Applying Hidden Markov Models in My Capstone Project

Project Overview

My capstone project is a web-based platform designed to help businesses automate the discovery of relevant and timely information about other companies they wish to collaborate with. Users enter details about their own business (for example, offering insurance services) and provide a list of target companies. The platform scrapes online sources for news and updates related to those companies, analyzes the content using **NLP**, and generates actionable insights such as recommendations on the right time and approach for outreach.

How HMM Will Be Used

A **Hidden Markov Model (HMM)** will be used to track how a company's business focus changes over time. For example, a company might start in logistics, shift to healthcare, and later expand into agriculture. These shifts aren't directly labeled but can be inferred from patterns in news articles. The HMM models these hidden "**business focus states**" and learns how companies transition between them. Unlike static company profiles, this approach allows the platform to predict future directions and support **timely, targeted outreach** based on evolving business trends.

Describe the Observations

The observations are sequences of **news headlines or article summaries** collected about a company in chronological order. These texts are transformed into numerical vectors using methods such as **TF-IDF** or **sentence embeddings** to allow the model to process them. For example, a company's news might include: "XYZ expands fleet operations," "XYZ partners with health NGOs," and "XYZ launches agricultural delivery pilot," each forming an observation in the time series.

Type of HMM Problem

This is an **unsupervised learning problem**, corresponding to the **second fundamental problem of HMMs**. The model does not receive labeled hidden states such as "healthcare focus" or "logistics." Instead, it must learn these hidden states, the transitions between them, and how likely each observed news item is generated from a particular state.

Training Algorithm

The model will be trained using the **Baum-Welch algorithm**, a variant of the **Expectation-Maximization (EM)** algorithm for HMMs. At the start, the **known values** include the observed sequences of news vectors and the number of hidden states chosen (e.g., 3 to 5 business focus areas). The **unknown values** to be learned during training are the hidden states themselves, the transition probabilities between states, the emission probabilities of observations given states, and the initial state probabilities.

Parameter Updates

During training, the Baum-Welch algorithm iteratively updates the following parameters: the **transition matrix (A)** — probabilities of moving from one hidden state to another, the **emission matrix (B)** — probabilities of producing specific observations from a given hidden state, and the **initial state distribution (π)**, probabilities of starting in each hidden state. These updates refine the model's ability to represent the evolution of a company's business focus.

Conclusion

By applying an HMM, the platform gains the capability to not only describe a company's current state but also to **anticipate future strategic directions**. This predictive insight transforms the tool from a reactive research aid into a **proactive business intelligence engine**, allowing users to target outreach efforts with greater relevance and timing.