Data is not flat

Working with the data is an art

Alisa Dammer
me: alisadammer.com

May 24, 2018

Structure

Introduction

Examples

Conclusion

Questions?

What do I need and how can I achieve it?

- From an idea to a MVP
- Steps required
- Things to consider beforehand

From the idea to a MVP

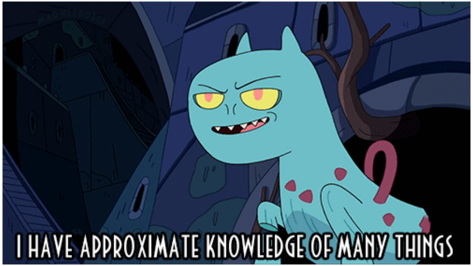
- Is it my core product?
  - Not that many...
- Is it an important feature?
  - Market advantage
  - Cost advantage
  - Hype advantage
- Is it a neat feature?
  - Is that feature so important?

Steps required

- Data (Legal issues, Cost, Data Mining, Big Data)
- Infrastructure (Data Engineers, Storage, DevOps, Big Data, Cost)
- Data Magic (Data Scientists, Statistics, Machine Learning, Deep Learning, Cost)
- Insights incorporation

Things to consider beforehand

- Fuzzy tasks
- Fuzzy QA
- Fuzzy results
- Cooooooost
- Data quality



Feature Engineering

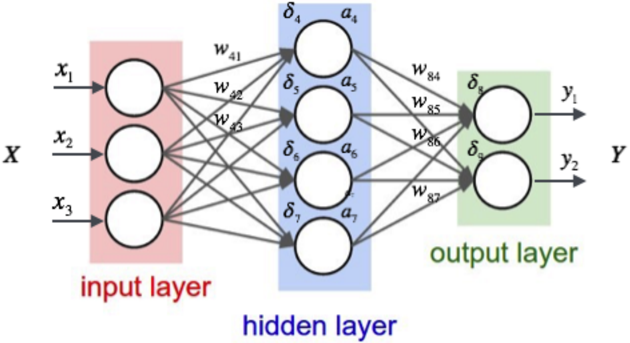
It is not:

- Part of the pre-processing
- Part of the data collection

It is:

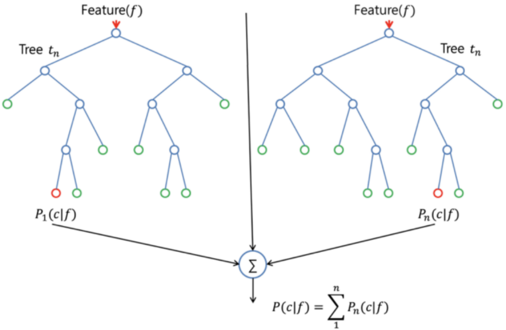
- Expert domain knowledge
- Manual and automated process
- Quite expensive to obtain
- More relevant than the model

Classification problem



`http://localhost:8888/notebooks/classification_example.ipynb`

Forecast problem



http://localhost:8888/notebooks/regression_case.ipynb#

Is the problem solved?

- What is feature engineering, again?
- Does it work?
- When to consider it

Thank you for attention!

Some useful information

- "Applied Predicitive Modeling" Max Kuhn, Kjell Johson
- https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-g
- https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
- https://elitedatascience.com/feature-engineering-best-practices
- https://en.wikipedia.org/wiki/Feature_learning
- https://en.wikipedia.org/wiki/Random_forest
- https://towardsdatascience.com/under-the-hood-of-neural-networks-part-1-fully-connected-5223b7f78528
- Coursera ML and Deep Learning courses from Andrew Ng
- https://medium.com/@curiousily/tensorflow-for-hackers-part-iv-neural-network-from-scratch-1a4f504dfa8
- https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Sources for images

- http://projectfreight.net/ws/project-freight-planning-part-3/
- https://www.workfromhomechristians.com.au/wp-content/uploads/2012/08/what-can-I-do1-300x300.jpeg
- https://qph.fs.quoracdn.net/main-qimg-a2bdcb6297b0091398767e4e4c69866
- https://cdn-images-1.medium.com/max/1600/1*QVIyc5HnGDWTNX3m-nIm9w.png
- https://cdn-images-1.medium.com/max/800/0*tG-IWcxL1jg7RkT0.png