Missing values

Data Mining

Edgar Roman-Rangel.
edgar.roman@itam.mx

Department of Computer Science.
Instituto Tecnológico Autónomo de México, ITAM.

Outline

Missing Values

Collaborative filtering

Linear regression

Outliers

Introduction

Finding missing values in real date is rather common.

Different alternatives can be put in place to handle missing values.

- Ignore.
- Fill in (imputation).
- Gather more data.

Ignore record

If only a small number of values are missing, we can simply ignore (delete) record from our analysis.

- This might become a problem as the number of missing values increases. E.g., Consider a dataset of 30 variables, with 5% of missing values, uniformly spread across all data.

Filling in

- Assign the mean, median or mode of such a variable.
- Use a constant value.
- Find a proxy variable.
- Find a highly correlated variable and use it for prediction (collaborative filtering, similar to linear regression).

Notice: No new information is added. It only allows us to perform computations.
Moreover, we induce some sort of bias when filling in the blanks.

Gathering more data

- Requires designing a data collection plan.
- In practice, it might be infeasible.

Missing same feature in many records

- Mean or median might lack of actual meaning.
- If feature is not crucial, drop it.
- Find a highly correlated variable.
- Gather more data.

Outline

Missing Values

Collaborative filtering

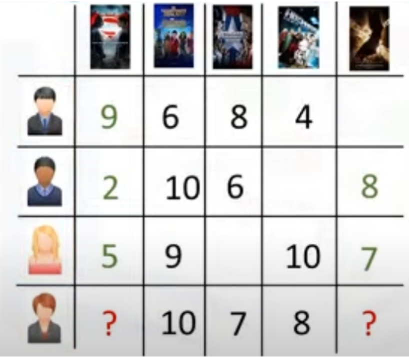
Linear regression

Outliers

Introduction

It is base on the assumption that there exists a relationship between points and variables.

Also used by recommender systems to find relationships between users and products: “I might like things that my friends like”.



	9	6	8	4	
	2	10	6		8
	5	9		10	7
	?	10	7	8	?

Process

For a given incomplete data point,

1. Using the available (filled in) features, find the most similar point in the dataset.
2. A common choice is to use the cosine similarity,

$$s_{i,j} = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}.$$

Fill in the missing value (n-feature) of the i-th point by,

- the value corresponding to the most similar point, or
- a weighted average of the values of (all) other points,

$$\mathbf{x}_{i,n} = \frac{\sum_j (s_{i,j})(\mathbf{x}_{j,n})}{\sum_j s_{i,j}}.$$

Outline

Missing Values

Collaborative filtering

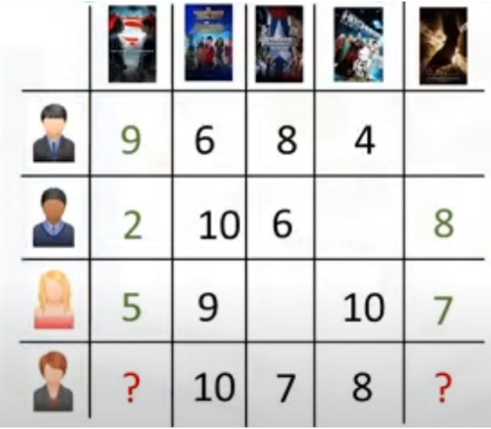
Linear regression

Outliers

Filling it by linear regression

We could also learn a regressor for our incomplete variable, using a complete feature from the dataset.

For instance, predict feature x_1 using x_2 as its predictor.



	9	6	8	4	
	2	10	6		8
	5	9		10	7
	?	10	7	8	?

Outline

Missing Values

Collaborative filtering

Linear regression

Outliers

Outlier detection

We can process outliers with similar treatments as missing values.

- Simply ignore them, if the dataset is large enough.
- Replace their value using a scaling process (to be seen).

References

Jiawei Han, Micheline Kamber, Jian Pei.
“Data Mining Concepts and Techniques”. Ch 2.2, and Ch 3.2.
Elsevier. 2012.

Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, Kenneth C.Lichtendahl, Jr.
“Data Mining for Business Analytics: Concepts, Techniques, and Applications in R”. Ch 3.
Wiley. 2018.

Q&A

Thank you!

`edgar.roman@itam.mx`