What is data science?

Fort Collins Data Science Meetup

2016-12-08

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured

https://en.wikipedia.org/wiki/Data_science

Data science

=

Processes and systems

+

data

Data science

=

knowledge from data

Steps of a data analysis

1. Define the question
2. Define the ideal data set
3. Determine what data you can access
4. Obtain the data
5. Clean the data
6. Exploratory analysis
7. Statistical prediction/modeling
8. Interpret results
9. Challenge Results
10. Report results
11. Create Reproducible code

https://github.com/DataScienceSpecialization/courses/blob/master/05_ReproducibleResearch/lectures/structureOfADataAnalysis1.pdf

1. Define a question

- What knowledge do you want

- Example: Can I separate spam from non-spam?

2. Define the ideal data set

- What do you need to know to answer your question?

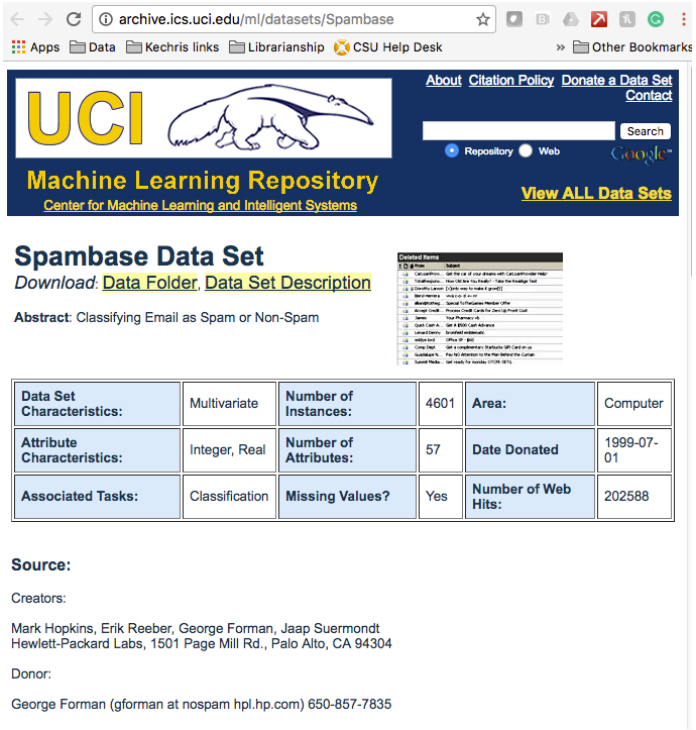
Type of question	Data
Descriptive	Whole population
Exploratory	Random sample with lots of variables
Inferential	Random sample of a particular population
Predictive	Predictive: training and test data
Causal	Data from randomized study
Mechanistic	Data about all components of a system

- Example: Which emails are spam?

3. What data can you access?

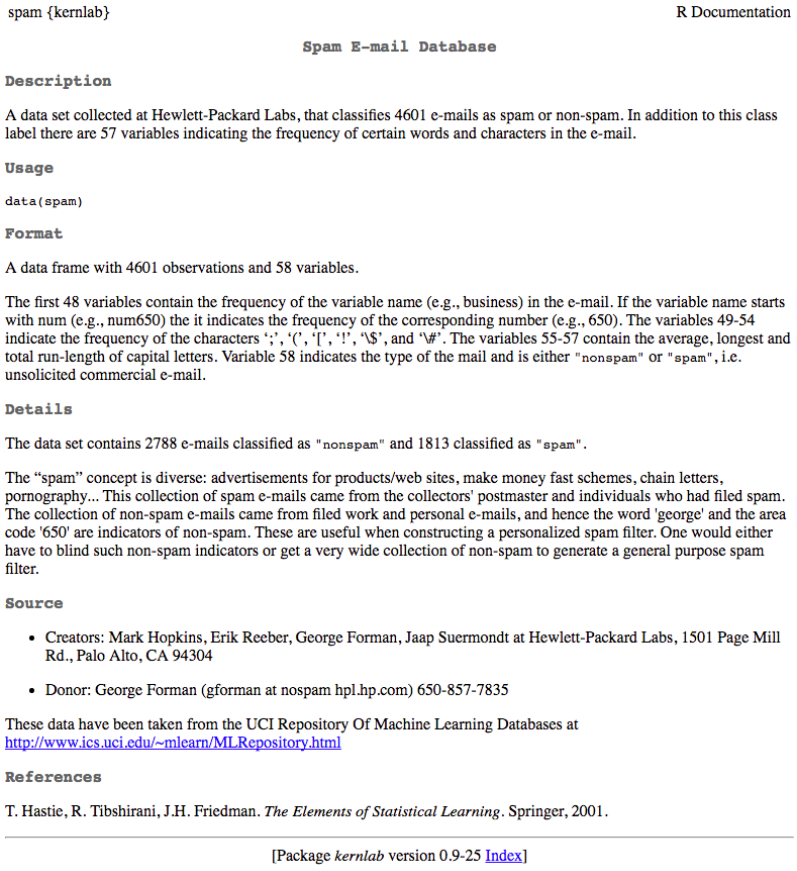
- Free vs. for a fee
- Terms of use
- Generate your own data

http://archive.ics.uci.edu/ml/datasets/Spambase



4. Obtain the data

- Optimal: Raw data
- Cite Source
- Ask permission
- Record URL if online



spam {kernlab} R Documentation

Spam E-mail Database

Description

A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail.

Usage

```
data(spam)
```

Format

A data frame with 4601 observations and 58 variables.

The first 48 variables contain the frequency of the variable name (e.g., business) in the e-mail. If the variable name starts with num (e.g., num650) the it indicates the frequency of the corresponding number (e.g., 650). The variables 49-54 indicate the frequency of the characters ';', '(', '[', '!', '\$', and '\#'. The variables 55-57 contain the average, longest and total run-length of capital letters. Variable 58 indicates the type of the mail and is either `"nonspam"` or `"spam"`, i.e. unsolicited commercial e-mail.

Details

The data set contains 2788 e-mails classified as `"nonspam"` and 1813 classified as `"spam"`.

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography... This collection of spam e-mails came from the collectors' postmaster and individuals who had filed spam. The collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

Source

- Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt at Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304
- Donor: George Forman (gforman at nospam hpl.hp.com) 650-857-7835

These data have been taken from the UCI Repository Of Machine Learning Databases at http://www.ics.uci.edu/~mlearn/MLRepository.html

References

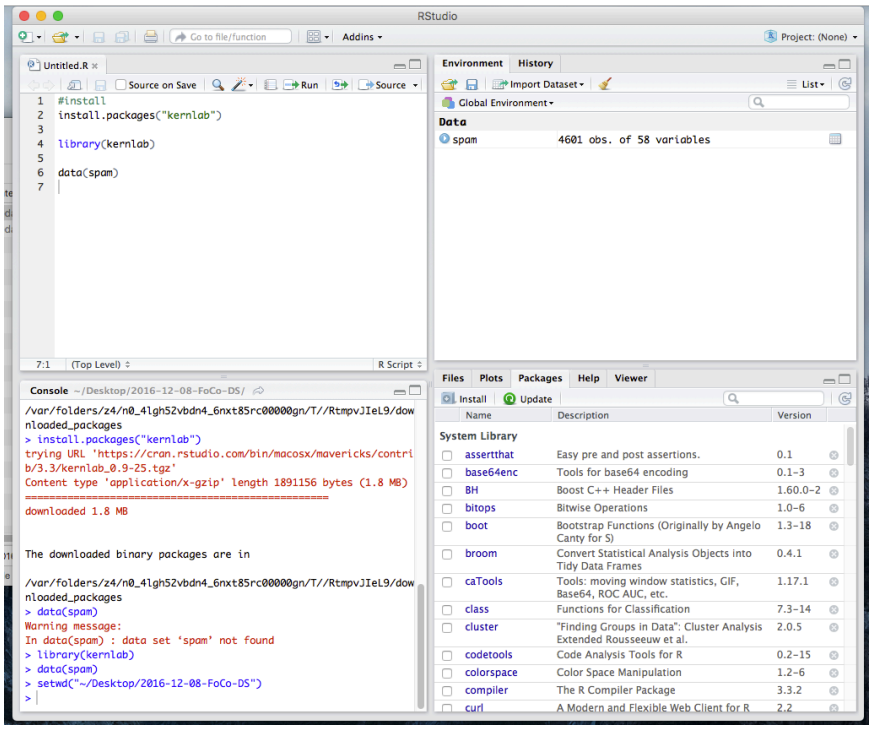
T. Hastie, R. Tibshirani, J.H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[Package *kernlab* version 0.9-25 Index]

http://search.r-project.org/library/kernlab/html/spam.html

5. Clean the data

- Where did it come from?
- Is it pre-processed?
- Reformatting/subsampling
  - need test and training set
- Do QC



6. Exploratory Analysis

- Look at summary stats
- Missing data?
- Exploratory plots
- Exploratory analysis

7. Statistical Prediction/modeling

- Based on exploratory analysis
- Methods depend on question
- Account for processing
- Report uncertainty

8. Interpret results

- Results match data and analysis type
- Explain
- Interpret coefficients
- Interpret measures of uncertainty

- The fraction of characters that are dollar signs can be used to predict if email is spam
- Anything with more than 6.6% dollar signs is classified as Spam.
- More dollar signs always means more spam in our model
- Our test set erroer rate was 22.4%

9. Challenge results

- Challenge all steps
- Challenge measures of uncertainty
- Challenge choices of terms in models
- Think of alternate analysis

10. Write up results

- Tell a story
  - State your question
  - Summarize relevant parts of your analysis
  - Order them to tell a story
  - Include pretty figures that support the story

- Can I use quantitative characteristics to classify email as spam?
- Collected data from UCI and created test and training sets. I explored relationships between spam status and variables. Chose logistic model on training set by cross validation.
- I applied this to test and got 78% accuracy
- # dollar signs is a reasonable predictor of spam
- We could probably do better than 78%. Maybe use a multivariate model. Maybe logistic regression wasn't the best choice?

11. Write reproducible code

- Learn markdown!
- Human readable
- Machine readable