

Case Study in Bayesian Analysis

Patrick Smith

Bayes Case Study II

OPENING

Bayes Case Study II

Today, we're going to focus on handling missing data

Bayes Case Study II

Today, we're going to focus on handling missing data

This is definitely more of an art than a science

Bayes Case Study II

Today, we're going to focus on handling missing data

This is definitely more of an art than a science

Our overall goal today will be to show there is some kind of "functional relationship" between the "missingness" of the data, and features found in our data. By doing this, we can categorize the kind of "missingness" we are dealing with for a particular data-set.

Bayes Case Study II

We'll briefly cover the 3 types of "missingness" and go straight to coding from there.

Type	Description
Missing Completely at Random	This is basically the best scenario, all NaN, NA, or blanks are distributed totally "at random" can be safely omitted
Missing at Random	This is less strong, but is "random" given the sample you are using. This is what we're aiming at for our analysis, functionally, we want to show that our missing data isn't dependent on data we haven't observed or accounted for in our data set
Missing not at Random	"There is a data generating process that yields missing values". Basically, it means there is some "pattern" to the 'missingness'

Overview: Introducing the Inclusion Indicator

Bayes Case Study II

What type of missingness are we most concerned with?

Bayes Case Study II

How can we define the data generating process?

Bayes Case Study II

The Inclusion Indicator

$$I = \begin{cases} 1 & x : \textit{missing} \\ 0 & x : \neg \textit{missing} \end{cases}$$

Bayes Case Study II

Codealong

Bayes Case Study II

How can we tell is if data is “missing at random”?

Bayes Case Study II

How can we tell is if data is “missing at random”?

- How can I apply what I've learned in Regressions to this problem?
- What are other metrics I could utilize to account for variation in data outside of regressions?

Bayes Case Study II

How can we tell is if data is “missing at random”?

Bayes Case Study II

Case 1: Using Logistic Regression to model the “missingness”

Bayes Case Study II

Case 2: Using KNN for Imputing Missing Data

Bayes Case Study II

Case 3: Imputing with Random Forest

Bayes Case Study II

Conclusion

More Unsupervised Learning Methods

Q & A