Fundamentals of Machine Learning

Instructor: Ekpe Okorafor
1. Accenture – Big Data Academy
2. Computer Science
African University of Science & Technology

Ekpe Okorafor PhD

Affiliations:

- **Accenture Digital – Big Data Academy**
  - ❑ Principal, Big Data & Analytics
- **African University of Science & Technology**
  - ❑ Professor, Computer Science / Data Science
  - ❑ Research Professor - High Performance Computing Center of Excellence

Research Interests:

- Big Data, Predictive & Adaptive Analytics
- Statistical Machine Learning
- Performance Modelling and Analysis
- Information Assurance and Cybersecurity.
- High Performance Computing & Network Architectures
- Distributed Storage & Processing
- Massively Parallel Processing & Programming
- Fault-tolerant Systems

Email: ekpe.okorafor@gmail.com; eokorafor@aust.edu.ng
Twitter: @EkpeOkorafor; @Radicube

Objectives

Objectives

- **What machine learning is**
- **What are three common machine learning techniques**
- **How organizations are applying these techniques**
- **What is the relationship between algorithms and data volume**

Outline

- **Overview**
- **The three C's of machine learning**
- **Importance of data and algorithms**
- **Essential points**
- **Conclusion**

Outline

- **Overview**
- **The three C's of machine learning**
- **Importance of data and algorithms**
- **Essential points**
- **Conclusion**

Fundamentals of Computer Programming

- **Let's first consider how a typical program works**
  - Hardcoded conditional logic
  - Predefined reactions when those conditions are met

```
$ cat spam-filter.py
#!/usr/bin/env python

import sys

for line in sys.stdin:
        if  Make MONEY Fa$t At Home!!!  in line:
                print  This message is likely spam
        if  Happy Birthday from Aunt Betty  in line:
                print  This message is probably OK
```

- **The programmer must consider all possibilities at design time**
- **An alternative technique is to have computers *learn* what to do**

What is Machine Learning

- **Machine learning is a field within artificial intelligence (AI)**
  - AI: the science and engineering of making intelligent machines
- **Machine learning focuses on automated knowledge acquisition**
  - Primarily through the design and implementation of algorithms
  - These algorithms require empirical data as input
- **Machine learning algorithms learn based on input provided**
  - Amount of data is often more important than the algorithm itself

What is Machine Learning (cont'd)

- **The output produced varies by application**
  - Product recommendations
  - Items grouped based on similarity
  - Possible diagnosis of a disease
- **These are examples of The Three C's of machine learning**

What is Machine Learning (cont'd)

- **The output produced varies by application**
  - Product recommendations
  - Items grouped based on similarity
  - Possible diagnosis of a disease
- **These are examples of 'The Three Cs' of machine learning**

Outline

- Overview
- **The three C's of machine learning**
- Importance of data and algorithms
- Essential points
- Conclusion

The ‘Three C’s’

- **Three established categories of machine learning techniques:**
  - Collaborative filtering (recommendations)
  - Clustering
  - Classification

Collaborative Filtering

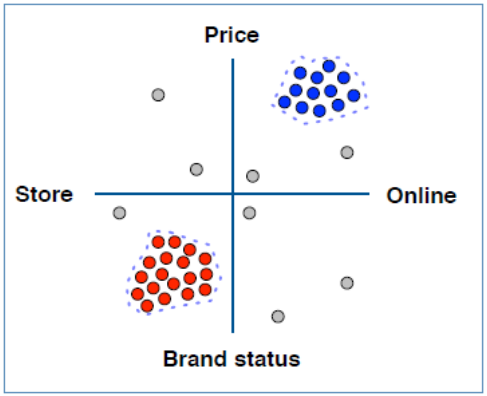
- **Collaborative filtering is a technique for recommendations**
  - It's one primary type of recommender system
  - We'll cover it in detail today
- **Helps users find items of relevance**
  - Among a potentially vast number of choices
  - Based on comparison of preferences between users

Applications Involving Collaborative Filtering

- **Collaborative filtering is domain agnostic**
- **Can use the same algorithm to recommend practically anything**
  - Movies (movielens, Netflix, etc)
  - Television (TiVO suggestions)
  - Music (Several popular music download and streaming services)
  - Colleges (Application to several colleges can be a aunting task)
- **Amazon uses CF to recommend a variety of products**

Clustering

- **Clustering algorithms discover structure in collections of data**
  - Where no formal structure previously existed
- **They discover what clusters ('groupings'), naturally occur in data**
  - By examining various properties of the input data
- **Clustering is often used for exploratory analysis**
  - Divide huge amount of data into smaller groups
  - Can then tune analysis for each group

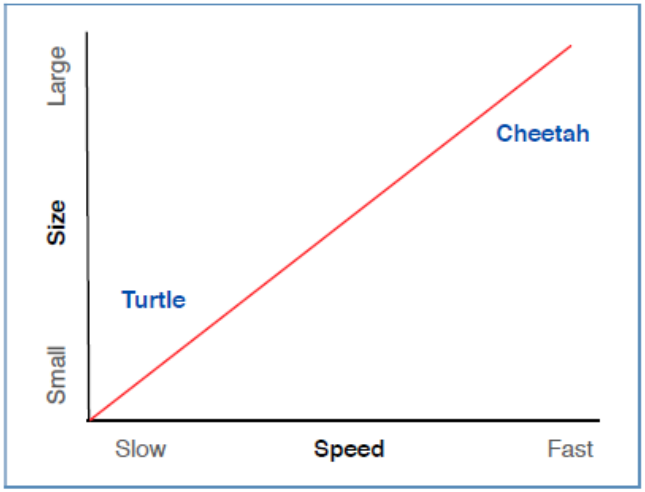


Applications Involving Clustering

- **Market segmentation**
  - Group similar customers in order to target them effectively
- **Finding related news articles**
  - Google News
- **Epidemiological studies**
  - For example, identifying cancer cluster and finding root cause
- **Computer vision (groups of pixels that cohere into objects)**
  - Related pixels clustered to recognize faces or license plates

Classification

- **The previous two techniques are *unsupervised learning***
  - The algorithm discovers recommendations or groups
- **Classification is a form of 'supervised' learning**
  - Requires training with data that has known labels
    - These are healthy cells, those are cancerous
  - Learns how to label new records based on that information



Applications Involving Classification

- **Spam filtering**
  - Train using a set of spam and non/spam messages
  - System will eventually learn to detect unwanted e/mail
- **Oncology**
  - Train using images of benign and malignant tumors
  - System will eventually learn to identify cancer
- **Risk Analysis**
  - Train using financial records of customers who do/don’t default
  - System will eventually learn to identify risk customers

Outline

- Overview
- The three C's of machine learning
- **Importance of data and algorithms**
- Essential points
- Conclusion

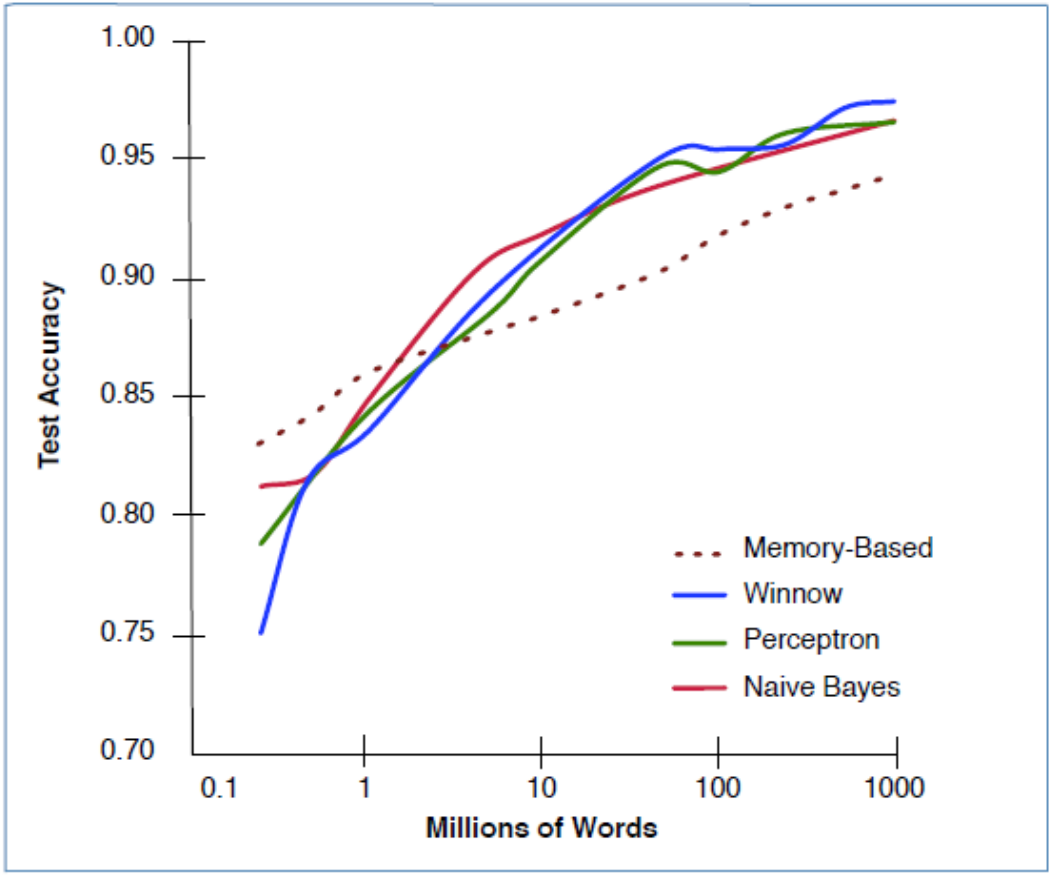
Relationship of Algorithms and Data Volume

- **There are many algorithms for each type of machine learning**
  - There is no overall best algorithm
  - Each algorithm has advantages and limitations
- **Algorithm choice is often related to data volume**
  - Some scale better than others
- **Most algorithms offer better results as volume increases**
  - Best approach = simple algorithm + lots of data

Relationship of Algorithms and Data Volume (cont'd)

It's not who has the best algorithms that wins.

It's who has the most data. [Banko and Brill, 2001]



Outline

- Overview
- The three C's of machine learning
- Importance of data and algorithms
- **Essential points**
- Conclusion

Essential Points

- **Machine learning algorithms learn based on data provided**
- **Collaborative filtering recommends items**
- **Clustering discovers how to group a set of items into subsets**
- **Classification is supervised learning that can identify item types**
- **More data is usually preferable to a better algorithm**

Outline

- Overview
- The three C's of machine learning
- Importance of data and algorithms
- Essential points
- **Conclusion**

Conclusion

In this section you have learned

- **What machine learning is**
- **What are three common machine learning techniques**
- **How organizations are applying these techniques**
- **What is the relationship between algorithms and data volume**