Comprehensive Guide to Statistical and Machine Learning Models

Sia Gao

Contents

1 Introduction

Understanding statistical and machine learning models is essential for analyzing and forecasting data in various domains, from finance and economics to marketing and engineering. This document provides a comprehensive guide to key concepts, mathematical formulations, validation techniques, and suitable scenarios for a wide range of models.

By presenting detailed descriptions and comparisons, this guide aims to demystify complex ideas and bridge the gap between theoretical knowledge and practical application. Whether you are preparing for exams, conducting research, or solving real-world problems, this resource equips you with the foundational tools to select and validate the right models for your needs.

2 Detailed Model Descriptions

2.1 ARIMA (Autoregressive Integrated Moving Average)

Mathematical Model:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Where:

- ϕ are autoregressive coefficients,
- θ are moving average coefficients,
- ϵ_t is the error term.

Parameters:

- p: Order of autoregression.
- d: Number of differencing steps (to make the series stationary).
- q: Order of moving average.

Suitable Scenarios:

- Time series data, especially stationary data or data made stationary through differencing.
- Short-term forecasting, e.g., stock prices or sales volumes.

Features:

- **AR (Autoregressive)**: Depends on past values.
- **I (Integrated)**: Removes non-stationarity.
- **MA (Moving Average)**: Captures dependencies in error terms.

Validation Methods:

- **ACF/PACF Plots:**
  - **ACF (Autocorrelation Function)**: Checks correlations and determines the need for differencing.
  - **PACF (Partial Autocorrelation Function)**: Helps identify the autoregressive order (p).
- **Residual Analysis:**
  - Ljung-Box Test: Verifies if residuals are white noise.
  - Residual ACF/PACF: Confirms the model captures the data's correlation.
- **Information Criteria:**
  - AIC/BIC: Compares models to select the optimal one.

Related Concepts:

- Differencing: Used to make non-stationary series stationary.

$$\Delta Y_t = Y_t - Y_{t-1}$$

- White Noise: A series with a mean of zero, constant variance, and no autocorrelation.

2.2 ETS (Error, Trend, Seasonality)

Mathematical Model:

1. Additive Model:
$$Y_t = (T_t + S_t) + E_t$$
2. Multiplicative Model:
$$Y_t = (T_t \times S_t) \times E_t$$

Where:

- T_t: Trend component,
- S_t: Seasonality component,
- E_t: Error term.

Suitable Scenarios:

- Time series data with trends and seasonality.
- Short-term and medium-term forecasting, such as consumer product sales.

Features:

- Decomposes time series into error, trend, and seasonality components.
- Highly flexible and adaptable to various types of time series.

Validation Methods:

- **Residual Analysis:**
  - Verify residuals are white noise.
  - Compute ACF for residuals.
- **Information Criteria:**
  - AIC/BIC for model selection.
- **Prediction Error:**
  - Evaluate using RMSE, MAE.

Related Concepts:

- Exponential Smoothing: Assigns exponentially decreasing weights to past data.
- **Smoothing Parameters:**
  - α: Level smoothing coefficient.
  - β: Trend smoothing coefficient.
  - γ: Seasonality smoothing coefficient.

2.3 Prophet

Mathematical Model:

$$Y_t = g(t) + s(t) + h(t) + \epsilon_t$$

Where:

- $g(t)$: Trend component,
- $s(t)$: Seasonal component,
- $h(t)$: Holiday effect,
- ϵ_t: Error term.

Suitable Scenarios:

- Data with nonlinear trends and multiple seasonal cycles.
- Data influenced by holidays or special events.
- Business data forecasting, such as website traffic and sales.

Features:

- Developed by Facebook, emphasizing usability and interpretability.
- Allows customization for holidays and special events.
- Handles missing and outlier values effectively.

Validation Methods:

- **Cross-Validation:**
  - Split data into training and validation sets to assess model performance.
- **Residual Analysis:**
  - Check residuals for white noise.
  - Analyze residual distribution.
- **Prediction Error:**
  - Evaluate using RMSE, MAE, MAPE.

Related Concepts:

- **Changepoints:** Automatically detects turning points in the trend.
- **Seasonal Components:** Accounts for yearly, quarterly, monthly, and weekly seasonality.
- **Holiday Effects:** Allows user-defined holiday effects.

2.4 Neural Network Autoregression (NNAR)

Mathematical Model:

$$\hat{Y}_t = f(W \cdot X_t + b)$$

Where:

- X_t: Input vector,
- W: Weight matrix,
- b: Bias vector,
- f: Activation function.

Suitable Scenarios:

- Time series with nonlinear and complex patterns.
- Data with both long-term and short-term dependencies.

Features:

- Uses neural networks to capture nonlinear relationships.
- Autoregressive inputs, leveraging past values to predict the future.

Validation Methods:

- Compare training and test set errors:
  - Verify whether the model overfits.
- Residual Analysis:
  - Check autocorrelation of residuals.
- Prediction Error:
  - Evaluate using RMSE, MAE, MAPE.

Related Concepts:

- **Activation Functions:**
  - Common options include Sigmoid and ReLU for enhancing nonlinearity.
- **Network Structure:**
  - Includes input, hidden, and output layers.
- **Overfitting and Regularization:**
  - Methods to prevent overfitting, such as Dropout and regularization techniques.

2.5 Holt-Winters

Additive Model:

$$\begin{aligned} l_t &= \alpha(Y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}), \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, \\ s_t &= \gamma(Y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}, \\ \hat{Y}_{t+h} &= l_t + hb_t + s_{t-m+h}. \end{aligned}$$

Multiplicative Model:

$$\begin{aligned} l_t &= \alpha\left(\frac{Y_t}{s_{t-m}}\right) + (1-\alpha)(l_{t-1} + b_{t-1}), \\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, \\ s_t &= \gamma\left(\frac{Y_t}{l_{t-1} + b_{t-1}}\right) + (1-\gamma)s_{t-m}, \\ \hat{Y}_{t+h} &= (l_t + hb_t) \times s_{t-m+h}. \end{aligned}$$

Where:

- l_t: Level,
- b_t: Trend,
- s_t: Seasonality,
- α, β, γ: Smoothing parameters,
- m: Seasonal length.

Suitable Scenarios:

- Time series with trends and seasonality.
- Periodic data, such as sales and tourism numbers.

Features:

- Models both trends and seasonality simultaneously.
- Provides additive and multiplicative forms for varying seasonality amplitudes.

Validation Methods:

- Residual Analysis:
  - Examine ACF and PACF of residuals.
- Prediction Error:
  - Evaluate using RMSE, MAE, MAPE.
- Information Criteria:
  - Use AIC/BIC for model selection.

Related Concepts:

- **Smoothing Parameter Adjustment:**
  - Optimize α, β, γ for better performance.
- **Seasonal Length (m):**
  - Determined by the periodic characteristics of the data.

2.6 GARCH (Generalized Autoregressive Conditional Heteroskedasticity)

Mathematical Model:

$$
\begin{aligned}
Y_t &= \mu + \epsilon_t, \\
\epsilon_t &= \sigma_t Z_t, \\
\sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,
\end{aligned}
$$

Where:

- σ_t^2: Conditional variance,
- Z_t: Random variable from a standard normal distribution.

Suitable Scenarios:

- Financial time series, such as stock returns and exchange rates.
- Data with volatility clustering.

Features:

- Models conditional heteroskedasticity to capture changing volatility.
- Explains autocorrelation in volatility.

Validation Methods:

- ACF/PACF of squared residuals:
  - Check for ARCH effects.
- Residual Normality Test:
  - Use the Jarque-Bera test to verify residual normality.
- Information Criteria:
  - AIC/BIC to determine the orders p and q.

Related Concepts:

- **ARCH Effect:**
  - Detects whether conditional variance changes over time.

- **Volatility Clustering:**
  - High-volatility periods tend to cluster.
- **Extended Models:**
  - EGARCH, TGARCH for capturing asymmetries in volatility.

2.7 VAR (Vector Autoregression)

Mathematical Model:

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \epsilon_t,$$

Where:

- Y_t: Vector of multiple time series,
- A_i: Coefficient matrices,
- ϵ_t: Error vector.

Suitable Scenarios:

- Multivariate time series analysis.
- Examining interdependencies among macroeconomic variables (e.g., GDP, inflation rate, interest rates).

Features:

- Captures dynamic relationships between variables.
- Does not require distinguishing endogenous and exogenous variables.

Validation Methods:

- Model Stability:
  - Check eigenvalues to ensure they lie within the unit circle.
- Residual Analysis:
  - Confirm residuals are white noise.
- Information Criteria:
  - Use AIC/BIC to select lag order p.

Related Concepts:

- **Granger Causality Test:**
  - Determines whether one variable helps predict another.
- **Impulse Response Function (IRF):**
  - Analyzes dynamic responses to shocks.
- **Variance Decomposition:**
  - Explains contributions of variables to forecast error.

2.8 Cointegration

Mathematical Model:

$$Y_t - \beta X_t = \epsilon_t,$$

Where:

- Y_t and X_t: Non-stationary variables,
- ϵ_t: Stationary linear combination of Y_t and X_t.

Suitable Scenarios:

- Analyzing long-term equilibrium relationships between non-stationary time series.
- Applications in finance and economics, such as stock prices, exchange rates, and interest rates.

Features:

- Captures long-term stable relationships between variables.
- Avoids spurious regression issues.

Validation Methods:

- **Engle-Granger Two-Step Method:**
  - Step 1: Regress variables to obtain residuals.
  - Step 2: Test residuals for stationarity.
- **Johansen Test:**
  - Determines the number of cointegrating relationships.

Related Concepts:

- **Error Correction Model (ECM):**

$$\Delta Y_t = \alpha(Y_{t-1} - \beta X_{t-1}) + \gamma \Delta X_t + \epsilon_t,$$

  Combines short-term dynamics with long-term equilibrium.

- **Unit Root Tests:**
  - ADF and PP tests to check for stationarity.

2.9 THIEF (Temporal Hierarchical Forecasting)

Model Concept:

- Models and forecasts time series at different temporal frequencies, then combines the results.

Suitable Scenarios:

- Scenarios requiring both long-term and short-term forecasts.
- Data with features across multiple time scales.

Features:

- Leverages hierarchical structure in time series to improve forecasting performance.
- Combines information from both high-frequency and low-frequency components.

Validation Methods:

- **Prediction Error:**
  - Evaluate using RMSE, MAE, MAPE.
- **Residual Analysis:**
  - Check if combined forecast residuals are white noise.

Related Concepts:

- **Temporal Decomposition:**
  - Decomposes time series into components at different frequencies.
- **Forecast Combination:**
  - Effectively integrates forecasts from different time scales.

2.10 MAPA (Multiple Aggregation Prediction Algorithm)

Model Concept:

- Aggregates the time series at multiple time scales (different time windows), forecasts at each scale, and combines the results.

Suitable Scenarios:

- Long-term forecasting.
- Data with multi-scale characteristics, where single-scale forecasting performs poorly.

Features:

- Reduces noise through aggregation, enhancing forecast stability.
- Combines information across different time scales.

Validation Methods:

- **Prediction Error:**
  - Evaluate using RMSE, MAE, MAPE.
- **Residual Analysis:**
  - Examine residuals from combined forecasts.

Related Concepts:

- **Time Aggregation:**
  - Studies the effect of different time windows on forecasting.
- **Forecast Combination:**
  - Effectively integrates forecasts across different aggregation levels.

2.11 Dynamic Harmonic Regression

Mathematical Model:

$$Y_t = \beta_0 + \sum_{k=1}^{K} \left[\alpha_k \cos\left(\frac{2\pi k t}{s}\right) + \beta_k \sin\left(\frac{2\pi k t}{s}\right) \right] + \epsilon_t,$$

Where:

- s: Seasonal cycle,
- K: Number of harmonics.

Suitable Scenarios:

- Complex seasonal patterns, such as energy consumption data with multiple daily peaks.
- Non-integer periodic data.

Features:

- Uses Fourier terms to capture complex seasonality.
- Handles non-standard seasonal patterns.

Validation Methods:

- **Spectral Analysis:**
  - Examines periodic components in the data.
- **Residual Analysis:**
  - Verifies residuals are white noise.
- **Information Criteria:**
  - Use AIC/BIC to determine the number of harmonics (K).

Related Concepts:

- **Fourier Transform:**
  - Converts time-domain signals into frequency domain.
- **Harmonics:**
  - Integer multiples of the fundamental frequency, capturing periodicity.

2.12 STL Decomposition (Seasonal-Trend Decomposition based on Loess)

Decomposition Model:

$$Y_t = T_t + S_t + R_t,$$

Where:

- T_t: Trend component,
- S_t: Seasonal component,
- R_t: Residual.

Suitable Scenarios:

- Data requiring separation of trend and seasonality.
- Time series with nonlinear trends and seasonal patterns.

Features:

- Nonparametric decomposition based on Loess (locally weighted regression).
- Robust to changes in seasonal components.

Validation Methods:

- **Decomposition Residual Analysis:**
  - Check if residuals are white noise.
- **Visualization:**
  - Analyze the trend, seasonality, and residual components graphically.

Related Concepts:

- **Loess Smoothing:**
  - Fits low-order polynomials locally.
- **Seasonal Adjustment:**
  - Removes seasonal effects from the data.

2.13 TBATS

Mathematical Model:

$$Y_t = \text{Box-Cox}(Y_t; \lambda) = l_t + \phi l_{t-1} + \sum_{i=1}^{k} s_t^{(i)} + d_t,$$

Where:

- l_t: Level,
- ϕ: Damped trend parameter,

- $s_t^{(i)}$: Seasonal components,
- d_t: ARMA error term.

Suitable Scenarios:

- Time series with complex multiple seasonalities (e.g., daily and weekly patterns).
- Long-period seasonality.

Features:

- Combines Box-Cox transformation, trend, seasonality, and ARMA error terms.
- Handles multiple and non-integer seasonal cycles.

Validation Methods:

- **Prediction Error:**
  - Evaluate using RMSE, MAE, MAPE.
- **Residual Analysis:**
  - Examine ACF and PACF of residuals.
- **Information Criteria:**
  - Use AIC/BIC for model selection.

Related Concepts:

- **Box-Cox Transformation:**
  - Stabilizes variance and normalizes data.
- **Damped Trend:**
  - Controls the growth rate of the trend.

2.14 Model Combinations

Mathematical Model:

$$\hat{Y}_t = \sum_{i=1}^{n} w_i \hat{Y}_{i,t},$$

Where:

- w_i: Weight of the i-th model,
- $\hat{Y}_{i,t}$: Prediction from the i-th model.

Suitable Scenarios:

- When improving prediction accuracy and stability is desired.
- Situations where different models have unique strengths.

Features:

- Combines predictions from multiple models to reduce bias and variance.
- Weights can be equal or adjusted based on historical performance.

Validation Methods:

- **Prediction Error:**
  - Evaluate using RMSE, MAE, MAPE.
- **Cross-Validation:**
  - Assess performance across different datasets.

Related Concepts:

- **Weighted Average:**
  - Adjusts weights based on model historical error.
- **Model Integration:**
  - Inspired by ensemble learning approaches such as Bagging and Boosting.

2.15 Monte Carlo Simulation

Basic Principle:

- Uses a large number of random samples to simulate a system or process, estimating expected values or probabilities.

Suitable Scenarios:

- Risk assessment in complex systems.
- Integration calculation and optimization problems.

Features:

- Does not require analytical solutions, making it suitable for high-dimensional problems.
- Accuracy improves with the number of simulations.

Validation Methods:

- **Convergence Check:**
  - Verify if results stabilize as the number of simulations increases.
- **Confidence Interval:**
  - Estimate the uncertainty range of results.

Related Concepts:

- **Random Number Generation:**
  - The quality of random numbers affects simulation results.
- **Variance Reduction Techniques:**
  - Methods like control variates and importance sampling to improve efficiency.

2.16 Markov Chains

Transition Probability Matrix:

$$P = [p_{ij}],$$

Where:

- $p_{ij} = P(X_{t+1} = j \mid X_t = i)$: Probability of transitioning from state i to state j.

Suitable Scenarios:

- Modeling system states, such as weather forecasting and customer behavior.
- Analyzing random processes.

Features:

- The next state depends only on the current state (Markov property).
- Transition probabilities describe movements between states.

Validation Methods:

- **Steady-State Distribution:**
  - Solve $\pi P = \pi$ to find steady-state probabilities.
- **Path Simulation:**
  - Simulate state transitions to validate the model.

Related Concepts:

- **Absorbing States:**
  - States that, once entered, cannot be left.
- **Periodicity:**
  - Whether the system has cyclic behavior.

2.17 MCMC (Markov Chain Monte Carlo)

Basic Principle:

- Constructs a Markov chain with a target distribution through random sampling.

Suitable Scenarios:

- Sampling from complex probability distributions.
- Computing posterior distributions in Bayesian statistics.

Features:

- Handles high-dimensional and complex distributions.
- Common algorithms include Metropolis-Hastings and Gibbs Sampling.

Validation Methods:

- **Convergence Diagnostics:**
  - Use Gelman-Rubin statistics to check chain convergence.
- **Autocorrelation Analysis:**
  - Examine independence of samples.

Related Concepts:

- **Burn-in Period:**
  - Discard initial samples to reduce the effect of starting values.
- **Autocorrelation Time:**
  - Estimates correlation between samples, adjusting sampling intervals.

2.18 State Space Models (Kalman Filtering)

Mathematical Model:

1. State Equation:
$$X_t = F_t X_{t-1} + G_t \epsilon_t$$
2. Observation Equation:
$$Y_t = H_t X_t + v_t$$

Where:

- X_t: State variable,
- Y_t: Observation variable,
- ϵ_t, v_t: Error terms.

Suitable Scenarios:

- Real-time estimation of dynamic systems.
- Target tracking, navigation, and economic forecasting.

Features:

- Recursive estimation with efficient computation.
- Handles noisy observational data.

Validation Methods:

- **Prediction Error Analysis:**
  - Check the distribution of Kalman filter prediction errors.
- **Model Comparison:**

  - Use AIC/BIC for model selection.

Related Concepts:

- **Kalman Filter:**
  - Minimizes the variance of estimation errors.
- **Extended Kalman Filter:**
  - Handles nonlinear systems.

3 Model Comparison Table

Model	Validation Methods	Related Concepts	Best-Suited Scenarios
ARIMA	AIC/BIC, Ljung-Box test, Residual Analysis	ACF/PACF, Differencing, White Noise	Stationary or differenced stationary data
ETS	AIC/BIC, Residual Analysis, Prediction Error	Exponential Smoothing, Trend, Seasonality	Time series with trend and seasonality
Prophet	Cross-Validation, Residual Analysis, Prediction Error	Changepoints, Seasonality, Holiday Effects	Nonlinear trends with holiday influences
NNAR	Train/Test Errors, Residual Analysis, Prediction Error	Activation Functions, Network Structure	Nonlinear and complex patterns in time series
Holt-Winters	Residual Analysis, Prediction Error, AIC/BIC	Smoothing Parameters, Seasonal Length	Cyclical data with trends and seasonality
GARCH	AIC/BIC, Squared Residual ACF/PACF, Normality Test	ARCH Effects, Volatility Clustering	Modeling volatility in financial time series
VAR	AIC/BIC, Residual Analysis, Model Stability	Granger Causality, Impulse Response	Dynamic relationships in multivariate data
Cointegration	Engle-Granger, Johansen Test	Error Correction Model, Unit Root Tests	Long-term equilibrium relationships
THIEF	Prediction Error, Residual Analysis	Temporal Decomposition, Forecast Combination	Multi-scale long-term forecasting
MAPA	Prediction Error, Residual Analysis	Time Aggregation, Forecast Combination	Long-term forecasts with multi-scale features
Dynamic Harmonic Regression	AIC/BIC, Spectral Analysis, Residual Analysis	Fourier Transform, Harmonics	Complex seasonal patterns
STL Decomposition	Residual Analysis, Visualization	Loess Smoothing, Seasonal Adjustment	Nonlinear trends and seasonal patterns
TBATS	AIC/BIC, Residual Analysis, Prediction Error	Box-Cox Transformation, Damped Trend	Complex multiple seasonalities

Continued on next page...

Model	Validation Methods	Related Concepts	Best-Suited Scenarios
Model Combinations	Prediction Error, Cross-Validation	Weighted Average, Model Integration	Enhancing forecast accuracy
Monte Carlo Simulation	Convergence Check, Confidence Interval	Random Number Generation, Variance Reduction	Risk assessment and simulation
Markov Chains	Steady-State Distribution, Path Simulation	Absorbing States, Periodicity	System state modeling
MCMC	Convergence Diagnostics, Autocorrelation Analysis	Burn-in Period, Autocorrelation Time	Bayesian statistics and complex sampling
State Space Models	Prediction Error Analysis, Model Comparison	Kalman Filter, Extended Kalman Filter	Real-time estimation of dynamic systems

Table 1: Comparison of Statistical and Machine Learning Models

4 Conclusion

This document has provided an in-depth exploration of various statistical and machine learning models, including their mathematical formulations, suitable applications, and validation methods. By breaking down these concepts into understandable components and offering a detailed comparison table, it serves as a practical resource for both theoretical understanding and practical implementation.

Whether you are tackling academic research, business analytics, or engineering challenges, the knowledge and insights presented here empower you to make informed decisions in model selection and application. The ability to connect the right model to the right problem is key to achieving reliable and actionable results. As you explore and apply these techniques, this document will remain a valuable reference for mastering data analysis and forecasting.