Machine Learning with MATLAB

Stefan Duprey, Application Engineer
stefan.duprey@mathworks.fr
What You Will Learn

- Overview of machine learning
- Algorithms available with MATLAB
- MATLAB as an interactive environment for evaluating and choosing the best algorithm
Machine Learning
Basic Concepts

- Start with an initial set of data

- “Learn” from this data
  - “Train” your algorithm with this data

- Use the resulting **model** to **predict** outcomes for **new** data sets
Machine Learning
Characteristics and Examples

- **Characteristics**
  - Lots of data (many variables)
  - System too complex to know the governing equation (e.g., black-box modeling)

- **Examples**
  - Pattern recognition *(speech, images)*
  - Financial algorithms *(credit scoring, algo trading)*
  - Energy forecasting *(load, price)*
  - Biology *(tumor detection, drug discovery)*
Model Development Process

- Exploration
- Modeling
- Evaluation
- Deployment
Exploratory Data Analysis

- Gain insight from visual examination
  - Identify trends and interactions
  - Detect patterns
  - Remove outliers
  - Shrink data
  - Select and pare predictors
  - Feature transformation

<table>
<thead>
<tr>
<th>MPG</th>
<th>Acceleration</th>
<th>Displacement</th>
<th>Weight</th>
<th>Horsepower</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>2000</td>
</tr>
<tr>
<td>4000</td>
<td>200</td>
<td>400</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>2000</td>
</tr>
<tr>
<td>4000</td>
<td>200</td>
<td>400</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>
Data Exploration
Interactions Between Variables

Plot Matrix by Group
Parallel Coordinates Plot
Andrews’ Plot

Glyph Plot
Chernoff Faces
Machine Learning Overview

Types of Learning, Categories of Algorithms

- **Unsupervised Learning**
  - Group and interpret data based only on input data

- **Supervised Learning**
  - Develop predictive model based on both input and output data

- **Categories of Algorithms**
  - Clustering
  - Classification
  - Regression
Unsupervised Learning
Clustering

Machine Learning

Unsupervised Learning

Group and interpret data based only on input data

Supervised Learning

Clustering

Classification

Regression

K-means, Fuzzy K-means
Hierarchical
Neural Network
Gaussian Mixture
Clustering Overview

- What is clustering?
  - Segment data into groups, based on data similarity

- Why use clustering?
  - Identify outliers
  - Resulting groups may be the matter of interest

- How is clustering done?
  - Can be achieved by various algorithms
  - It is an iterative process (*involving trial and error*)
Dataset We’ll Be Using

- Cloud of randomly generated points
  - Each cluster center is randomly chosen inside specified bounds
  - Each cluster contains the specified number of points per cluster
  - Each cluster point is sampled from a Gaussian distribution
- Multi-dimensional dataset
Example Cluster Analysis

K-Means

- K-means is a partitioning method
- Partitions data into $k$ mutually exclusive clusters
- Each cluster has a centroid (or center) – Sum of distances from all objects to the center is minimized
Distance Metrics & Group Quality

- Distance measures choices
- Many built-in distance metrics, or define your own

`>> doc pdist
>> distances = pdist(data,metric); %pdist = pairwise distances
>> squareform(distances)
>> kmeans(data,k,'distance','cityblock') %not all metrics supported`

- Create silhouette plots

`>> silhouette(data,clusters)`
Clustering Neural Network

- Networks are comprised of one or more layers

- Outputs computed by applying a nonlinear transfer function with weighted sum of inputs

- Trained by letting the network continually adjust itself to new inputs (*determines weights*)
Clustering
Neural Network

- Neural Network Toolbox provides interactive apps for easily creating and training networks

- Multi-layered networks created by cascading (provide better accuracy)

- Example architectures for clustering:
  - Self-organizing maps
  - Competitive layers
Self Organising Map Neural Net

How it Works

- Started with a regular grid of ‘neurons’ laid over the dataset
- Size of the grid determined the number of clusters
- Neurons competed to recognize data points (by being close to them)
- Winning neurons were moved closer to the data points
- Repeated until convergence
Gaussian Mixture Models

- Good when clusters have different sizes and are correlated
- Assume that data is drawn from a fixed number $K$ of normal distributions
Cluster Analysis

Summary

- Segments data into groups, based on data similarity

- No method is perfect (depends on data)

- Process is iterative; explore different algorithms

- Beware of local minima (global optimization can help)
Model Development Process

- Exploration
- Modeling
- Evaluation
- Deployment
Supervised Learning
Classification for Predictive Modeling

Machine Learning

Unsupervised Learning

Supervised Learning

Classify based on both input and output data
Classification
Overview

- What is classification?
  - Predicting the best group for each point
  - “Learns” from labeled observations
  - Uses input features

- Why use classification?
  - Accurately group data never seen before

- How is classification done?
  - Can use several algorithms to build a predictive model
  - Good training data is critical
Example Classification

Decision Trees

- Builds a tree from training data
  - Model is a tree where each node is a logical test on a predictor

- Traverse tree by comparing features with threshold values

- The “leaf” of the tree specifies the group
Ensemble Learners
Overview

- Decision trees are “weak” learners
  - Good to classify data used to train
  - Often not very good with new data
  - Note rectangular groups

- What are ensemble learners?
  - Combine many decision trees to create a “strong” learner
  - Uses “bootstrapped aggregation”

- Why use ensemble methods?
  - Classifier has better predictive power
  - Note improvement in cluster shapes
Decision Trees
How do I build them with MATLAB?

- **Build tree model**
  
  ```matlab
  >> tree = classregtree(x,y);
  >> view(tree)
  ```

- **Evaluate the model on new data**
  
  ```matlab
  >> tree(x_new)
  ```
Enhancing the model: Ensemble Learning

- Combine weak learners into a stronger learner
  
  `>> ens = fitensemble(x,y,'AdaBoostM2',200,'Tree');`

- Bootstrapped aggregated trees forest
  
  `>> ens = fitensemble(x,y,'Bag',200,'Tree','type','classification');`
  `>> y_pred = predict(ens,x);`

- Visualise class boundaries
K-Nearest Neighbor Classification

- One of the simplest classifiers
- Takes the $K$ nearest points from the training set, and chooses the majority class of those $K$ points
- No training phase – all the work is done during the application of the model
MATLAB Helps to Manage Complexity

- A single calling syntax for all methods
- Documentation helps you choose an appropriate algorithm for your particular problem
Support Vector Machines
Overview

- Good for modeling with complex boundaries between groups
  - Can be very accurate
  - No restrictions on the predictors

- What does it do?
  - Uses non-linear “kernel” to calculate the boundaries
  - Can be computationally intensive

- Version in Statistics Toolbox only classifies into two groups
Classification

Summary

- No absolute best method
- Simple does not mean inefficient
- Watch for overfitting
  - Decision trees and neural networks may overfit the noise
  - Use ensemble learning and cross-validation
- Parallelize for speedup
Supervised Learning
Regression for Predictive Modeling

Machine Learning

Unsupervised Learning

Supervised Learning

Develop predictive model based on both input and output data

Regression

Linear
Non-linear
Non-parametric
Regression

- Why use regression?
  - Predict the continuous response for new observations

- Type of predictive modeling
  - Specify a model that describes Y as a function of X
  - Estimate coefficients that minimize the difference between predicted and actual

- You can apply techniques from earlier sections with regression as well *(e.g., Neural Network)*
Linear Regression

- Y is a *linear* function of the regression coefficients
- Common examples:

  **Straight line**
  \[ Y = B_0 + B_1 X_1 \]

  **Plane**
  \[ Y = B_0 + B_1 X_1 + B_2 X_2 \]

  **Polynomial**
  \[ Y = B_0 + B_1 X_1^3 + B_2 X_1^2 + B_3 X_1 \]

  **Polynomial with cross terms**
  \[ Y = B_0 + B_1 X_1^2 + B_2 (X_1 \ast X_2) + B_3 X_2^2 \]
Nonlinear Regression

- Y is a *nonlinear* function of the regression coefficients
- Syntax for formulas:

  **Fourier Series**
  \[ y \sim b_0 + b_1 \cos(b_3 x) + b_4 \sin(b_3 x) \]

  **Exponential Growth**
  \[ @(b, t) (b(1) \times \exp(b(2) \times t)) \]

  **Logistic Growth**
  \[ @(b, t) \left( \frac{1}{b(1) + \exp(-b(2) \times x)} \right) \]
Generalized Linear Models

- Extends the linear model
  - Define relationship between model and response variable
  - Model error distributions other than normal

- Logistic regression
  - Response variable is binary (true / false)
  - Results are typically expressed as an odd’s ratio

- Poisson regression
  - Model count data (non-negative integers)
  - Response variable comes from a Poisson distribution
Machine Learning with MATLAB

- Interactive environment
  - Visual tools for exploratory data analysis
  - Easy to evaluate and choose best algorithm
  - Apps available to help you get started
    (e.g., neural network tool, curve fitting tool)

- Multiple algorithms to choose from
  - Clustering
  - Classification
  - Regression
Learn More: Machine Learning with MATLAB

Data Driven Fitting with MATLAB

Multivariate Classification in the Life Sciences

Classification with MATLAB

Electricity Load and Price Forecasting

Regression with MATLAB

Credit Risk Modeling with MATLAB