Predictive Analytics
Regression and Classification

Lecture 1 : Part 1

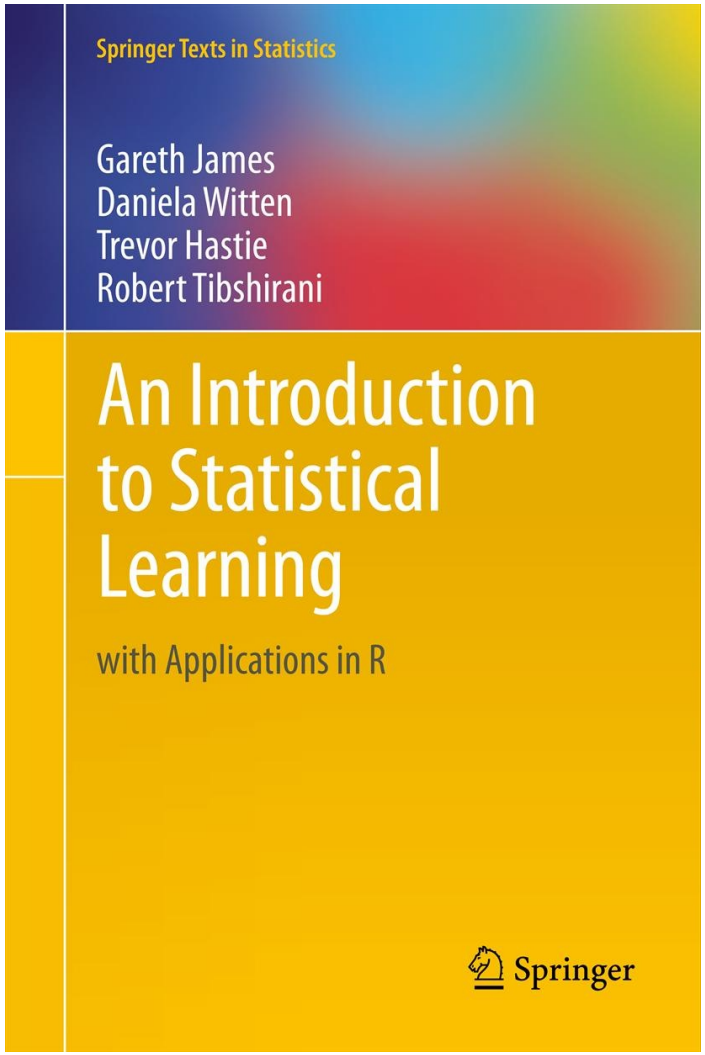
Sourish Das

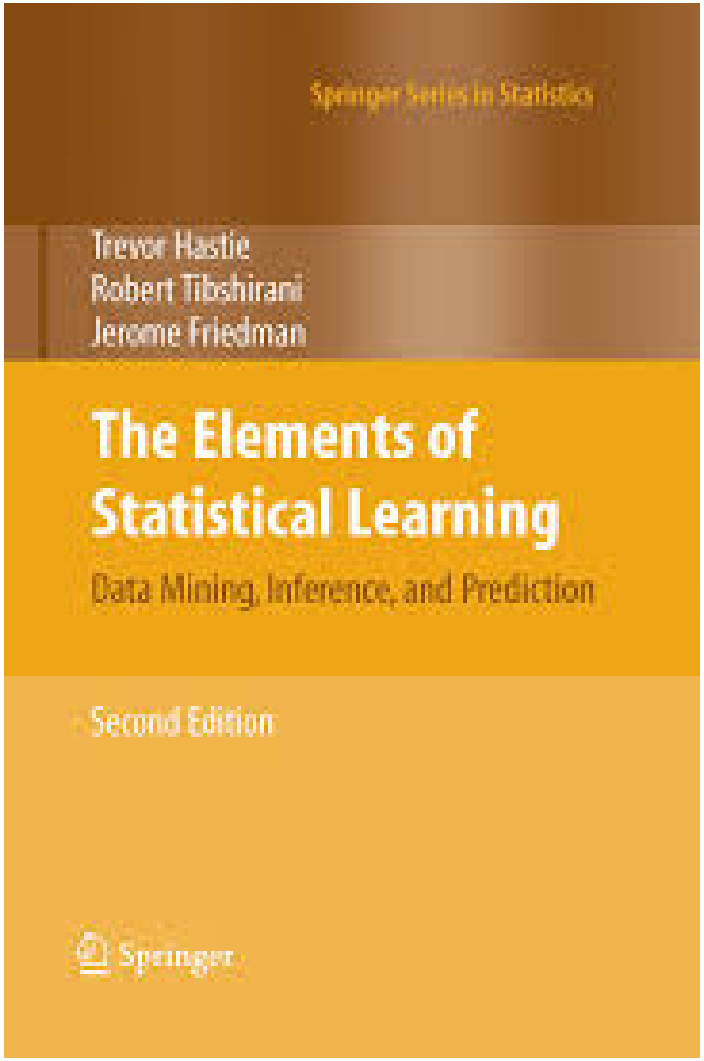
Chennai Mathematical Institute

Aug-Nov, 2020

Introduction

Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Reading material

- *Data Mining; Concepts and Techniques*, Jiawei Han and Micheline Kamber, Morgan Kaufman (2006).
- *Web Data Mining*, Bing Liu, Springer Verlag (2007).

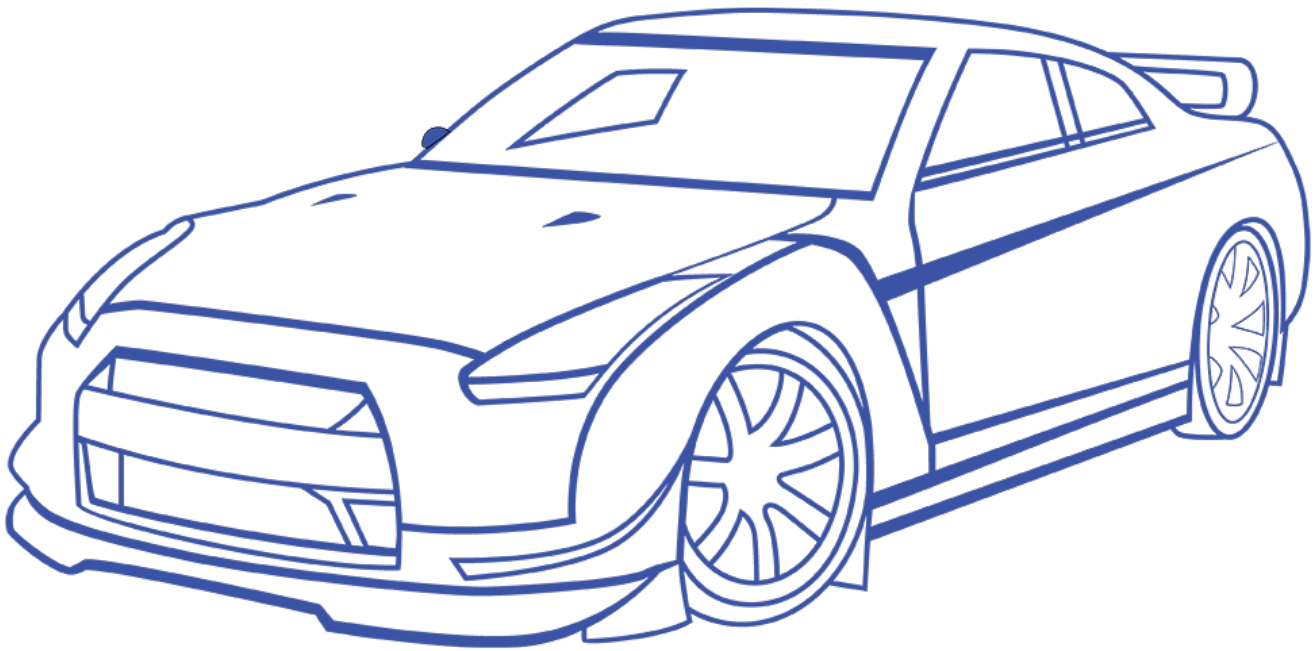
For a good introduction to text mining and information retrieval, please see.

- *An Introduction to Information Retrieval*, Christopher D Manning, Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press (2009). (Available online at `http://www-nlp.stanford.edu/IR-book`).

Supervised Learning

Motivating Examples of Supervised Learning

Ex 1 Given the different features of a new prototype car, can you predict the mileage or 'miles per gallon' of the car?



Motivating Examples of Supervised Learning

Ex 1 Given the different features of a new prototype car, can you predict the mileage or 'miles per gallon' of the car?

	mpg	cyl	disp	hp
Mazda RX4	21.0	6	160	110
Mazda RX4 Wag	21.0	6	160	110
Datsun 710	22.8	4	108	93
Hornet 4 Drive	21.4	6	258	110
.....				
Prototype	?	4	120	100

- Note that your objective is to predict the variable `mpg`.

- We are going to use `mtcars` data set in `R`.

Motivating Examples of Supervised Learning

Ex 2 Given the credit history and other features of a loan applicant, a bank manager want to predict if loan application would become good or bad loan!!

- Note that your objective is to predict the label of the loan good or bad!

How to identify if a problem is predictive analytics problem?

- Ask a question to your client or collaborator: "**Do you want to predict something**?"
- If the answer is 'yes' - then ask which variable?
- Check if that variable is available in the database.
- if yes - then you can consider it as a predictive analytics problem.

Supervised learning

- Supervised learning algorithms are trained using labeled data.
- For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs).
- Typically,

$$y = f(X),$$

where y is target variable and X is feature matrix

- Objective: Learn $f(.)$

Supervised learning

- Supervised learning

$$y = f(X)$$

typically are of two types:

1. Regression : target variable y is continuous variable - e.g., income, blood pressure, distance etc.

2. Classification: target variable y is categorical or label variable - e.g., species type, color, class etc.

Data : Quantitative Response

x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n
x^*_{11}	x^*_{12}	$\cdots$	x^*_{1p}	$y^*_1 =?$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x^*_{m1}	x^*_{m2}	$\cdots$	x^*_{mp}	$y^*_m =?$

- $D_{train} = (X, y)$, is the traing dataset, where X is the matrix of predictors or features, y is the dependent or target variable.
- $D_{test} = (X^*, y^* =?)$ is the test dataset, where X^* is the matrix of predictors or features, and y^* is missing and we want to forecast or predict y^*

Data : Qualitative Response

x_{11}	x_{12}	$\cdots$	x_{1p}	G_1
x_{21}	x_{22}	$\cdots$	x_{2p}	G_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	G_n
x^*_{11}	x^*_{12}	$\cdots$	x^*_{1p}	$G^*_1 = ?$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x^*_{m1}	x^*_{m2}	$\cdots$	x^*_{mp}	$G^*_m = ?$

- Qualitative variables are also referred to as *categorical* or *discrete* variables as well as *factors*.

In the next part of Lecture 1,

we will initiate the discussion on Regression.

Thank You

sourish@cmi.ac.in