Knowledge Discovery and Data Mining

Lab 3 Data Cleaning I Handling Missing Data

Xuan Song
Songx@sustech.edu.cn

Intro

For this class we will be looking at the most common issue of any real data set: random data missing.

Basics

Let’s start by the very basics:

Reading csv file into python pandas:

pandas.read_csv()

See some example of data rows:

pandas.DataFrame.sample(n)

Data used in this class

First step: Melbourne Housing Snapshot

Second step: San Francisco Building Permits

How does MAR occur?

MAR: (data) Missing At Random

Common causes:

- Invalid/Bad probe readings
- Error/Loss of data packet during transmission

There are more types of data missing, but for this class, we only look at the most common MAR.

Check your data

Say, by using sampling, can you spot some missing data in the given dataset?

We can further investigate the amount of missing data with the following helpful functions:

pandas.DataFrame.isnull()

pandas.DataFrame.sum()
Also try **pandas.DataFrame.isnull().sum()**!

pandas.DataFrame.shape

numpy.product()

Methods

1.Remove row or column
2.Fill with pre-defined value
3.Use back-fill or forward-fill
4.Fill with median or mean

1. Remove row or column with missing data

There are many ways to deal with missing values, one easiest way is to drop the problematic data.

Try using:

pandas.DataFrame.dropna()

or

pandas.DataFrame.dropna(axis=1)

2. Fill missing data with predefined value

Another way is to add a default value, such as 0, to the data.

Try:

pandas.DataFrame.fillna(0)

3. Using forward-fill or back-fill to auto complete the data

Sometimes, 0 or 1990 does not feel ‘natural’, so we might want to fix the missing data by some better way.

Try:

pandas.DataFrame.fillna(method=‘ffill’,axis=0)

pandas.DataFrame.fillna(method=‘bfill’,axis=0)

4. Fill missing data with either mean or median value

For some data values, it is appropriate to fill missing by the statistically meaningful mean or median value.

Try:

pandas.DataFrame.mean()

pandas.DataFrame.median()

*Hint: you will still need fillna

Class Work

Apply all 4 different ways of missing data removal/imputation and compare the results on both datasets.

Practice Data Cleaning technics now as they will be very useful in the later part of the course.

End of Lab 3