Chapter 1

Chapter 1. Introduction

predictive modeling: the process of developing a mathematical tool or model that generates an accurate prediction

reasons why predictive models fail, 1. inadequate pre-processing of the data, 2. inadequate model validation 3. unjustified extrapolation (application of the model to data that reside in a space which the model has never seen), 4. over-fitting the model to the existing data

1.1 Prediction versus interpretation

There is a trade-off between the prediction accuracy and causality interpretation

1.2 Key ingredients of predictive models

Intuition and deep knowledge of the problem context + relevant data + versatile computational toolbox

In the end, predictive modeling is not a substitute for intuition, but rather a complement

Humans usually make better predictions when they are provided with the results of statistical prediction

1.3 Terminology

Sample, data pointm observation, instance Training set, test or validation sets Predictors, independent variables, attributes, descriptors Outcome, dependent variables, target, class, response Continuous data and categorical data Model building, model training, parameter estimation

1.4 Example data sets and typical data scenarios

Music genre Grant applications Hepatic injury Permeability Chemical manufacturing process Fraudulent financial statements

Dimensions: Number of samples, number of predictors Response characteristics: Categorial or continious, balanced/symmetric, unbalanced/skewed, independent Predictor characteristics: continious, count, categorical, correlated/associated, different scales, missing values, sparse, symmetric/skewed, balanced,unbalanced

1.5 Overview

1.6 Notation