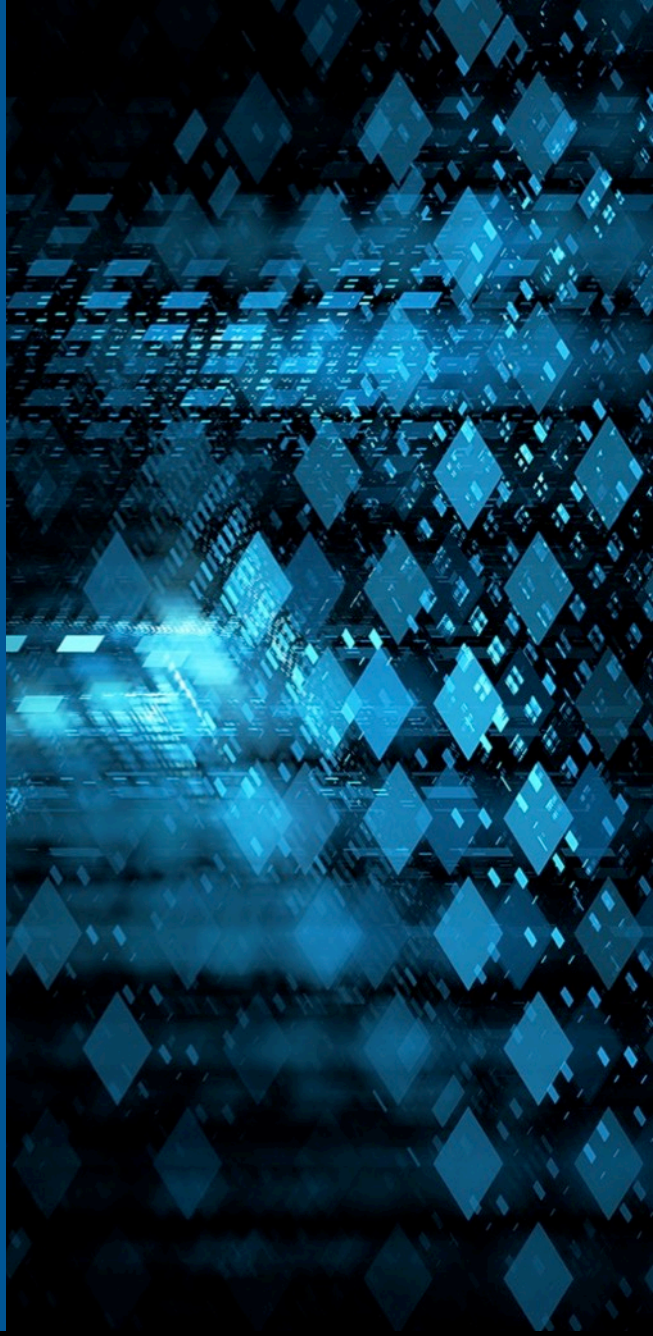
Myth-busting Machine Learning

Angela Horneman

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213



This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM18-0465

Introduction

What is Machine Learning?

Any method that lets a computer take data and learn rules from it for classifying or clustering other similar data.*

*See appendix for source and other accepted definitions

May, 2018

Take-away 1

When people claim they are using machine learning, get clarification on what type of method they are using

- classification or regression
- clustering
- hybrid clustering and classification
- statistical method
- time series analysis
- graph analysis

https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice

https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

http://www.ibmbigdatahub.com/blog/what-graph-analytics

May, 2018

Topics

Vocabulary

Machine learning steps

Classification

Clustering

Machine learning in cyber today

What I didn’t talk about

May, 2018

Vocabulary

Data: a set of records, containing a set of *features,* and each record corresponds to an instance of a "thing of interest".	Coin
Feature: a field in a record that tells something about that particular instance of a "thing of interest".	Color
Label: the "class" for a specific instance of a "thing of interest".	Penny
Model: the rules learned from training data that will be used to classify or cluster new data	Lincoln = Penny
Prediction: the class or cluster that an instance of new data will belong to according to the model.	Washington = ?

May, 2018

Example 1

Thing of interest = person

Classifications = person's occupation

Location| Pay range| Work hours| Work days| Occupation

Restaurant| 8-10/hr| 5am-10am| SMTW| Cashier

Office building| 12-15/hr| 8am-5pm| MTWHF| Admin assistant

May, 2018

Example 2

Thing of interest = access control logs

Classifications = cyber policy violation

Username| Device | Role| Time |Login Tries | Violation

jdoe| server1| user| 8:30AM| 1| no

jdoe| server1| root| 8:30PM| 4| yes

May, 2018

Machine Learning Steps

Machine Learning Steps

1. Use case development: figure out what you want to do and how you would go about doing it.
2. Data collection: get the data you need to work with.
3. Data cleaning: make the data you get usable by your machine learning algorithm.

https://insights.sei.cmu.edu/sei_blog/2017/06/machine-learning-in-cybersecurity.html

May, 2018

Machine Learning Steps cont.

4. Feature engineering: determine if any of the features in the original data can be used to make new features.
5. Model building and validation: use the chosen machine learning algorithm to build the model (train) and test the model (validate).
6. Deployment and monitoring: implement the model and watch the results. If accuracy or performance begins to decrease, go back to previous steps.

May, 2018

Myth 1

Myth: As long as there is a lot of data related to a “thing of interest” machine learning will provide new insight that will solve a problem.

Reality: Every ML algorithm is a tool. Each

- works with specific types of data
- in specific ways
- to give specific types of insight

To get useable insight, you have to choose the right data and the right tool.

May, 2018

Take-away 2

Well-defined use cases are key!

May, 2018

Classification

Classification

AKA: supervised learning

Requires a labeled dataset.

Looks at labels and determines criteria of features' values that correspond to specific labels

Color	Image	Grooves	Type
Copper	Lincoln	No	Penny
Silver	Washington	Yes	Quarter
Silver	Roosevelt	Yes	Dime
Silver	Jefferson	No	Nickel
Silver	Eisenhower	Yes	Half-dollar

May, 2018

Classification Example

Color| Image| Grooves

May, 2018

Classification Example

Color| Image| Grooves| Type

Copper | Lincoln | No | Penny

Silver | Washington | Yes | Quarter

Silver | Roosevelt| Yes | Dime

Silver | Jefferson | No | Nickel

Silver | Eisenhower| Yes | Half-dollar

Take-away 3

Machine learning is only as good as the data that was used to build the model.

- If data wasn't representative to begin with, the model is not accurate to end with.
- If features cannot distinguish between the classification, the resulting model cannot classify.

May, 2018

Myth 2

Myth: Lots of data or more data make ML work.

Reality: Data that is representative for the use case and environmental context PLUS data with discriminating features is required to make ML work.

Quality is more important than quantity!

Getting quality data and making it useable for machine learning is the most time consuming and expensive part of the analytical task.

May, 2018

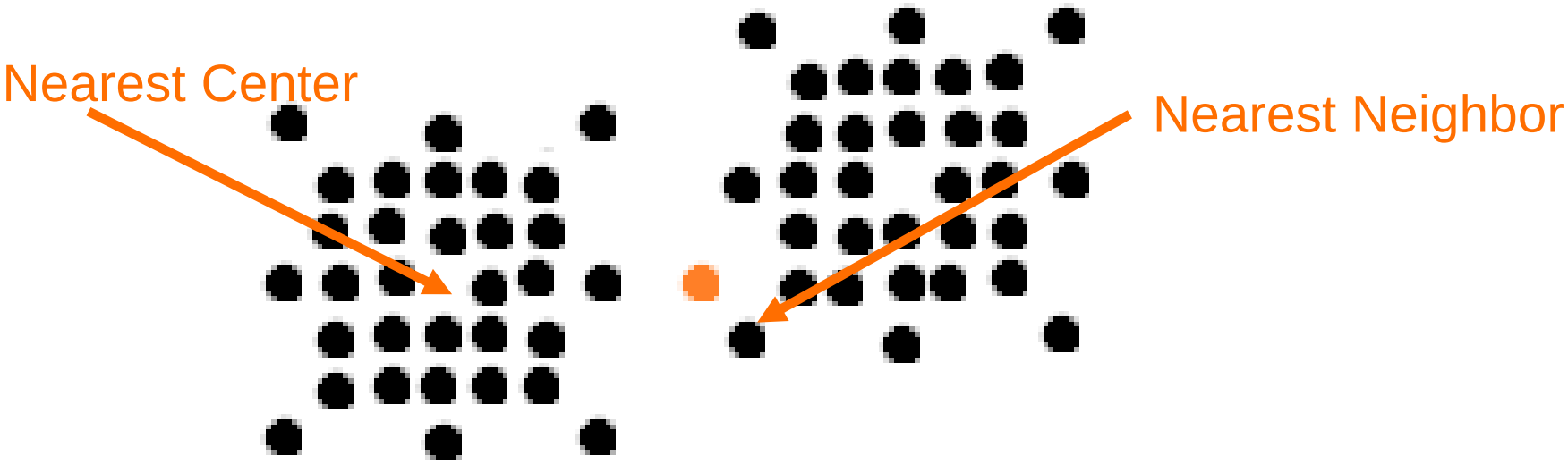
Clustering

Clustering

AKA: unsupervised learning

Does not require a labeled dataset.

Looks at records and determines some type of “closeness” between them.



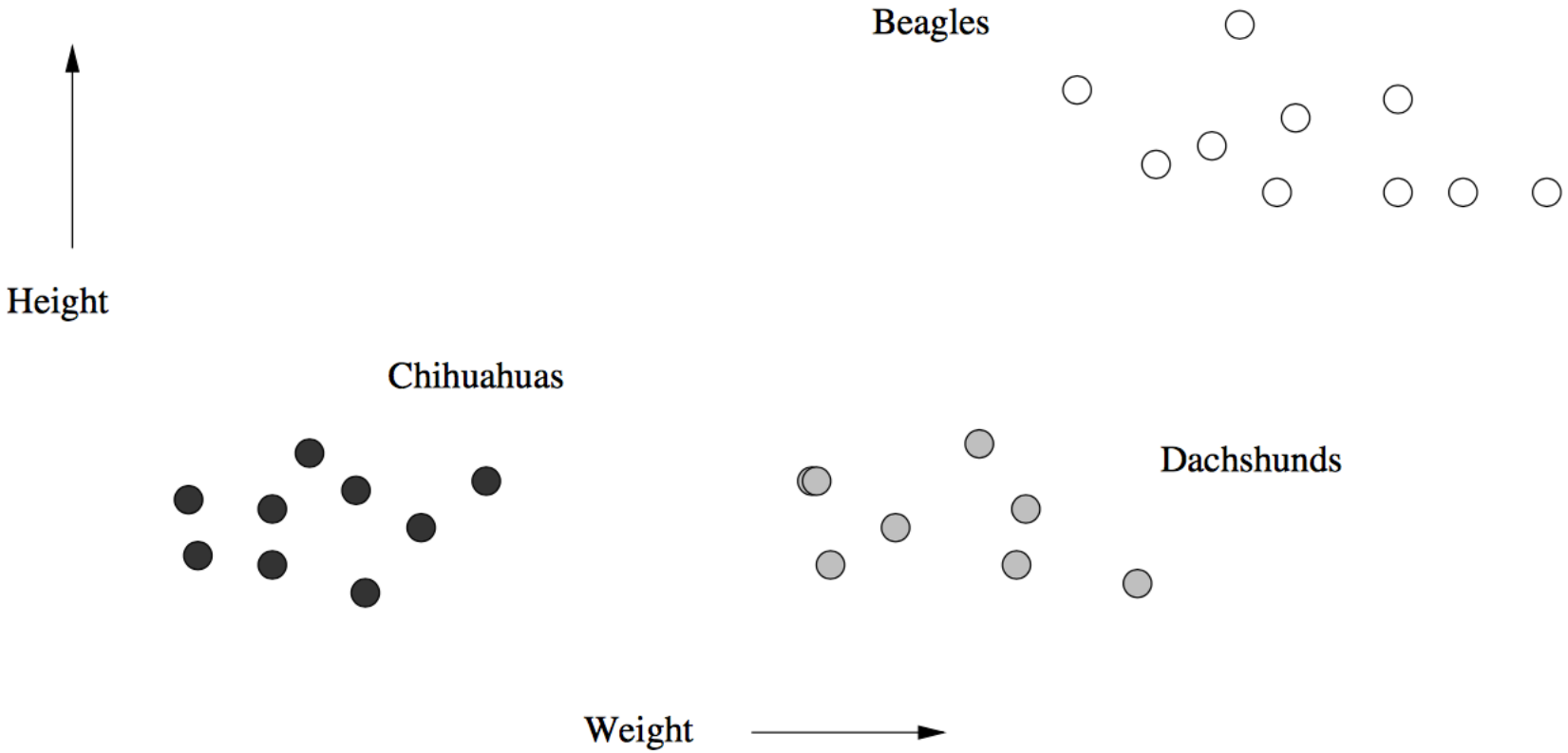
Myth 3

Myth: If an algorithm works with the type of data in the dataset, it is an appropriate choice. Furthermore, if there are multiple algorithms, the one that performs best on my test data is the best option.

Reality: How an algorithm works is an important consideration. Just because an algorithm was the ”best” out of a selection, does not mean it is an appropriate tool.

May, 2018

Clustering Example



This image comes from
http://infolab.stanford.edu/~ullman/mmds/ch7.pdf

May, 2018

Machine Learning in Cyber Today

In Academic and Research Literature

Not much that is practical.

Lots of use cases for retrospective Distributed Denial of Service (DDos) and network scan detection.

If you know of anything, I would love to know!

May, 2018

Myth 4

Myth: When machine learning can do something, it should be used to do it.

Reality: There is a lot that machine learning could do that can be implemented more simply, effectively, or with less overhead using other methods.

May, 2018

Commercial

Seems to be used in

- anti-malware
- website reputation
- User and Entity Behavior Analytics (UEBA)
- fraud detection

May, 2018

Take-away 4

ML is most successful when it is viewed as part of a larger process and not a stand-alone solution.

May, 2018

Scalability

Machine learning applications can run into many scalability related issues

- storage for data
- hardware expenses for processing (CPUs and RAM)
- processing time
  - training that may take days and need redone on a regular bases
  - processing that takes hours for each new batch of data

May, 2018

Take-away 5

Successful machine learning requires sufficient hardware to store and process data.

May, 2018

What I Didn’t Talk About

What I Didn’t Talk About

Issues of implementation (e.g. overfitting a model)

Privacy and security concerns

Artificial Intelligence, deep learning, active learning, reinforcement learning

May, 2018

Take-away Recap

1. Clarify what people mean by ML. Is it what you would expect?

2. Have a use case. Can someone explain how the application fulfills *your* use case?

3. Representative examples of the right data is required for building models. What data is needed to build the machine learning model so it works for *your* use case in *your* environment? Can someone tell you how you (or they) get that data and what it takes to make the data suitable for training the model?

May, 2018

Take-away Recap cont.

4. Use ML as part of a larger process and not a stand-alone solution. What human intervention or other analyses are needed to make the ML solution successful?

5. Scalability matters. What and how much hardware is required for your environment now and in two year?

May, 2018

Bibliography

Black 2016]

Black, Jason. Cutting Through the Machine Learning Hype. *Forbes*. November 16, 2016. http://www.forbes.com/sites/valleyvoices/2016/11/16/cutting-through-the-machine-learning-hype/#7150af5f7e96

[Litan 2016]

Litan, Avivah; Bussa, Toby; & Ahlm, Eric. The Fast-Evolving State of Security Analytics. ID: G00298030. *Gartner.* 2016. https://www.gartner.com/doc/3274217/fastevolving-state-security- analytics-

[McLellan 2016]

McLellan, Charles. Inside the Black Box: Understanding AI Decision-Making. *ZDNet*. December 1, 2016. http://www.zdnet.com/article/inside-the-black-box-understanding-ai-decision-making/

[Ossola 2016]

Ossola, Alexandra. How Machine Learning Works. *Popular Science*. February 23, 2016. http://www.popsci.com/how-machine-learning-works-interactive

Bibliography cont.

[Pauli 2017]

Pauli, Darren. Infosec Industry to Drive Machine Learning Spend Surge Says Analyst. *The Regis- ter*. January 31, 2017. http://www.theregister.co.uk/2017/01/31/infosec_ai_push/

[Scarpati 2014]

Scarpati, Jessica. Context-Aware Security: Big Benefits for Networks, but No Shortcuts. *Tech- Target*. August 2014. http://searchnetworking.techtarget.com/feature/Context-aware-security-Big- benefits-for-networks-but-no-shortcuts

[Sullivan 2015]

Sullivan, Danny. How Machine Learning Works, As Explained By Google. *MARTECH Today*. November 4, 2015. https://martechtoday.com/how-machine-learning-works-150366

[Thomas 2008]

Thomas, Ciza; Sharma, Vishwas; & Balakrishnan, N. Usefulness of DARPA Dataset for Intrusion Detection System Evaluation. International Society for Optics and Photonics. In *SPIE Defense and Security Symposium, Data Mining, Intrusion Detection, Information Assurance, and Data Networks*. Volume 6973. Pages 69730G 1–8. March 2008. doi:10.1117/12.777341
http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=837103

Appendix: Machine Learning Definition

The definition provided at the beginning is my synthesis of several commonly accepted definitions.

Arthur Samuel (1959) in *Computer Games*. “the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning. … Programming computers to learn from experience…”

https://www.mathworks.com/discovery/machine-learning.html

https://www.sas.com/en_us/insights/analytics/machine-learning.html

https://www.investopedia.com/terms/m/machine-learning.asp

https://www.techopedia.com/definition/8181/machine-learning

May, 2018