Notes - Machine Learning

Friday, September 15, 2017 10:44 PM

Quick Overview

NVIDIA Blog - What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?
NVIDIA Webinar - Deep Learning Demystified (23 min)

Microsoft Azure: Data Science for Beginners Video Series

The Five Questions Data Science can Answer (5 min)

1) Is this A or B? - Classification
2) Is this weird? - Anomaly Detection
3) How mush - or - How many? - Regression - Numerical Predictions
4) How is this organized? - Clustering
5) What should I do next? - Reinforcement Learning

Is Your Data Ready for Data Science? (5 min)

- Data Must be:
  - Relevant
  - Connected
  - Accurate
  - Enough to work with

Ask a Question You can Answer With Data (4 min)

- Ask a very specific that can be answered with a name or number
- How you ask a question can give you direction on what algorithm to use

Predict an Answer with a Simple Model (8 min)

- Example of using regression and SSE to fit parameters to a model
- Not that insightful of video

Copy other People's work to do Data Science (3 min)

- Sales pitch for Microsoft Azure Machine Learning and Cortana Machine Learning Gallery
- Not that insightful of video

Conceptual Resources

Brandon Rohrer's Blog - Data Science and Robots

- Blog filled with videos and posts about ML and data science
- Very good conceptual explanations without getting very technical

3 Brown 1 Blue - Neural Network Video Series

- Great animations and visualizations
- A little technical but not overwhelming
- Play at 1.5x speed

1) But what *is* a Neural Network? (19 min)
2) Gradient Descent, How Neural Networks Learn (21 min)
3) What is Backpropagation and What is it Actually Doing? (14 min)

4) Backpropagation Calculus (10 min)

YouTube Channel - DeepLearning.TV

- Cartoon animations on the concepts behind Artificial Neural Networks

Technical Resources

YouTube Channel - Mathematicalmonk

- Very advance lectures on ML algorithms, information theory, and statistics
- Play at 2x speed

Course Website - AP Monitor

- Dr. Hedengren's course website on dynamic optimization and control
- Modeling/optimizing for physical systems

Tutorials

Google Developers - Machine Learning Recipes Tutorial Series (python)

1) Hello World - Machine Learning Recipes #1 (7 min)
2) Visualizing a Decision Tree - Machine Learning Recipes #2 (7 min)
   - You might not have the library to perform the visualization but that's okay
3) What Makes a Good Feature? - Machine Learning Recipes #3 (6 min)
4) Let's Write a Pipeline - Machine Learning Recipes #4 (8 min)
5) Writing Our First Classifier - Machine Learning Recipes #5 (9 min)
6) Train an Image Classifier with TensorFlow for Poets - Machine Learning Recipes #6 (7 min)
   - Requires TensorFlow
7) Classifying Handwritten Digits with TF.Learn - Machine Learning Recipes #7 (7 min)
   - Requires TensorFlow
8) Let's Write a Decision Tree Classifier from Scratch: Machine Learning Recipes #8 (10 min)
   - A much more technical tutorial involving the inner workings of a decision tree

Buzzwords - Have these on your resume

Artificial Intelligence (AI)

- Computers being programmed to do something "smart"

Machine Learning (ML)

- Subfield of AI
- Statistical analysis of data, lots and lots of data
    - Can be very simple
        - Example: Plotting points and drawing a line through it
    - Can be very complex
        - **Artificial neural networks (ANN)**
- Essential it's optimization

Deep Learning (DL)

- Subfield of ML
- An ANN that is many layers "deep" in complexity
- Requires sample data and GPUs to "train" the ANN

Big Data

- Three V's: Velocity, Variety, & Volume

        - Velocity: Data comes in at a high rate
        - Varity: Incoming data might be different structure
        - Volume: Data is large in size
    - Massive amounts of data
        - 100s of Terabytes or so in size

Cloud Computing

- Outsourcing your computation/hardware to a third party service that you connect to through the internet
    - Amazon Web Services (AWS) - Dominant Cloud Provider
    - Microsoft Azure - Closest competitor to AWS
    - IBM and Google Cloud - Trying to catch up to AWS and Azure
- The Cloud providers offer many different services that range from storage, to webhosting, to computation and such
- Three main "levels" of service
    - Infrastructure as a Service (IaaS)
    - Platform as a Service (PaaS)
    - Software as a Service (SaaS)

Internet of Things (IoT)

- Connecting things to the internet that normally aren't connected
    - Examples: fridge, thermostat, car engine
    - Examples: General Electric is starting to embed its products with sensors and internet connection
    - Examples: "Smart" devices - Smart phone, Smart Watch, Smart Home - Basically it's all IoT
- Arduinos and Raspberry Pis are good at this
- **Industrial Internet of Things (IIoT)**
    - IoT but for industry application

Edge Computing

- Collect small amounts data from sensors and perform some computation onsite before sending it to a larger collection of data for storage or further computation - Usually sent to the Cloud

Hadoop Distributed File System (HDFS)

- Sometimes referred to as "Hadoop"
- An opensource database framework designed for storing data across a cluster of machines
    - YARN and HDFS are the core components in the architecture
    - YARN
        - The Data Operating System that runs on top of HDFS
        - It does Cluster Resource Management
        - Just a technical portion of Hadoop to be aware of
- HDFS is a distributed data base - The data is distributed across a cluster of machines
    - Good with Big Data
- There are many appendage-like services that may or may not be installed on top of HDFS and YARN
    - **Apache Spark**
        - An important in-memory data processing engine that people use
        - It's kind of like a HDFS version of a SQL database
        - Experience with Apache Spark is considered very valuable

Data Science

- Basic Workflow: Data Capture -> Storage -> Processing -> Visualization
    - Data Capture

        - Implementing sensors to collect data - usually IoT
    - Storage
        - Basically storing data
        - There might be some quick preprocessing done to format the data
    - Processing
        - Basically it's high performance computing
        - Ideally you would skip the storage step and feed your data into processing as it is captured
    - Visualization
        - Making the processed data in human readable form - usually charts, graphs, and heat maps
            - This can be very hard for high dimensional results
- Any parts of the workflow could be considered Data Science