Comparing 2025-Era Small LLMs and HMMs for Time Series Forecasting

CS/DS-GA 1170 – Fundamental Algorithms Project Proposal

Group Members: Zeyuan Ding (zd2466@nyu.edu), Keyu Hong (kh4300@nyu.edu)

1. Research Question

This project compares recent compact Large Language Models (LLMs) and classical Hidden Markov Models (HMMs) on a real-world time series forecasting task. We ask: *Can new-generation small LLMs, when treating numerical sequences as text tokens, outperform HMMs in short- and medium-term air quality forecasting?* We investigate whether the contextual reasoning and pattern generalization of modern LLMs (e.g., Phi-3-mini, Llama-3.2, Qwen2.5) enable superior predictive accuracy over the probabilistic temporal structure captured by HMMs.

2. Dataset

We use the **Beijing PM2.5 Air Quality Dataset**[1], containing hourly PM2.5 concentrations and meteorological variables (temperature, humidity, wind speed, etc.) collected between 2010 and 2014. After cleaning and normalization, PM2.5 levels serve as the prediction target, while the other features provide auxiliary context.

3. Methods and Algorithms

Hidden Markov Model (HMM). A Gaussian-emission HMM will be trained to identify latent pollution regimes and forecast future PM2.5 levels through learned state transitions and emission probabilities.

Large Language Model (LLM). We transform numeric time series into tokenized sequences by quantizing PM2.5 values and representing them as discrete symbols. A small LLM (e.g., Phi-3-mini or Qwen2.5 1.5B) will be fine-tuned in an autoregressive next-token prediction setup, using sliding time windows to capture temporal dependencies. The model will generate future tokens iteratively, which are then converted back into continuous forecasts. Both approaches will be trained and tested on identical rolling temporal splits for fair comparison.

4. Evaluation Plan

We adopt modern forecasting metrics to comprehensively evaluate accuracy and stability:

- Mean Absolute Error (MAE)
- Symmetric Mean Absolute Percentage Error (SMAPE)
- Mean Absolute Scaled Error (MASE)

We will compare one-step and multi-step forecasting and analyze interpretability (HMM latent state transitions vs. LLM token attention patterns).

5. Task Allocation

- **Zeyuan Ding:** Data preprocessing, HMM implementation, metric computation.
- **Keyu Hong:** LLM tokenization pipeline, fine-tuning, and sequence generation.

Both members will jointly analyze experimental results and co-author the final report.

6. Expected Outcome

We expect HMMs to perform competitively in stable short-term predictions, while 2025-era small LLMs may excel in capturing nonlinear contextual dynamics across longer temporal horizons. The study aims to highlight how recent lightweight LLMs can generalize linguistic reasoning capabilities to structured numerical forecasting.

[1] https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data