Predictive Modeling

Department of Mathematical Sciences
Ying Sha, Professor

Michigan Tech

Publicly Available Data Sets

http://archive.ics.uci.edu/ml

The University of California (Irvine) is a well-known location for classification and regression data sets.

http://www.kdnuggets.com/datasets

The Association for Computing Machinery (ACM) has a special interest group on Knowledge Discovery in Data (KDD). The KDD group organizes annual machine learning competitions.

http://fueleconomy.gov

A web site run by the U.S. Department of Energy’s Office of Energy Efficiency and Renewable Energy and the U.S. Environmental Protection Agency that lists different estimates of fuel economy for passenger cars and trucks.

http://www.cheminformatics.org

This web site contains many examples of comput

Publicly Available Data Sets

http://www.ncbi.nlm.nih.gov/geo

The NCBI GEO web site is "a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high throughput functional genomic data submitted by the scientific community."

https://www.kaggle.com

Kaggle is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

Competitions on predictive modeling:

In 2019, Rose-Hulman Institute of Technology host the 36th Undergraduate Mathematics Conference on April 20 - 21, 2019. The link is http://www.rose-hulman.edu/mathconf/

About Predictive Modeling:

- Deal with *data analysis* with a specific focus on the *practice of predictive modeling*.
- May stir associations such as **machine learning**, **pattern recognition**, and **data mining**.
- Involves the process of developing a model in a way that we can understand and **quantify the model's prediction accuracy on future, yet-to-be-seen data**.
- This class guides you to the **predictive modeling process**. You will learn about **the approach** and to **gain intuition** about the many **commonly used and modern, powerful models**.
- Statistical and mathematical techniques are discussed, but our motivation is to **describe** the techniques way that helps develop **intuition for its strengths and weaknesses** instead of its **mathematical underpinnings**.

- You should have some knowledge of **basic statistics**, including **variance, correlation, simple linear regression, and basic hypothesis testing** (e.g. *p*-values and test statistics).

- **R is used** throughout this class. An introduction and start-up guide for R are in the Appendix B.

Chapter 1. Introduction

What is predictive modeling?

- People are faced with questions to know future events.
- We earnestly want to make the best decisions towards that future.
- We are predicting future events given the information and experience we currently have (predictors).
- As information has become more readily available via the internet and media, we routinely use this information to help us make decisions, for example:
  - Google
  - WebMD
  - E*TRADE

These tools take our current information, sift through data looking for patterns that are relevant to our problem, and return answers.

The process of developing these kinds of tools has been called

- Machine Learning
- Artificial Intelligence
- Pattern Recognition
- Data Mining
- Predictive Analytics
- Knowledge Discovery

Each field approaches the problem using different perspectives and tool sets,

The ultimate objective is the same: *to make an accurate prediction*. For this book, we will pool these terms into the commonly used phrase ***predictive modeling***.

- What is Predictive Modeling

  - ➢ Geisser (1993) -- the process by which a model is created or chosen to try to best predict the probability of an outcome.

  - ➢ Kuhn and Johnson -- the process of developing a mathematical tool or model that generates an accurate prediction.

- There is increasing presence of predictive models (Levy 2010):

  - ➢ Google global machine uses it to interpret cryptic human queries.
  - ➢ Credit card companies use it to track fraud.
  - ➢ Netflix uses it to recommend movies to subscribers.

Conclusion: Predictive models now *permeate our existence.*

While predictive models guide us towards more satisfying products, better medical treatments, and more profitable investments, they regularly generate inaccurate predictions and provide the wrong answers.

For example:

- Have not received an important e-mail (incorrectly identified the message as spam),
- Misdiagnose diseases, and
- Erroneously buy and sell stocks.

Predictive Models regularly fail because

- ➢ Do not account for complex variables such as human behavior,
- ➢ Inadequate pre-processing of the data,
- ➢ Inadequate model validation,

  Validation process can involve

  - ✓ Analyzing the goodness of fit of the model,
  - ✓ Checking whether the model's predictive performance deteriorates substantially when applied to data that were not used in model estimation.

- ➢ Over-fitting the model to the existing data, or
- ➢ Only explore relatively few models.

This class will try to help you to

- Produce reliable, trustworthy models, and
- Provide intuitive knowledge of a wide range of common models.

The objectives of this course are to provide:

- Foundational principles for building predictive models,
- Intuitive explanations of many commonly used predictive modeling methods for both classification and regression problems,
- Principles and steps for validating a predictive model
- Computer code to perform the necessary foundational work to build and validate predictive models

1.1 Prediction versus Interpretation

- There is trade-off between prediction and interpretation.

- The objective of predictive model is not to understand why something will (or will not) occur.

- We are primarily interested in accurately projecting the chances that something will (or will not) happen.

- Focus of this type of modeling is to optimize prediction accuracy.

  - ➢ The primary interest of predictive modeling is to generate accurate predictions.
  - ➢ A secondary interest may be to interpret the model and understand why it works.
  - ➢ The unfortunate reality is that as we push towards higher accuracy, models become more complex and their interpretability becomes more difficult.
  - ➢ This is almost always the trade-off we make when prediction accuracy is the primary goal.

1.2 Key Ingredients of Predictive Models

- If a predictive signal exists in a set of data, many models will find some degree of that signal regardless of the technique or care placed in developing the model.

- But the best, most predictive models are fundamentally influenced by a modeler with expert knowledge and context of the problem.

- This expert knowledge should first be applied in obtaining *relevant* data for the desired research.

- Irrelevant information can drive down predictive performance of many models.

- Subject-specific knowledge can help separate potentially meaningful information from irrelevant information, eliminating detrimental noise and strengthening the underlying signal.

- To summarize,
  - The foundation of an effective predictive model is laid with *intuition* and *deep knowledge of the problem context*.
  - That process begins with *relevant* data, another key ingredient.
  - The third ingredient is a *versatile* computational tools.

1.3 Terminology

- *Sample*, *data point*, *observation*, or *instance* refer to a single, independent unit of data, such as a customer, patient, or compound.

- *Sample* can also refer to a subset of data points, such as the training set sample.

- *Training set* consists of the data used to develop models

- *Test* or *validation* sets are used solely for evaluating the performance of a final set of candidate models.

- *Predictors*, *independent variables*, *attributes*, or *descriptors* are the data used as input for the prediction equation.

- *Outcome*, *dependent variable*, *target*, *class*, or *response* refer to the outcome event or quantity that is being predicted.

- *Continuous* data have natural, numeric scales.

- *Categorical* data, otherwise known as *nominal*, *attribute*, or *discrete* data, take on specific values that have no scale.

- *Model building*, *model training*, and *parameter estimation* all refer to the process of using data to determine values of model equations.

1.4 Example Data Sets and Typical Data Scenarios

- In this section, we briefly explore a few examples of predictive modeling problems and the types of data used to solve them.
- We can see the diversity of the problems as well as the characteristics of the collected data.

1. ***Music***

- Published as a contest data set on the TunedIT web site.
- The objective was to develop a predictive model for classifying music into six categories.

- Data:
  - There were 12,495 music samples;
  - 191 characteristics were determined;
  - The response categories were not balanced (Fig. 1.1)
  - All predictors were continuous; many were highly correlated; the predictors spanned different scales of measurement.

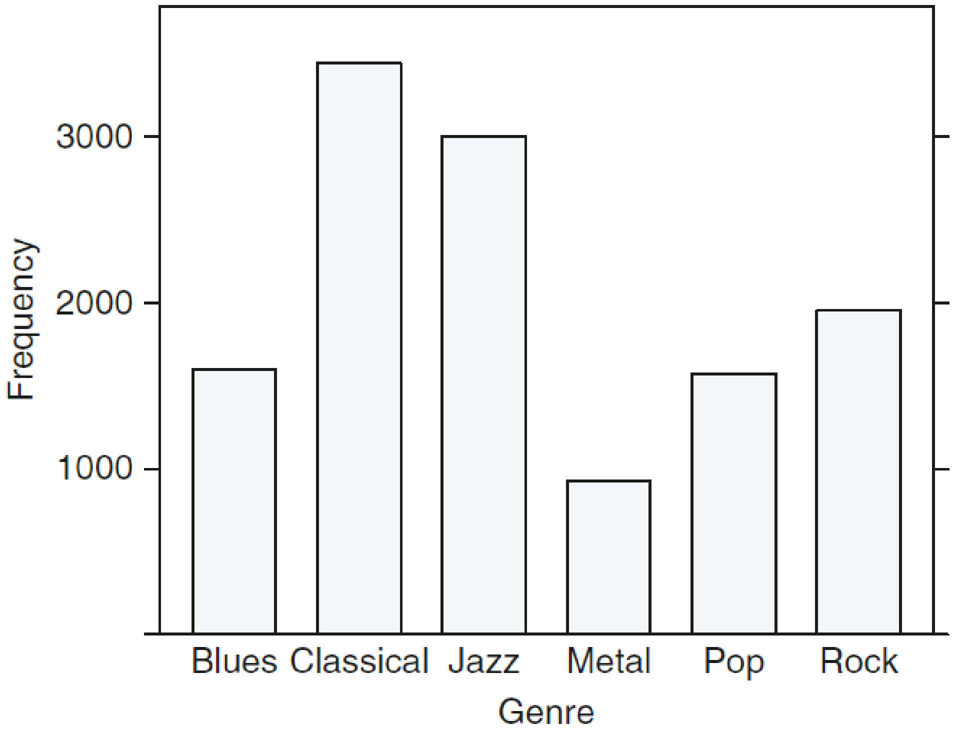

Fig. 1.1: The frequency distribution of genres in the music data

2. Grant Applications

- From a competition on the Kaggle web site.

- The objective was to develop a predictive model for the probability of success of a grant application.

- Data:
  - The database consisted of 8,707 grants between the years 2005 and 2008 for model building.
  - The test set contained applications from 2009 to 2010.
  - There are 249 predictors. Grant status (either "unsuccessful" or "successful") was the response and was fairly balanced (46% successful).

    The web site notes that current Australian grant success rates are less than 25%. Hence the historical database rates are not representative of Australian rates.

- Predictors include measurements and categories such as Sponsor ID, Grant Category, Grant Value Range, Research Field, and Department and were continuous, count, and categorical.
- Another notable characteristic of this data set is that many predictor values were missing.

3. Comparisons Between Data Sets

- Response data
  - Continuous: the distribution may be symmetric or skewed
  - Categorical response data: the distribution may be balanced or unbalanced.

- Predictors
  - Continuous, count, and/or categorical;
  - Missing values?
  - On different scales of measurement?
  - High correlation or association? (redundant information)
  - Sparse?

    A majority of samples contain the same information while only a few contain unique information.

- The relationship between the number of samples (*n*) and number of predictors (*P*).

  - ➢ If n $\gg$ P,
    - ✓ all predictive models handle this scenario.

  - ➢ If n < P
    - ✓ multiple linear regression or linear discriminant analysis cannot be directly used.
    - ✓ recursive partitioning and *K*-nearest neighbors (*K*NNs) can be used directly.

- Different kinds of models handle different data sets

  - ➢ **Partial least squares** can be used to handle correlated predictors.

  - ➢ **Recursive partitioning** is unaffected by predictors of different scales, can be used for missing data; but has a less stable partitioning structure when predictors are correlated.

  - ➢ **Multiple linear regression** cannot handle missing predictor information.

- **In summary**
  - We must have understanding of the predictors and the response for any data set prior to attempting to build a model.
  - Most data sets will require some degree of pre-processing in order to expand the universe of possible predictive models and to optimize each model’s predictive performance.

1.5 Overview

There are four parts in this book.

Part I: General Strategies

- Data pre-processing (Chap. 3)
  - Data transformations
  - The addition and/or removal of variables
  - Binning continuous variables
- Data resampling (Chap. 4)

  Data spending and methods for spending data in order to appropriately tune a model and assess its performance.

Part II: Traditional and modern regression techniques when modeling a continuous outcome

- Regression models using linear combinations of the predictors, including
  - Linear regression,
  - Partial least squares, and
  - L_1 regularization.
- Regression models that are not based on simple linear combinations of the predictors, including
  - Neural networks,
  - Multivariate adaptive regression splines (MARS),
  - Support vector machines (SVMs), and
  - *K-nearest neighbors* (*K*NNs).

- Tree-based models (not covered in this class)
    - Regression trees,
    - Bagged trees,
    - Random forests,
    - Boosting, and
    - Cubist.

Part III: Traditional and modern regression techniques when modeling a categorical outcome -- Predictive classification models

- Classification models using linear combinations of the predictors, including
  - Linear discriminant analysis,
  - Quadratic discriminant analysis,
  - Regularized discriminant analysis, and
  - Partial least squares discriminant.
- Regression models that are not based on simple linear combinations of the predictors, including
  - Flexible discriminant analysis,
  - Neural networks,
  - SVMs,
  - *K*NNs,
  - Naive Bayes, and
  - Nearest shrunken centroids.
- Tree-based models (not covered)

Part IV: Other important considerations when building a model or evaluating its performance (not covered)

- Feature selection techniques to find the most relevant predictors.
- Various methods for quantifying predictor importance.

There is a computing section at the end of each chapter.

1.6 Notation

$$\begin{aligned}
n &= \text{ the number of data points} \\
P &= \text{ the number of predictors} \\
y_i &= \text{ the } i\text{th observed value of the outcome, } i = 1 \ldots n \\
\widehat{y_i} &= \text{ the predicted outcome of the } i\text{th data point, } i = 1 \ldots n
\end{aligned}$$

$\bar{y} =$ the average or sample mean of the n observed values of the outcome

$\mathbf{y} =$ a vector of all n outcome values

$x_{ij} =$ the value of the jth predictor for the ith data point, $i = 1 \ldots n$ and $j = 1 \ldots P$

$\bar{x}_j =$ the average or sample mean of n data points for the jth predictor, $j = 1 \ldots P$

$\mathbf{x}_i =$ a collection (i.e., vector) of the P predictors for the ith data point, $i = 1 \ldots n$

$\mathbf{X} =$ a matrix of P predictors for all data points; this matrix has n rows and P columns

$\mathbf{X}' =$ the transpose of $\mathbf{X}$; this matrix has P rows and n columns

$C =$ the number of classes in a categorical outcome

$C_\ell =$ the value of the ℓth class level

$p =$ the probability of an event

$p_\ell =$ the probability of the ℓth event

$Pr[.] =$ the probability of event

$\sum_{i=1}^{n} =$ the summation operator over the index i

$\mathbf{\Sigma} =$ the theoretical covariance matrix

$E[\cdot] =$ the expected value of $\cdot$

$f(\cdot) =$ a function of .; $g(\cdot)$ and $h(\cdot)$ also represent functions throughout the text

$\beta =$ an unknown or theoretical model coefficient

$b =$ an estimated model coefficient based on a sample of data points