Monday Group Presentation on Missing Data

Sam Shi

Ryerson University

November 10, 2016

Overview

1. Missing data
   - Types of Missing Data
   - Why it is important
   - Common ways to deal with missing data

2. Methods

Types of Missing Data

supposed you want to model weight Y as a function of gender X and you do a survey asking for Y and X, in the end there are some missing values(data point missing or data attribute missing), below are possible scenarios,

- Missing completely at random(MCAR)
  No particular reasons why the data is missing, such as, it can be someone dropped the survey paper, hard recognizable hand-writting, etc. (data are rarely MCAR)
- Missing at random(MAR)
  It can happen that one gender X would less likely to disclose their weight information than the other.
- Missing not at random(MNAR)
  Missing value itself is related to why it is missing, e.g. a person with higher weight Y would more likely not fill out the weight blank on survey.

why it is important

- Easy to occur very common

	Division		
	Alpha	Bravo	Charlie
Sales		$ 11,500,000	
Net operating income		$ 920,000	$ 210,000
Average operating assets	$ 800,000		
Margin	4 %	%	7 %
Turnover	5		
Return on investment (ROI)	%	20 %	14 %

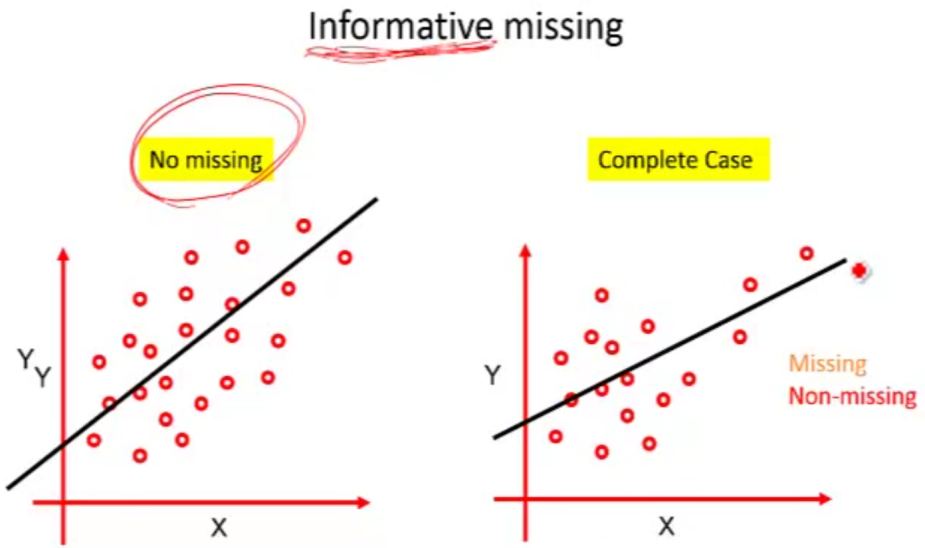
- Nearly all standard statistical methods presume complete information for all the variables included in the analysis; Machine learning models need to have complete input.[Wiki, C4.5]

Improvements from ID.3 algorithm [edit]

C4.5 made a number of improvements to ID3. Some of these are:

- Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the thresh
- Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Handling attributes with differing costs.
- Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

- Create bias [Missing Data Analysis]



1. Loss of statistical power
2. Regression slope is biased

Common ways to deal with missing data

A quick summary before we introudce dealing methods, [Computerphile]

- Listwise deletion (or complete case analysis)
- Imputation methods
- Multiple Imputation
- Maximum Likelihood
- Bayesian simulation methods
- Hot deck imputation methods

Blocks of Highlighted Text

Block 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Block 2

Pellentesque sed tellus purus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis magna at risus dictum tempor eu vitae velit.

Block 3

Suspendisse tincidunt sagittis gravida. Curabitur condimentum, enim sed venenatis rutrum, ipsum neque consectetur orci, sed blandit justo nisi ac lacus.

Multiple Columns

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

References


C4.5: https://en.wikipedia.org/wiki/C4.5_algorithm


Missing Data Analysis: https://www.youtube.com/watch?v=QAvSj2TWZy0


The Trouble with Missing Data
https://www.youtube.com/watch?v=oCQbC818KKU