Performance of Hypothesis Classes on Time Series Data

Sasmit Datta Debangshu Chowdhary Netra Poonia

Jishnu Shenoi

1 Problem Statement

This report aims to evaluate the predictive capabilities of three distinct models: a linear model, a decision tree, and a neural network, to determine which of them provides the most accurate and reliable forecasts for a given time series dataset. Each model comes with its own strengths and weaknesses concerning time series data - linear models are simple and interpretable but may fail to capture complex patterns; decision trees offer a more nuanced approach to non-linearity but may overfit; and neural networks offers a parametrized way to fit non-linear data but can suffer from high computational demands.

1.1 Formulation

For given time series data $\mathbf{x}_{T-k}, \mathbf{x}_{T-k+1}, \ldots, \mathbf{x}_T$ where $\mathbf{x}_t \in \mathbb{R}^n$ we fit a model

$$f(\mathbf{x}_{T-k}, \mathbf{x}_{T-k+1}, ..., \mathbf{x}_{T-1}) = \mathbf{x}_T^{(n)} \tag{1}$$

where $\mathbf{x}_t^{(i)} \in \mathbb{R}$ is the i-th value of the vector $\mathbf{x}_t$. In essence what we are trying to do is given the multivariate time-series data points of the k previous time steps, we try to predict the scaler value of the n-th index of data point of the next time-step.

2 Methodology

2.1 Dataset

We focused on the Air Quality Index (AQI) of Delhi, utilizing a dataset spanning from 2010 to 2023 sourced from Kaggle. The initial steps involved data preprocessing which included feature reduction to streamline the dataset, removing unnecessary columns such as "agency" and "station-location", and ensuring that only relevant data for Delhi was consolidated.

Further preprocessing included handling missing values by merging similar columns, dropping rows and columns with all missing values, and renaming columns to more intuitive names. To address the missing data within the retained features, interpolation techniques were employed to maintain the continuity of the time series data.

Once the data was clean and structured, we calculated the AQI based on PM2.5 levels using a formula that maps PM2.5 concentrations to AQI values.

The features retained and refined for analysis after preprocessing included key pollutants such as PM2.5, PM10, nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3). These pollutants are standard measures for assessing air quality and have significant health and environmental implications. Alongside these, meteorological factors like temperature (Temp), relative humidity (RH), wind speed (WS), barometric pressure (BP), and wind direction (WD) were also considered because of their impact on pollutant dispersion and concentration levels.

We predict the AQI of the next time-step using all the features (including AQI) of $k = 10$ previous time-steps.

2.2 Linear Model

For a the linear model we used a simple linear regressor with following formulation

$$f(\mathbf{x}_{T-k}, \mathbf{x}_{T-k+1}, ..., \mathbf{x}_{T-1}) = \mathbf{w}_0 \cdot \mathbf{x}_{T-k} + \mathbf{w}_1 \cdot \mathbf{x}_{T-k+1} + ... + \mathbf{w}_{k-1} \cdot \mathbf{x}_{T-1} + b \tag{2}$$

where $\mathbf{w}_i \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are the tunable parameters of our model.

2.3 Neural Network

We use a simple multi-layer perceptron with two hidden layers of 256 dimensions each. So, we first take the k previous data points and concatenate them:

$$\mathbf{x} = \text{concat}(\mathbf{x}_{T-k}, \mathbf{x}_{T-k+1}, ..., \mathbf{x}_{T-1}) \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^{kn}$. We pass this $\mathbf{x}$ through our MLP, $f_\theta(\mathbf{z})$ to get $f(\mathbf{x}_{T-k}, \mathbf{x}_{T-k+1}, ..., \mathbf{x}_{T-1})$.

2.4 Regression Trees

Similar to before, we take k time series vectors and concatenate them like eq (**??**) and formulate our regression tree

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m \cdot I(\mathbf{x} \in R_m) \tag{4}$$

where c_m is the predicted value for the region R_m, and I is the indicator function, which equals 1 if $\mathbf{x}$ falls into the region R_m and 0 otherwise.

3 Experimental Results

We separated our data into a 90:5:5 train:validation:test split. For the linear model and neural network, we normalised all the inflowing data with the mean and standard deviation of our train dataset. MSE Loss and coefficient of determination (R^2) were used as metrics. We trained each model, performing hyper-parameter optimization to maximise performance on the validation dataset.

Regression Tree Model: The Regression Tree model demonstrates a near-perfect fit with an R^2 score of 0.999 and an MSE of 3.417 on the test data. This means that it captures the inherent pattern and variance of the underlying data quite well.

Model	R^2	MSE
Regression Tree	0.999	3.417
Neural Network	0.815	708.812
Linear Model	0.793	790.919

Table 1: Model Performance Metrics

Neural Network Model: With an R^2 score of 0.815 and an MSE of 708.812, the Neural Network model indicates a good, but not exceptional, level of accuracy. The neural network lacks am inductive bias towards sequence since the vectors are concatenated and aren't fed with any sort of sequential information.

Linear Model: The Linear Model lags slightly behind with an R^2 score of 0.793 and an MSE of 790.919. The linear model has been able to capture the variance in the data as well as the neural network but in terms of MSE, it is quite behind. Although, a higher MSE like this is natural due to the almost negligible parameter count as compared to the neural network (121 vs 97,024). This also alludes to the fact that the neural network may be over-parametrised for the task at hand and might be overfitting to the data.

4 Conclusion and Future Work

In conclusion, the models evaluated in this study reveal distinct predictive capabilities for the Delhi Air Quality Index (AQI) time series data. The Regression Tree model exhibited remarkable accuracy on the test set, as indicated by an R^2 score of 0.999 and an MSE of 3.417. The Neural Network and Linear Model, while demonstrating good performance, did not reach the exceptional levels of the Regression Tree, which could indicate that the additional complexity of the Neural Network may not be necessary for this dataset.

Looking ahead, future work should incorporate other statistical methods for data normalization to potentially enhance model accuracy. Considering the time series nature of the data, methods such as Z-score normalization, Min-Max scaling, or Box-Cox transformation could stabilize the data variance and mean. Introducing sequential models like RNNs, LSTMs, and Transformers might also prove beneficial. These models are inherently capable of capturing temporal dependencies and could provide a more sophisticated understanding of the temporal dynamics present in the data. Further exploration into advanced techniques like attention mechanisms, which have been highly successful in domains like natural language processing, may also offer new insights for improving time series forecasting in the context of air quality.