HMM-Based Clinical Note Abstraction Project

Project Definition

Healthcare professionals spend countless hours reviewing lengthy clinical notes to extract critical patient information. This project aims to develop an automated system using Hidden Markov Models to identify and extract key medical events, diagnoses, and treatments from unstructured physician notes.

The system would process raw clinical text such as "Patient presented with chest pain and shortness of breath. History of hypertension and diabetes. Physical exam revealed elevated blood pressure 160/95. EKG showed ST elevation. Administered aspirin and transferred to cardiac catheterization lab. Angioplasty performed successfully. Patient stable post-procedure." From this narrative, the model would generate a structured abstract organizing information into categories like Chief Complaint (chest pain, shortness of breath), Medical History (hypertension, diabetes), Key Findings (elevated BP, ST elevation), Procedures (angioplasty), and Outcome (stable post-procedure).

The HMM approach models clinical notes as sequences where hidden states represent six different types of clinical information being documented: chief complaint, medical history, physical examination findings, clinical assessment, treatment interventions, and patient outcomes. The goal is to automatically discover what type of clinical information is being discussed at each point in the medical narrative.

HMM Analysis

1. Describe the Observations: What measurable data would the model use?

The model would process several types of measurable features from clinical text:

- **Medical terminology**: Disease names ("pneumonia"), medications ("metformin"), procedures ("angioplasty")
- **Clinical indicators**: Phrases like "complains of" (symptoms), "diagnosed with" (assessments), "prescribed" (treatments)
- **Temporal markers**: "on admission," "post-operative," "at discharge"
- **Quantitative values**: Vital signs ("120/80"), dosages ("5mg twice daily")
- **Supporting features**: Word position, medical entity labels, sentence structure, section headers

2. Type of HMM Problem: If you don't know the hidden states in advance, what kind of HMM task is this?

This represents an **unsupervised hidden state discovery problem** since we lack pre-labeled training sequences indicating which words belong to specific clinical information categories. The model must independently learn to recognize patterns in clinical documentation and identify different types of medical information without explicit supervision. This makes it a structure discovery task where the algorithm uncovers latent organizational patterns in how physicians document patient care.

3. Training Algorithm:

a. What values are known at the start?

- **Clinical note text sequences**: Raw medical documentation
- **Vocabulary size**: Number of unique medical terms in dataset
- **Number of hidden states**: Six clinical information types (predetermined)
- **Medical domain knowledge**: Clinical workflows and terminology

b. What values are unknown and need to be learned?

- **Initial state probabilities (π)**: Which clinical information types typically begin medical notes
- **State transition probabilities (A)**: How clinical information flows between categories
- **Emission probabilities (B)**: Which words characterize each clinical information type
- **Hidden state sequences**: Optimal progression of clinical information types for each note

4. Parameter Updates: Which HMM parameters will your training algorithm update?

The Baum-Welch algorithm will systematically update three essential parameter sets:

- **π (Initial states)**: 6 values capturing likelihood each clinical information type starts a note
- **A (Transitions)**: 6×6 matrix (36 values) modeling information flow between categories
- **B (Emissions)**: 6×vocabulary size matrix determining word probabilities for each state

The training process follows an iterative approach beginning with random parameter initialization. The expectation step calculates the probability of being in each state at every position within the text sequence. The maximization step updates the initial, transition, and emission parameters based on these calculated probabilities. This cycle repeats until the model converges, meaning the likelihood of the training data stops improving significantly.

Success will be measured through clinical accuracy validation by medical professionals, assessment of information coverage representing the percentage of key clinical facts

captured, quantification of time savings for healthcare professionals, and evaluation of extraction consistency across different physicians' documentation styles. This approach combines the interpretability advantages of Hidden Markov Models with practical healthcare applications, creating both an academically sound and clinically relevant machine learning project.