MACHINE LEARNING: CLASSIFICATION

PREMANANDA INDIC, PH.D.

DEPARTMENT OF ELECTRICAL ENGINEERING
ANALYSIS PLATFORM

MATLAB Access for Everyone at
University of Texas at Tyler

ANALYSIS PLATFORM

University of Texas at Tyler

MATLAB Access for Everyone at
University of Texas at Tyler

OUTLINE

➢ INTRODUCTION

➢ DIFFERENT CLASSIFIERS

➢ EXAMPLES
OUTLINE

➢ INTRODUCTION

➢ DIFFERENT CLASSIFIERS

➢ EXAMPLES
What is Machine Learning?

- Machine Learning is a field of study that gives computers the ability to “learn” without being explicitly programmed
  - Prediction
  - Classification
INTRODUCTION

➢ What is Machine Learning?

- Machine Learning is a field of study that gives computers the ability to “learn” without being explicitly programmed
  - Prediction
  - Classification

Samuel AL, IBM J. Research & Development, 1959, vol. 3 (3), 210-229
OUTLINE

➢ INTRODUCTION

➢ DIFFERENT CLASSIFIERS

➢ EXAMPLES
APPROACHES

➢ SUPERVISED LEARNING

➢ UNSUPERVISED LEARNING
APPROACHES

➢ SUPERVISED LEARNING (Classification / Prediction)

Provide training set with features and solutions
APPROACHES

➢ STANDARD MACHINE LEARNING

➢ ADVANCED MACHINE LEARNING

  Based on Artificial Neural Networks (Deep Learning)
APPRAOCHES

CLASSIFICATION

- Logistic Regression
- Support Vector Machine
APPROACHES

CLASSIFICATION

• Logistic Regression

• Support Vector Machine
APPROACHES

➢ Linear Regression

\[ \hat{y}^i = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i + \ldots + \theta_n x_n^i \quad i = 1, 2, \ldots, m \]

\[ \hat{Y} = \Theta^T X \]

- Gradient Descent by Louis Augustin Cauchy in 1847

Cost Function to Minimize

\[ J = \left( (\hat{y}^i - y^i)^2 \right) = (\hat{Y} - Y)^T (\hat{Y} - Y) = \frac{1}{m} \sum_{i=1}^{m} (\theta^T X^i - y^i)^2 \]
APPROACHES

➢ Linear Regression

\[ \theta^{k+1} = \theta^k - \gamma \nabla_{\theta} J(\theta) \]

\[ \nabla_{\theta} J(\theta) = \frac{2}{m} X^T (X\theta - y) \]
APPROACHES

➢ Logistic Regression

Two class \( y = 1 \) or \( y = 0 \)

\[
\hat{p} = f(\Theta^T X) = \frac{1}{1 + e^{-\Theta^T X}}
\]

\( \hat{y} = 1 \) if \( \hat{p} < 0.5 \); \( \hat{y} = 0 \) if \( \hat{p} \geq 0.5 \)

\[
J = \frac{1}{m} \sum_{i=1}^{m} [y^i \log(\hat{p}^i) + (1 - y^i) \log(1 - \hat{p}^i)]
\]
APPROACHES

➢ Logistic Regression

\[ \theta^{k+1} = \theta^k - \gamma \nabla_{\theta} J(\theta) \]

\[ \frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (f(\theta^T X^i) - y^i) x_j^i \]
APPROACHES

➢ Support Vector Machine

\[
f(X) = w^T X - b
\]

\[
G(x_j, x_k) = \exp(-\|x_j - x_k\|^2)
\]

\[
G(x_j, x_k) = (1 + x_j'x_k)^q, \text{ where } q \text{ is in the set \{2,3,...\).}
\]

https://medium.com/@LSchultebraucks/introduction-to support-vector-machines-9f8161ae2fcb
APPROACHES

➢ SUPERVISED LEARNING (Classification / Prediction)

  • Support Vector Machine (SVM)

Used for regression as well as classification

https://www.mathworks.com/matlabcentral/fileexchange/62061-multi-class-svm
APPROACHES

➢ SUPERVISED LEARNING (Classification)
  • Logistic Regression
  • Support Vector Machines
  • k-Nearest Neighbors
  • Decision Trees and Random Forests
SECTION 1: Learner App

➢ Home Value Classification: 9 features to classify high vs low medianHouseValue

- longitude: A measure of how far west a house is; a higher value is farther west
- latitude: A measure of how far north a house is; a higher value is farther north
- housingMedianAge: Median age of a house within a block; a lower number is a newer building
- totalRooms: Total number of rooms within a block
- totalBedrooms: Total number of bedrooms within a block
- population: Total number of people residing within a block
- households: Total number of households, a group of people residing within a home unit, for a block
- medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
  
  **medianHouseValue**: Median house value for households within a block (measured in US Dollars)

- oceanProximity: Location of the house w.r.t ocean/sea

https://www.kaggle.com/camnugent/california-housing-prices

Demo with N=5000
70% Training Data
30% Test Data
Models Trained:
Logistic Regression
SVM
SECTION 1: Learner App

 Prediction of House Price Classification Problem

Confusion Matrix

<table>
<thead>
<tr>
<th>True Class</th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive</td>
</tr>
<tr>
<td>1</td>
<td>False Negative</td>
</tr>
<tr>
<td>0</td>
<td>False Positive</td>
</tr>
<tr>
<td>0</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

True Positive Rate = True Positive / Total Positive

True Negative Rate = True Negative / Total Negative = 1 – False Positive Rate
SECTION 1: Learner App

DATA IMPORT & CLASSIFICATION LEARNER INITIALIZATION

The Learner App is designed to facilitate data import and classification tasks. It provides a user-friendly interface for initializing learners. The app supports data set uploading, variable selection, and various validation techniques, including cross-validation, holdout validation, and resubstitution validation. The interface allows users to select predictors, specify data types, and set ranges for each variable. The app also highlights the importance of distinguishable labels for response variables.
Demo Learner App in MATLAB - logistic regression and linear SVM

classificationLearner(Ttrain,'hi_lo_label');
SECTION 2: Raw Data Analysis

Visualize the data, Summarize variables, data cleaning, pre-processing if needed

207 Missing values, replace with median values

ocean_proximity: 20636×1 categorical Values:
- <1H OCEAN 9135
- INLAND 6550
- ISLAND 5
- NEAR BAY 2289
- NEAR OCEAN 2657
SECTION 3: Correlation Analysis

FIND VARIABLE CORRELATIONS TO EACH OTHER AND THE MEDIAN HOUSE VALUE

\[ R_{pp} = \text{corr}(\text{table2array}(T1(:, \text{select\_vars}))) \]
SECTION 4: Logistic Regression

SPLIT INTO TRAINING AND TEST DATA AND FIT LOGISTIC REGRESSION MODEL

mdl = fitglm([Ttrain(:,1:9)
table(y)],'Distribution','binomial');

Estimated Coefficients:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Estimate</th>
<th>SE</th>
<th>tStat</th>
<th>pValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-154.19</td>
<td>14.421</td>
<td>-10.692</td>
<td>1.106e-26</td>
</tr>
<tr>
<td>longitude</td>
<td>-1.7683</td>
<td>0.17448</td>
<td>-10.356</td>
<td>3.875e-24</td>
</tr>
<tr>
<td>latitude</td>
<td>-1.8133</td>
<td>0.18905</td>
<td>-9.4714</td>
<td>7.854e-22</td>
</tr>
<tr>
<td>housing_median_age</td>
<td>0.044239</td>
<td>0.0051484</td>
<td>8.5928</td>
<td>8.490e-19</td>
</tr>
<tr>
<td>total_rooms</td>
<td>0.0003444</td>
<td>5.7387e-05</td>
<td>5.3364</td>
<td>0.00040561</td>
</tr>
<tr>
<td>total_bedrooms</td>
<td>0.00080298</td>
<td>0.00084259</td>
<td>0.95299</td>
<td>0.3406</td>
</tr>
<tr>
<td>population</td>
<td>-0.0023529</td>
<td>0.00020995</td>
<td>-11.207</td>
<td>3.774e-29</td>
</tr>
<tr>
<td>households</td>
<td>0.0039573</td>
<td>0.00094559</td>
<td>4.185</td>
<td>2.851e-05</td>
</tr>
<tr>
<td>median_income</td>
<td>1.0172</td>
<td>0.053904</td>
<td>18.87</td>
<td>2.010e-79</td>
</tr>
<tr>
<td>ocean_proximity_INLAND</td>
<td>-0.053285</td>
<td>0.24937</td>
<td>-0.21368</td>
<td>0.8308</td>
</tr>
<tr>
<td>ocean_proximity_ISLAND</td>
<td>0</td>
<td>0</td>
<td>NaN</td>
<td>NaN</td>
</tr>
<tr>
<td>ocean_proximity_NEAR_BAY</td>
<td>-0.10616</td>
<td>0.19861</td>
<td>-0.53449</td>
<td>0.593</td>
</tr>
<tr>
<td>ocean_proximity_NEAR_OCEAN</td>
<td>0.11076</td>
<td>0.15948</td>
<td>0.6945</td>
<td>0.40737</td>
</tr>
</tbody>
</table>

3500 observations, 3498 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 1.83e+03, p-value = 0

Remove Insignificant features
SECTION 5: Outliers

DIAGNOSTICS OF MODELS - IDENTIFY OUTLIERS

mdl1 = fitglm([Ttrain(:,[1:4 6:8])
    table(y,'variablenames',{'Hi_lo_label'})],'
    Distribution','binomial');

plotDiagnostics(mdl1,'leverage')

Demo with MATLAB
**SECTION 6: Classification (Clean Data)**

**TEST MODEL FOR TWO CLASS CLASSIFICATION** (Logistic Regression)

Test Data N = 1500
(30% of 5000)

Missing Values
Insignificant Features
Outliers

Diagram: 2 class Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>high</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>252</td>
<td>104</td>
</tr>
<tr>
<td>low</td>
<td>69</td>
<td>1075</td>
</tr>
</tbody>
</table>
SECTION 7: SVM Classification

REGULARIZATION OF VARIABLES DONE AUTOMATICALLY, NO NEED TO CHOOSE FEATURES SEPARATELY AS WAS DONE EARLIER FOR LOGISTIC REGRESSION

Test Data N = 1500
(30% of 5000)

Linear SVM

Demo Logistic Regression and SVM binary classification with cleaned up data - PYTHON

SVMModel = fitcsvm(Ttrain(:,1:9), y, 'standardize', true);
SECTION 8: SVM Classification

LINEAR vs RADIAL BASIS FUNCTION (RBF) KERNEL

fitcsvm([x1 x2],y1);
fitcsvm([x1 x2],y1,'KernelFunction','rbf');

x1: Age of House
X2: Median Income

Demo SVM decision boundaries with MATLAB
SECTION 9: Multiclassification (SVM)

ONE CLASS vs REST

Also perform one to one class

Mdl = fitcecoc(Ttrain(:,1:8),y,'Learners',t,'Coding',coding,'ResponseName',responseName,... 'PredictorNames',predictorNames,'ClassNames',classNames);
SECTION 10: Multiclassification (SVM)

LOW vs MOD vs HIGH CLASS

\[
\text{Mdlp} = \text{fitcecoc(Ttrain(:,1:8),y,'Learners',t,'FitPosterior',true,...}
\]

'ClassNames',{'low','mod','high'},...

'Verbose',2);

Demo SVM Multi-class classification with MATLAB
CONCLUSION

➢ Classification divides the data into different groups

➢ Look at the raw data and understand features in relation to class designation

➢ Several codes are available to perform classification
THANK YOU

SBIR: RAE (Realize, Analyze, Engage) - A digital biomarker based detection and intervention system for stress and cravings during recovery from substance abuse disorders.

*Pis: M. Reinhardt, S. Carreiro, P. Indic*

Department of Veterans Affairs

Design of a wearable sensor system and associated algorithm to track suicidal ideation from movement variability and develop a novel objective marker of suicidal ideation and behavior risk in veterans.

Clinical Science Research and Development Grant (approved for funding).

*P. Indic (site PI, UT-Tyler)*

*E.G. Smith (Project PI, VA)*

*P. Salvatore (Investigator, Harvard University)*

ORS Research Design & Data Analysis Lab

Office of Research and Scholarship

National Institute Of Health Grant

*P. Indic (Analytical Core PI, UT-Tyler)*

*N. Ambal (PI, Univ. of Alabama, Birmingham)*

Pre-Vent

National Science Foundation Smart & Connected Health Grant

*P. Indic (Lead PI, UT-Tyler)*

*D. Paydarfar (Co PI, UT-Austin)*

*H. Wang (Co PI, UMass Dartmouth)*

*Y. Kim (Co PI, UMass Dartmouth)*

SBIR: RAE (Realize, Analyze, Engage) - A digital biomarker based detection and intervention system for stress and cravings during recovery from substance abuse disorders.

*Pis: M. Reinhardt, S. Carreiro, P. Indic*

Department of Veterans Affairs

Design of a wearable sensor system and associated algorithm to track suicidal ideation from movement variability and develop a novel objective marker of suicidal ideation and behavior risk in veterans.

Clinical Science Research and Development Grant (approved for funding).

*P. Indic (site PI, UT-Tyler)*

*E.G. Smith (Project PI, VA)*

*P. Salvatore (Investigator, Harvard University)*

ORS Research Design & Data Analysis Lab

Office of Research and Scholarship

National Institute Of Health Grant

*P. Indic (Analytical Core PI, UT-Tyler)*

*N. Ambal (PI, Univ. of Alabama, Birmingham)*

Pre-Vent

National Science Foundation Smart & Connected Health Grant

*P. Indic (Lead PI, UT-Tyler)*

*D. Paydarfar (Co PI, UT-Austin)*

*H. Wang (Co PI, UMass Dartmouth)*

*Y. Kim (Co PI, UMass Dartmouth)*

SBIR: RAE (Realize, Analyze, Engage) - A digital biomarker based detection and intervention system for stress and cravings during recovery from substance abuse disorders.

*Pis: M. Reinhardt, S. Carreiro, P. Indic*

Department of Veterans Affairs

Design of a wearable sensor system and associated algorithm to track suicidal ideation from movement variability and develop a novel objective marker of suicidal ideation and behavior risk in veterans.

Clinical Science Research and Development Grant (approved for funding).

*P. Indic (site PI, UT-Tyler)*

*E.G. Smith (Project PI, VA)*

*P. Salvatore (Investigator, Harvard University)*

ORS Research Design & Data Analysis Lab

Office of Research and Scholarship

National Institute Of Health Grant

*P. Indic (Analytical Core PI, UT-Tyler)*

*N. Ambal (PI, Univ. of Alabama, Birmingham)*

Pre-Vent

National Science Foundation Smart & Connected Health Grant

*P. Indic (Lead PI, UT-Tyler)*

*D. Paydarfar (Co PI, UT-Austin)*

*H. Wang (Co PI, UMass Dartmouth)*

*Y. Kim (Co PI, UMass Dartmouth)*
QUESTIONS