HMM, MEMM, CRF

PRELIMINARIES

• MARKOV CHAINS

Probabilistic models that represent a sequence of random variables $\{q_t\}_{t=1,\dots,T}$ satisfying Markov property.

→ row-stochastic transition matrix $\underline{\underline{A}} \in [\emptyset; 1]^{N \times N}$

→ set of states $S = \{s_1, \dots, s_N\}$

→ initial probability distribution $\underline{\pi} = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_N \end{bmatrix}$

• MIXTURE MODELS

Distribution models that arise from convex combinations of simple distributions.

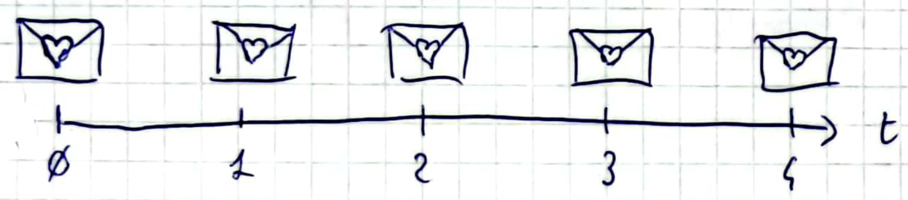
→ probability distributions $b_i(x)$

→ observations x

TOTAL PROBABILITY THEOREM $\Longrightarrow \mathbb{P}(x) = \sum_i \mathbb{P}(x \mid s_i)\, \mathbb{P}(s_i) = \sum_i b_i(x)\, p_i$

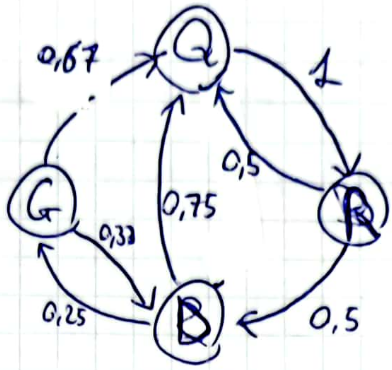
HIDDEN MARKOV MODELS

Imagine a sequence of love letters (or messages):



Each message can ~~be~~ contain a question, can reply, a general consideration, or a love declaration.

$S = \{ Q, G, D, R \}$



Question → surely Reply

Reply → maybe another Question, maybe a random Declaration

Declaration → likely a "do you really love me" Question, or a General statement

General → Question more likely than random Declaration

Problem: we can't observe the state itself, but we can measure the characteristics of the messages.

$$\underline{X}_t = \begin{bmatrix} \text{length} \\ \text{count}(\text{"love"}) \\ \text{count}(\text{"you"}) - \text{count}(\text{"I"}) \\ \text{count}(\text{"?"}) \\ \vdots \\ \text{TFIDF}(\text{some words}) \end{bmatrix}$$

FINAL GOAL:

WE WANT TO ESTIMATE THE HIDDEN STATE (Q, R, D OR G) OF EACH MESSAGE GIVEN THE FEATURES $\underline{X}_t$ OF THE MESSAGE.

Lawrence Rabiner (1989) systematised tasks on such problems:

1. LIKELIHOOD: $\mathbb{P}(\{X_t\}_t \mid \lambda)$ → parameters $\underline{\underline{A}}, \underline{\pi}, b$

3. DECODING: $\{q_t^*\}_t = \text{argmax}\ \mathbb{P}(\{q_t\}_t \mid \{X_t\}_t, \lambda)$

2. LEARNING: $\lambda^* = \text{argmax}\ \mathbb{P}(\lambda \mid \{\underline{X}_t\}_t)$

① LIKELIHOOD (model $\underset{\underline{A}, b(\cdot), \pi, \{x_t\}, S}{}$ ⟶ prob. of observed sequence $\mathbb{P}(\{x_t\} | \lambda)$)

Recall:

- X_t depends only on q_t
- q_t depends only on q_{t-1} (Markov property)

If we had $\{q_t\}_{t=1,\dots,T}$, then:

$$\mathbb{P}(\{X_t\} | \{q_t\}, \lambda) = \prod_{t=1}^{T} \mathbb{P}(X_t | q_t \ \lambda)$$

but we don't know them. So:

$$\mathbb{P}(\{X_t\}, \{q_t\} | \lambda) = \mathbb{P}(\{X_t\} | \{q_t\}, \lambda) \cdot \mathbb{P}(\{q_t\} | \lambda) =$$

$$= \prod_{t=1}^{T} \mathbb{P}(X_t | q_t, \lambda) \prod_{t=1}^{T} \mathbb{P}(q_t | q_{t-1}, \lambda)$$

for each possible hidden state sequence $\{q_t\}$. $\# = N^T$

⇓

INTRACTABLE

DYNAMIC PROGRAMMING FRAMEWORK

→ Break down a complex problem in overlapping sub-problems that can be solved recursively.

FORWARD ALGORITHM

Let $\alpha_t(j) = \mathbb{P}(q_t = s_j, x_t, \dots, x_1 | \lambda) =$

$$= \sum_i \alpha_{t-1}(i) A_{ij} \, b_j(x_t) =$$

$$= \sum_i \mathbb{P}(q_{t-1} = s_i, x_{t-1}, \dots, x_1 | \lambda) \, \mathbb{P}(q_t = s_j | q_{t-1} = s_i) \, \mathbb{P}(X_t | q_t = s_j$$

since $\mathbb{P}(q_t = S_j, x_t, x_{t-1}, \dots, x_1) =$

$$= \sum_i \mathbb{P}(q_t = S_j, q_{t-1} = S_i, x_t, x_{t-1}, \dots, x_1) =$$

$$= \sum_i \mathbb{P}(q_t = S_j, \quad x_t \mid q_{t-1} = S_i, x_{t-1}, \dots, x_1) \cdot \mathbb{P}(\underbrace{\quad}_{q_{t-1}=S_i, x_{t-1}, \dots}) =$$

$$= \sum_i \mathbb{P}(q_t = S_j, x_t \mid q_{t-1} = S_i) \; \alpha_{t-1}(i)$$

At the first step $\alpha_1(j) = \pi_j b_j(x_1)$.

In the end:

$$\mathbb{P}(\{x_t\}_{t=1,\dots,T} \mid \lambda) = \sum_{j=1}^{N} \alpha_T(j)$$

③ DECODING $\left(\underset{\underline{\underline{A}}, \underline{\pi}, \underline{b}(x)}{\text{model}}, \underset{\{x_t\}}{\text{observed sequence}}, \underset{S}{\overset{\text{state set}}{}} \right) \longrightarrow \underset{\{q_t\}}{\text{hidden state sequence}}$

Naive solution: for each of the N^T possible state sequences $\{q_t\}_{t=1\dots T}$ run the Forward Algorithm and take the sequence with highest likelihood.

$\longrightarrow$ INTRACTABLE

$\boxed{\text{VITERBI ALGORITHM}}$

Let $$\delta_t(j) = \max_{q_1, \dots, q_{t-1}} \mathbb{P}(q_1, \dots, q_{t-1}, q_t = S_j, x_1, \dots, x_t \mid \lambda) =$$

$$= \max_{q_1, \dots, q_{t-1}} \underbrace{\mathbb{P}(x_t, q_t = S_j \mid q_{t-1} \cdots q_1 \; x_{t-1}, \dots, x_1)}_{\substack{\text{doesn't depend on } q_1, \dots, q_{t-2} \text{ for} \\ \text{Markov property and since } \underline{\underline{A}} \text{ and } b(\cdot) \text{ are given}}} \mathbb{P}(q_1 \cdots q_{t-1}, x_1, \dots, x_{t-1})$$

$$\Rightarrow \delta_t(j) = \max_i \big(\underbrace{A_{ij}}_{\mathbb{P}(q_t = S_j \mid q_{t-1} = S_i)} \delta_{t-1}(i) \big) \underbrace{b_j(x_t)}_{\mathbb{P}(x_t \mid q_t, S_j)}$$

At the first step: $\delta_1(j) = \pi_j b_j(x_1)$

By iteratively computing $\delta_t(j)$ and the state S_{j^*} maximizing it, it is possible to reconstruct the optimal state sequence (not in the forward run, but in BACKTRACKING).

$$\begin{cases} q_T^* = \arg\max_j \delta_T(j) \\ q_t^* = \psi_{t+1}(q_{t+1}^*) \quad \text{for } t < T \end{cases}$$

where $\psi_{t+1}(j) = \arg\max_i \big(\delta_t(i) A_{ij} \big)$

↳ notice it's outgoing, while in the definition of δ it was incoming

② LEARNING (observed $\{x_t\}$ sequence, $\underset{S}{\text{state set}} \longrightarrow$ model parameters) $\underline{\underline{A}}, \underline{\pi}, \underline{b}(\cdot)$

Unsupervised learning (even supervised learning is possible).

$\longrightarrow$ learning $\underline{\underline{A}}$, $\underline{b}(\cdot)$ and $\underline{\pi}$

Naive solution: Knowing the state sequence, we would be able to estimate parameters immediately.

$\longrightarrow$ but we don't know $\{q_t\}$

Scansionato con CamScanner

Let's define:

$\boxed{\text{BACKWARD PROBABILITY}}$

$$\beta_t(j) = \mathbb{P}(x_T, \dots, x_{t+1} \mid q_t = j, \lambda) \quad \text{with } \beta_T(j) = 1 \ \forall s_j$$

Note: forward probability $\alpha_t(j)$ is the probability of a certain state and observations up to it; backward prob. is the probability of observations after a certain state given it.

Recursively:

$$\beta_t(j) = \mathbb{P}(x_T, \dots, x_{t+1} \mid q_t = j, \lambda) =$$

$$= \sum_{i=1}^{N} \mathbb{P}(x_T, \dots, x_{t+1} \mid q_{t+1} = i, \cancel{q_t = j}, \lambda)\, \mathbb{P}(q_{t+1} = i \mid q_t = j)$$

doesn't count anymore

$$= \sum_{i=1}^{N} \mathbb{P}(x_T, \dots, x_{t+2} \mid q_{t+1} = i, \lambda)\, \mathbb{P}(x_{t+1} \mid q_{t+1} = i, \lambda)\, A_{ji}$$

$$= \sum_{i=1}^{N} \beta_{t+1}(i)\, b_i(x_{t+1})\, A_{ji}$$

Maximum likelihood estimation in an iterative way:

$\boxed{\text{EXPECTATION-MAXIMIZATION ALGORITHM}}$

E-STEP (EXPECTATION):

use current estimates $\hat{\underline{A}}, \hat{\underline{b}}, \hat{\underline{\pi}}$ to compute

$$\mathbb{P}(q_t = s_i, q_{t+1} = s_j \mid \{x_t\}) =: \xi_t(i,j)$$

$$\xi_t(i,j) = \frac{\mathbb{P}(q_t = s_i, q_{t+1} = s_j, x_1, \dots, x_T)}{\mathbb{P}(\{x_t\}_t)}$$

where $\mathbb{P}(q_t = s_i, q_{t+1} = s_j, x_1, \dots, x_T) =$

$$= \mathbb{P}(x_1, \dots, x_T \mid q_{t+1} = s_j, q_t = s_i)\, \mathbb{P}(q_{t+1} = s_j, q_t = s_i)$$

given q_t and q_{t+1}, we can split the sequence in independent segments

$$= \mathbb{P}(x_1, \dots, x_t \mid \cancel{q_{t+1} = s_j}, q_t = s_i)\, \mathbb{P}(x_{t+1} \mid q_{t+1} = s_j, \cancel{q_t = s_i})$$
$$\cdot\, \mathbb{P}(x_{t+2}, \dots, x_T \mid q_{t+1} = s_j, \underbrace{\cancel{q_t = s_i}}_{\text{irrelevant}})\, \mathbb{P}(q_{t+1} = s_j, q_t = s_i)$$

$$= \mathbb{P}(x_1, \dots, x_t \mid q_t = s_i)\, b_j(x_{t+1})\, \beta_{t+1}(j)$$
$$\cdot\, \mathbb{P}(q_{t+1} = s_j \mid q_t = s_i)\, \mathbb{P}(q_t = s_i)$$

$$= \alpha_t(i)\, b_j(x_{t+1})\, \beta_{t+1}(j)\, A_{ij}$$

↳ with $\xi_t(i,j)$ we can estimate $\hat{A}_{ij}$

Then, we define:

$$\gamma_t(j) = \mathbb{P}(q_t = s_j \mid \{x_t\}) =$$
$$= \frac{\mathbb{P}(q_t = s_j, x_1, \dots, x_T)}{\mathbb{P}(x_1, \dots, x_T)} = \frac{\alpha_t(j)\, \beta_t(j)}{\mathbb{P}(x_1, \dots, x_T)}$$

↳ with $\gamma_t(j)$ we can estimate $\hat{b}_j$

M-STEP

Update parameters:

$$\hat{A}_{ij} = \frac{\mathbb{E}\left[\sum_{t=1}^{T-1} \mathbb{1}\{q_{t+1}=s_j,\ q_t=s_i\} \,\middle|\, \{x_t\}\right]}{\mathbb{E}\left[\sum_{t=1}^{T-1} \mathbb{1}\{q_t=s_i\} \,\middle|\, \{x_t\}\right]} = \frac{\sum_{t=1}^{T-1} \mathbb{P}(q_{t+1}=s_j,\ q_t=s_i \mid \{x_t\})}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \mathbb{P}(q_{t+1}=s_j,\ q_t=s_i \mid \{x_t\})}$$

probability of transitioning from s_i to s_j

$$= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \xi_t(i,j)}$$

$$\hat{b}_j(x) = \frac{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{x_t=x;\ q_t=s_j\} \,\middle|\, \{x_t\}\right]}{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{q_t=s_j\} \,\middle|\, \{x_t\}\right]} =$$

same symbol

$$= \frac{\sum_{t=1}^{T} \mathbb{P}(x_t=x;\ q_t=s_j \mid \{x_t\})}{\sum_{t=1}^{T} \mathbb{P}(q_t=s_j \mid \{x_t\})} = \frac{\sum_{t=1}^{T} \gamma_t(j)\, \overbrace{\mathbb{P}(x_t=x \mid \{x_t\})}^{\text{either } \emptyset \text{ or } 1}}{\sum_{t=1}^{T} \gamma_t(j)}$$

Thorough discussion about EM is complex, but in general it is about maximizing expected likelihood (expected since we don't know hidden states) iteratively: