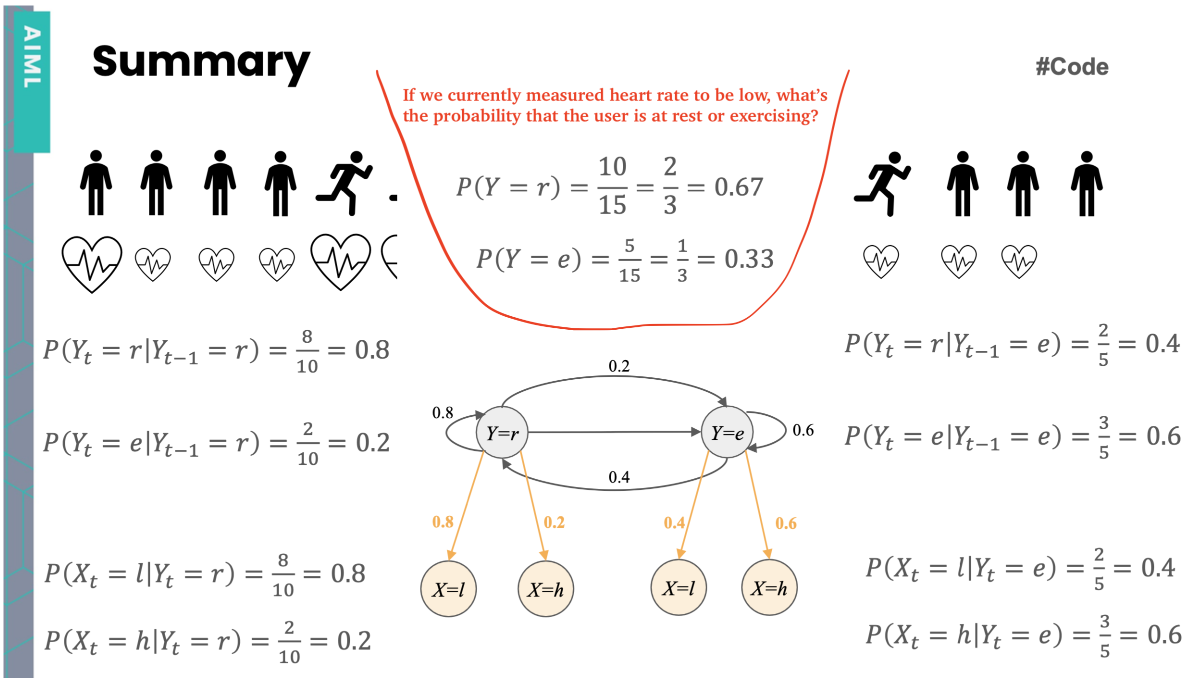
- Training data (**transition probabilities**)

 $P(Y_t = r|Y_{t-1} = r) = \frac{8}{10} = 0.8$

 $P(Y_t = r|Y_{t-1} = e) = \frac{2}{5} = 0.4$

 $P(Y_t = e|Y_{t-1} = r) = \frac{2}{10} = 0.2$

 $P(Y_t = e|Y_{t-1} = e) = \frac{3}{5} = 0.6$

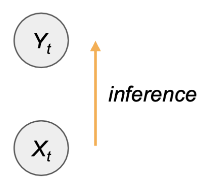


- **Encoding (focus on this)**

Sequence modelling: single evaluation

#Code

- given fixed model parameters and observed data, compute the probability of the hidden state



- If we currently measured heart rate to be low, what's the probability that the user is at rest or exercising?

$$P(Y=r)=\frac{10}{15}=\frac{2}{3}=0.67$$

$$P(Y=e)=\frac{5}{15}=\frac{1}{3}=0.33$$

Bayes' Theorem:

The symbol ∝ represents "proportional to" in mathematics.
We read x ∝ y as "x is directly proportional to y."

$$P(Y=r|X=l)\propto \underline{P(X=l|Y=r)}P(Y=r)=0.8\times 0.67=0.536$$

$$P(Y=e|X=l)\propto \underline{P(X=l|Y=e)}P(Y=e)=0.4\times 0.33=0.132$$

Decision: $y^{*}=\arg\max_{y\in\Omega_Y} P(X|Y)P(Y=y)=r$

HMM sequence modelling problems

#Code

- In applications of HMMs, typically need to solve the following problems
  - **Model fitting**: given observed data for $X_0, X_1, \ldots, X_T$, estimate the distribution functions
X_t=value, Y_t situation $P(X_t|Y_t)$, $P(Y_t|Y_{t-1})$; Y_t=current state, Y_t-1=previous situation
  - **Evaluation**: given fixed model parameters and observed data, compute the probability of the data, $P(X)$;
  - **Decoding**: given fixed model parameters and data compute the most probable sequence of hidden states $y = [y_0^*, y_1^*, y_2^*, \ldots, y_T^*]$.
- Solving these problems requires evaluating **all possible sequences of hidden states**; if there are K hidden states, this requires $O(K^T)$ (exponential complexity)
- Use of **dynamic programming** makes this tractable in order $O(TK^2)$.

Bellman recursion for optimal sequence probability

#Code

- Reading off PGM, at time step $t-1$, optimal sequence probability:

$$P^\star(X_0, \ldots, X_{t-1}, Y_{t-1}) = \max_{y' \in \mathcal{Y}_{t-2}} P(X_0, \ldots, X_{t-1}, Y_0 = y'_0, \ldots, Y_{t-2}, Y_{t-1})$$

  where $\mathcal{Y}_{t-2}$ is set of all possible state sequences, up to time $t-2$.

- Optimal sequence probability, as a function of y up to time t,

$$p_t^\star(y) = P^\star(X_0, \ldots, X_t, Y_t = y)$$

  is obtained using **Bellman recursion**,

$$p_t^\star(y) = \max_{y' \in \Omega_Y} [p_{t-1}^\star(y')\, P(Y_t = y | Y_{t-1} = y') P(X_t = x_t | Y_t = y)]$$

- Decoding (won't be tested) - Viterbi algorithm

HMM Viterbi decoding: algorithm

#Code

- **Step 1**. *Initialization*: Compute the initial optimal probability function,
$$p_0^\star(y) = P(X_0 = x_0 | Y_0 = y)\, P(Y_0 = y)$$

- **Step 2**. *Forward recursion*: Sequence of optimal probability functions,
$$p_t^\star(y) = \max_{y' \in \Omega_Y} p_{t-1}^\star(y')\, P(Y_t = y | Y_{t-1} = y')\, P(X_t = x_t | Y_t = y)$$
for $t = 1, 2, \ldots, T$, keeping track of the corresponding decision,
$$Y_t^\star(y) = \arg\max_{y' \in \Omega_Y} p_{t-1}^\star(y')\, P(Y_t = y | Y_{t-1} = y')$$

- **Step 3**. *Backtrack*: Find optimal sequence in reverse, for $t = T-1, T-2, \ldots, 1$,
$$y_T^\star = \arg\max_{y \in \Omega_Y} p_T^\star(y),\; y_{t-1}^\star = Y_t^\star(y_t^\star)$$

- Emission probabilities
- Transition probabilities

Understand

$$P_0^* = P(X_0 = x_0 | Y_0 = y) P(Y_0 = y)$$

- Recurrent Neural Network - RNN
- Transformer
  - Self-attention
    - Contextual information
  - Multi-head attention