Data Science: Logistic Regression

CS531 - Python Applications - SFBU

Ricardo Naoki Horiguchi

Table of Contents

Introduction

a. The Concept
   i. To show different ways of analysing data and displaying it using Data Science techniques.

b. Background
   i. The need to process, store, analyze, and comprehend large volumes of diverse data in meaningful ways has increased in the last years due to internet being available to many people, specially through mobile phone. The need to interpret and analyse those data is also urgent, and can offer much advantage and insights. Through data science it is possible to find patterns and relationship inside raw data. This is where data science techniques have proven to be extremely useful.

c. Area of Study
   i. Data Science is a compilation of techniques that can provide value from raw data. It is used inside many different fields of knowledge such as Machine Learning or Artificial Intelligence.
   ii. It uses many techniques such as decision trees, bayesian learning, KNN, visualization, statistics, data preparation and process mining.

Process: EDA

i. EDA - Exploratory Data Analysis
   1. It is the approach of analyzing data sets to summarize their main characteristics, using statistical graphics and other visualization methods.

ii. Getting Insights:
   1. Selection
   2. Preprocessing
   3. Transformation
   4. Data Mining
   5. Interpretation / Evaluation
   6. Knowledge

Process: EDA

A. EDA - Exploratory Data Analysis
   a. It is the approach of analyzing data sets to summarize their main characteristics, using statistical graphics and other visualization methods.

B. Getting Insights:
   a. Selection
   b. Preprocessing
   c. Transformation
   d. Data Mining
   e. Interpretation / Evaluation
   f. Knowledge

C. Data Cleaning:
   a. It is very important to look through the data and make sure it is clean, and begin exploring relationships between features and target variables.
      i. Looking at Data Types
      ii. Checking for Missing Values
      iii. Statistical Overview

Process: EDA

H. Visualizing

a. Correlating

b. Creating the Scatter Plot

i. To start looking at the relationships between features, we can create scatter plots to further visualize the way the different classes relates with each other.

E. Modeling

a. A model is the abstract representation of the data and the relationships in a given dataset

F. Train Test Split

a. Once we separate the features from the target, we can create a train and test class. As the names suggest, we will train our model on the train set, and test the model on the test set. We will randomly select 80% of the data to be in our training, and 20% as test.

G. Standardize

a. This puts the numbers on a consistent scale while keeping the proportional relationship between them.

Process: EDA

H. Logistic Regression Model

   a. In statistics, the logistic model is a statistical model that models the probability of one event taking place by having the log-odds for the event be a linear combination of one or more independent variables.

I.