Model Diagnostics, Selection, and Evaluation in Regression and Classification

A Comprehensive Approach for Robust Models

Dr. Martin Pius

January 2025

https://github.com/martinpius/PG_training

Boston Housing Dataset

- **Objective**: Predict median house prices in Boston suburbs.
- **Features** (13 predictors):
  - CRIM: Per capita crime rate.
  - ZN: Proportion of residential land zoned for large lots.
  - RM: Average number of rooms per dwelling.
  - LSTAT: Percentage of lower status population.
  - PTRATIO: Pupil-teacher ratio, etc.
- **Target Variable**:
  - MEDV: Median value of owner-occupied homes ($1000s).
- **Applications**:
  - Housing market analysis.
  - Urban planning and development.

Iris Dataset

- **Objective**: Classify iris flowers into species based on physical measurements.
- **Features** (4 predictors):
  - Sepal length and width.
  - Petal length and width.
- **Target Variable**:
  - Species: Setosa, Versicolor, Virginica.
- **Applications**:
  - Pattern recognition and classification.
  - Benchmarking classification algorithms.

Diagnostics for Regression Models

- **Multicollinearity**:
  - **Definition**: Occurs when predictors are highly correlated, making coefficient estimation unreliable.
  - **Test**: Variance Inflation Factor (VIF). A VIF greater than 5 indicates significant multicollinearity.
  - **Solution**: Remove correlated predictors or use regularization methods (Ridge/Lasso).
- **Residual Analysis**:
  - **Tests**:
    - Normality: Q-Q plots and histograms.
    - Heteroscedasticity: Scatterplot of residuals vs. predictions.
  - **Interpretation**:
    - Normal residuals ensure unbiased predictions.
    - Random variance indicates consistent estimates.
- **Outlier Detection**:
  - **Tests**: Cook's Distance and Leverage Statistics.
  - **Interpretation**: High values suggest data points influencing model parameters.

- **Metrics for Selection**:
  - **AIC (Akaike Information Criterion)**:
    - Penalizes model complexity.
    - Lower values indicate better models.
  - **BIC (Bayesian Information Criterion)**:
    - Stronger penalty for complexity than AIC.
    - Lower values preferred for simpler models.
- **R-squared and RMSE**:
  - R-squared: Measures variance explained by the model.
  - RMSE: Root mean square error to assess prediction error.

Example:

- Linear vs Polynomial Regression: Use AIC, BIC, and Cross-Validation RMSE to select the best fit.

Model Selection: Classification

- **Metrics for Evaluation**:
  - **Precision**: Fraction of correctly predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

  - **Recall**: Fraction of actual positives correctly identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  - **F1-Score**: Harmonic mean of precision and recall.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC Curve and AUC**:
  - ROC Curve visualizes the trade-off between sensitivity and specificity.
  - Higher AUC indicates better performance.

Cross-Validation

- **Definition**: Resampling technique for robust model evaluation.
- **Benefits**:
  - Reduces overfitting.
  - Provides reliable performance estimates.
- **Types**:
  - K-Folds Cross-Validation: Split data into K folds; train on $K-1$, test on the remaining fold.
  - LOOCV (Leave-One-Out Cross-Validation): Use one data point as a test set in each iteration.
  - Split data into K folds.
  - Train on $K-1$, test on the remaining fold.
  - Average performance across folds.

Summary

- Diagnostics ensure model reliability:
  - Multicollinearity: Address with VIF or dimensionality reduction.
  - Residual Analysis: Check for normality and constant variance.
  - Outliers: Investigate influential data points.
- Model Selection:
  - Regression: Use AIC/BIC for simplicity and accuracy.
  - Classification: Use Precision, Recall, F1-Score, and AUC.
- Cross-Validation improves model robustness and generalization.

Thank You for Listening

Questions?

https://github.com/martinpius/PG_training