Introduction to Data Science: Communicating Results of Data Analysis

Héctor Corrada Bravo

University of Maryland, College Park, USA

CMSC320: 2020-05-03

For Today

Data Analysis Deliverables:

- Written analyses
- R packages

Written analyses

1. Title
2. Introduction and motivation
3. Description of dataset
4. Description of statistical and machine learning models used (Methods)
5. Results (including measures of uncertainty)
6. Conclusions (including potential problems)
7. References

https://leanpub.com/datastyle

Written analyses

Introduction and Motivation

Always lead with the question (task) you are addressing.

E.g.: "Can we use tweets about stocks to predict stock prices?" Not: "Can we use the Random Forest algorithm to learn a classifier that predicts stock prices"

E.g: "What are good predictors of student performance?" Not: "Can we use linear regression to predict student performance"

Written analyses

Description of dataset

Size: entities and attributes

Important: describe what you did to

1) obtain,
2) tidy the dataset.

Written analyses

Description of data analysis methods

Be specific, use equations when appropriate:

$$W = a + bH + e$$

where W is *weight*, H is *height* and e is an error term.

When appropriate mention distributional assumptions on e.

Written analyses

Description of data analysis methods

When using ML methods, describe:

- preprocessing (e.g., feature selection, transformations)
- algorithm choice (why is it appropriate)
- model selection and assessment (e.g., which classification metric and why)

Written analyses

Results

- Report estimates in the appropriate units
- Report estimates with uncertainty

We saw confidence intervals on our previous lectures with specific advise regarding their presentation. (*Note*: this also applies to prediction metrics)

Written analyses

Results

Important: Summarize importance of estimate (i.e., refer to the question you originally posed in introduction)

Why does this estimate address your question?

Written analyses

End matter

- Include potential problems with the analysis you carried out.
- Include references to the analysis methods used.

Graphics

Karl Broman's presentation on effective graphics:

http://tinyurl.com/graphs2017

Graphics

A few other notes on style:

- Make titles legible
- Annotate in plot if possible (see example data analysis early in semester)
- Include units in axis titles when appropriate
  - E.g., not appropriate in PC scatterplot

R packages

Case study: suppose you used data to create a classifier for diagnostic purposes. How do you share?

R packages is a reproducible, high-visibility way of publishing these types of results

- Consistent organization
- Standardized deployment

R packages

Case study: suppose you used data to create a classifier for diagnostic purposes. How do you share?

R packages is a reproducible, high-visibility way of publishing these types of results

Hadley's presentation on R packages

http://www.slideshare.net/hadley/r-packages

The book