

Monday Group Presentation on Missing Data

Sam Shi

Ryerson University

November 11, 2016

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-writing, 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	Beta	Charlie
Sales		\$ 11,500,000	
Net operating income		\$ 820,000	\$ 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]

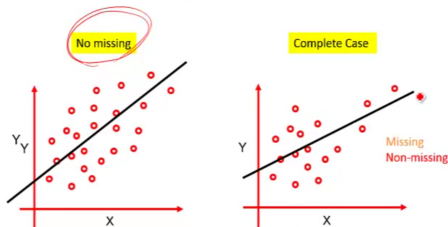
Improvements from ID3 algorithm [\[edit\]](#)

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 threshold
- 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]

Informative missing



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

Common ways to deal with missing data

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

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

Listwise deletion

Subject	Age	Gender	Income
1	29	M	\$40,000
2	45	M	\$36,000
3	81	M	--missing--
4	22	--missing--	\$16,000
5	41	M	\$98,000
6	33	F	\$60,000
7	22	F	\$24,000
8	--missing--	F	\$81,000
9	33	F	\$55,000
10	45	F	\$80,000

Advantage:

- Easy to implement, no special computation method requires
- It is valid if the missing data is MCAR
- If the proportion of deleted data is small, e.g. $< 5\%$

Disadvantage:

- Can exclude a large portion of data
- Missing data are MCAR rarely happens in reality
- Introduce bias

– picture from Wiki

Mean imputation: Use mean value of available data to represent missing data.

- dasdas

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>