MATLAB을 이용한 고급 데이터 fitting 기법

Application Engineer
엄 준 상 대리
Agenda

- Regression and Curve Fitting in R2012a
- Don’t know which type of model to use
- Need to identify which parameters are most important
What is regression?

- Type of predictive modeling
- Specify a model that describes Y as a function of X
- Estimate a set of coefficients that minimize the difference between predicted and actual
- Typically, minimize the sum of the squared errors (the sum of the squared residuals)

\[ y = mx + b \]
Types of Regression: *Linear Regression*

- Implies that $Y$ is a *linear* function of the regression coefficients

- Common examples:

  - **Straight line**
    \[ Y = B_0 + B_1X_1 \]

  - **Plane**
    \[ Y = B_0 + B_1X_1 + B_2X_2 \]

  - **Polynomial**
    \[ Y = B_0 + B_1X_1^3 + B_2X_1^2 + B_3X_1 \]

  - **Polynomial with cross terms**
    \[ Y = B_0 + B_1X_1^2 + B_2(X_1 \times X_2) + B_3X_2^2 \]
Types of Regression: *Linear Regression*

- Implies that $Y$ is a *linear* function of the regression coefficients

- Syntax for formulas:
  - **Straight line**
    - default
  - **Plane**
    - default
  - **Polynomial**
    - `poly3` or $y \sim x1^3$
  - **Polynomial with cross terms**
    - $y \sim x1:x1 + x1:x2 + x2:x2$
## Regression Analysis in Statistics Toolbox

<table>
<thead>
<tr>
<th>Model</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before R2012a</td>
</tr>
<tr>
<td><strong>Linear Regression</strong></td>
<td>polyfit</td>
</tr>
<tr>
<td></td>
<td>regress</td>
</tr>
<tr>
<td></td>
<td>regstats</td>
</tr>
<tr>
<td><strong>Nonlinear Regression</strong></td>
<td>nlinfit</td>
</tr>
<tr>
<td><strong>Generalized Linear Model</strong></td>
<td>glmfit</td>
</tr>
</tbody>
</table>

- **Linear Regression**
  - *Before R2012a*: polyfit, regress, regstats
  - *New in R2012a*: LinearModel

- **Nonlinear Regression**
  - *Before R2012a*: nlinfit
  - *New in R2012a*: NonLinearModel

- **Generalized Linear Model**
  - *Before R2012a*: glmfit
  - *New in R2012a*: GeneralizedLinearModel
Demo: Modeling House Prices in Boston

- **Dependent variable**
  - Housing price

- **Independent variables**
  - Tax rate
  - Age of house
  - Student / teacher ratio
  - Geographic location
  - Air quality
  - Etc.

- **Goal**
  - Identify most significant variables
  - Build an accurate predictive model

*Created with MATLAB and Mapping Toolbox*
Agenda

- Regression and Curve Fitting in R2012a
- Don’t know which type of model to use
- Need to identify which parameters are most important
Regression Techniques

- Regression techniques require that the user specify a model.
- Model specification describes the dynamics of the system.

Example - population models

Logistic Growth

\[ N_t = \frac{N_0 \times K}{N_0 + (K - N_0) \times \exp(-r_0 \times t)} \]

Exponential Growth

\[ N_t = N_0 \times e^{(r \times t)} \]
Don’t know what type of model to use?

Line

Quadratic

Rational
“Black-Box” Models

Low Dimensional Data

\[ X_1 \rightarrow \cdots \rightarrow X \rightarrow X_2 \rightarrow \cdots \rightarrow X_n \rightarrow \text{?} \rightarrow Y \]

High Dimensional Data
Low Dimensional Data
Localized Scatter Plot Smoothing (LOWESS)

- Subsets of data are fit using weighted linear least squares.
- Localized subsets are defined by span value.
- Optimal span value depends on data and requires experimentation to find.
Demo: Fit-It

- **Challenge:**
  - Data set consisting of X and Y
  - No first principles information
  - Need curve fit that best describes the relationship between X and Y

- **Technique:**
  - *Localized regression* to fit the data
  - *Cross-validation* to find optimal span
  - *Bootstrapping* to calculate confidence intervals
Cross-Validation
To Determine Optimal Span

- Technique to determine whether a model can be generalized to other similar data sets

- Approach
  - Divide data set into training and test data sets
  - Fit model to training set
  - Use test set to evaluate goodness of fit

- Specifics for our example
  - Use cross validation to evaluate many different spans
  - Choose the span that produces the most accurate result
Bootstrap
To Derive Confidence Intervals

- Technique to create a new data set that has similar statistical properties to our original data set

- Approach
  - Sample with replacement from original data set
  - Number of samples is identical to length of data set

- Specifics for our example
  - Generate a new data set
  - Fit a curve to the new data set
  - Repeat multiple times
  - Derive confidence intervals from these fits
High-Dimensional Data

- Localized regression inefficient with high-dimensional data

- Other algorithms available to solve this problem
  - Decision trees
  - Neural networks
  - Support vector machines

- Decision trees
  - Set of if-then statements that lead to a prediction

- Bagged decision tree
  - “Ensemble” of trees
  - Uses boot-strapping to create multiple trees
Demo: Short-Term Energy Load Forecaster

- **Challenge:**
  - Implement a “black-box” model for energy load forecasting

```
Weather
  - Dry Bulb
  - Dew Point

Seasonality
  - Hour, Weekday
  - Holidays

Historical Load
  - Previous Day
  - Previous Week
```

![Diagram](image_url)
Demo: Short-Term Energy Load Forecaster

- **Challenge:**
  - Implement a “black-box” model for energy load forecasting

- **Techniques:**
  - *Decision trees* for non-parametric modeling
  - *Bagging* to improve the model
Demo: Texture Classification

- Identify features appropriate for classification
- Extract features for training and test data
- Train classifier with features
- Test classifier and analyze results

- Using KTH-TIPS database

“On the significance of real-world conditions for material classification,”

“Classifying materials in the real world,” B. Caputo, E. Hayman, M. J.
Typical Classification Workflow

1. Select a classification method
2. Train your classifier
3. Measure classifier accuracy
4. Work with your model
5. Simplify your model
Available Classification Algorithms

- **Statistics Toolbox**
  - Discriminant Analysis
  - Logistic Regression
  - Classification Trees
  - Naïve Bayes Classifier
  - Bagged Decision Trees

- **BioInformatics Toolbox**
  - Support Vector Machines

- **Neural Network Toolbox**
  - Neural Networks
Agenda

- Regression and Curve Fitting in R2012a
- Don’t know which type of model to use
- Need to identify which parameters are most important
Why do we care?

- Predictive accuracy
  - Describe the signal, not the noise
  - Bias/variance tradeoff

- Interpretability
  - Simple models are easier to interpret
  - Occam’s razor

- Design considerations
  - Reduce memory
  - Improve speed
How do we do this?

- **Feature selection (relatively low # of predictors)**
  - Test all possible subset (brute force)
  - Sequential feature selection
  - ReliefF

- **Shrinkage/regularization (relatively large # of predictors)**
  - Ridge regression
  - Lasso
  - Elastic net

- **Feature transformation (relatively large # of predictors)**
  - Principal component regression
  - Partial least-squares regression
Sequential Feature Selection

*Searches for an Optimal Subset of Variables*

Start → Add the best predictor to the model → Add the predictor that improves the accuracy the most → Statistically significant?

Yes →

No → Remove the insignificant predictor → Stop
Demo: Predicting Wine Quality

- **Challenge:**
  - Determine which features (e.g., alcohol, sugar, sulfur content) are important for predicting the wine’s quality level

- **Techniques:**
  - *Naive Bayes classifier* (to predict wine quality)
  - *Sequential feature selection* (to determine optimal subset of features necessary to predict quality)
Lasso

- Introduces an additional term to the minimization
  - Prevents overfitting
  - Enables feature selection

\[
\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|
\]

Model coefficients are driven toward zero based on a tuning parameter (\(\lambda\)).
Demo: Lasso

- **Challenge**: Generate an accurate predictive model with a “tricky” data set

- **Approach**: Create a data set (and use Lasso to accurately model it)
  - Large number of predictors (8) compared with number of observations (20)
  - Predictors (X) are correlated with one another
  - Generate observations (Y) from predictors (X) with a linear relationship
  - 5 of the 8 predictors have zero influence on the generated observations
    - $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$

- **Techniques**:
  - **Lasso** (Subset selection, coefficient shrinkage)
  - **Cross-validation** (to identify correct $\lambda$ value)

---

Where to Go For More Information

- Demo code is available on MATLAB Central
Resource

- Code used in seminar
  - MATLAB File Exchange

- Recorded webinars
  - Data-Driven Fitting with MATLAB
  - Fitting with MATLAB: Statistics, Optimization, and Curve Fitting
  - Computational Statistics: Feature Selection, Regularization, and Shrinkage with MATLAB
  - New Capabilities for Regression and Curve Fitting