TDS

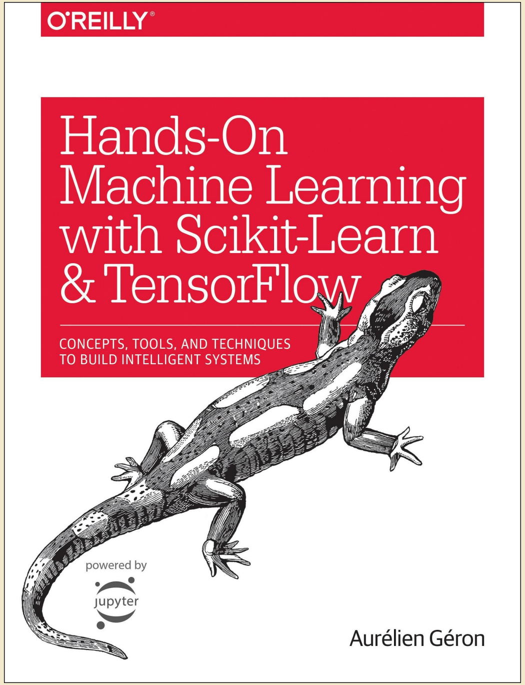
Intro to Machine Learning

October 1st, 2018

YouTube Link: https://youtu.be/PAqWw-PrySc

Resources

- *Hands-On Machine Learning with Scikit-Learn and Tensorflow*
- *The author says I can!*



O'REILLY

Hands-On Machine Learning with Scikit-Learn & TensorFlow

CONCEPTS, TOOLS, AND TECHNIQUES
TO BUILD INTELLIGENT SYSTEMS

powered by jupyter

Aurélien Géron

Resources (Cont.)

- Tom Mitchell’s book on machine learning: https://www.cs.ubbcluj.ro/~gabis/ml/ml-books/McGrawHill%20-%20Machine%20Learning%20-Tom%20Mitchell.pdf

Overview

- What is Machine Learning?
- Types of Machine Learning
    - Supervised vs Unsupervised vs Reinforcement
    - Batch vs Online
    - Instance vs Model-based
- Main Challenges of Machine Learning
    - Insufficient Quantity of data
    - Poor Quality of data
    - Non-representative-ness of training data
    - Feature Relevance: selection, extraction, gathering
    - Model Complexity: over- and under-fitting
- Testing and Validation
    - Training and test sets
    - Generalization Error
    - Cross Validation

What is Machine Learning?

The science (and art) of programming computers so they can *learn from data*

Technical definition:

> “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves w/ experience E” - Tom Mitchell

Our goal with the image recognition project, using this definition:

> “A program that, after **observing each handwritten letter** (E), it **predicts the character** (T), and **with backpropagation and gradient descent** (P), it’s capacity to **predict** (T) improves with each **observation** (E).”

Types of ML - Supervision, Reinforcement

- Supervised: Data comes with labels
- Unsupervised: ...Data comes without labels (cluster on similarity, and others)
- Reinforcement: learning system observes, decides, and is rewarded/penalized

Supervised Learning

- Data comes with labels
- Two kinds of prediction algorithms:
  - Classification: algorithm *classifies* an observation into a category
  - Regression: algorithm predicts a target numerical value
- Tag yourself: Classification, or regression?
  - Median Household income by zip code?
  - Our project: Neural Network reads an image and guesses the letter or number it represents?
- ‘Important supervised learning algorithms:’
  - k-Nearest Neighbors, Linear Regression, Logistic Regression, Support Vector Machines (SVMs), Decision Trees and Random Forests, Neural Networks*

Unsupervised Learning

- You guessed it - data comes *without* labels
- Main unsupervised learning algorithms: clustering, visualization (dimensionality reduction), association rule learning
- ‘Important unsupervised learning algorithms’:
  - Clustering: k-means Hierarchical Cluster Analysis (HCA), Expectation Maximization
  - Visualization and Dimensionality Reduction: Principal Component Analysis (PCA), Kernel PCA, Locally-Linear Embedding (LLE), t-distributed Stochastic Neighbor Embedding (t-SNE)
  - Association rule learning: Apriori, Eclat

Test Yourself: Supervised vs Unsupervised

- Which belong to supervised and which belong to unsupervised?
    - Facebook user data (age, dob, liked pages, connections)
    - Iris Dataset: https://archive.ics.uci.edu/ml/datasets/iris
    - Gapminder Data: https://www.gapminder.org/tools/#$chart-type=bubbles

Reinforcement Learning

- The learning system observes, acts, and is rewarded (or penalized)
- MarI/O: https://www.youtube.com/watch?v=qv6UVOQ0F44

Challenges of ML - Insufficient Data

“For a toddler to learn what an apple is, all it takes is for you to point to an apple and say ‘apple’ (possibly repeating this procedure a few times). Now the child is able to recognize apples in all sorts of colors and shapes. Genius.” - Géron

However, machine learning algorithms take lots and LOTS of data for it to work properly.

Challenges of ML - Non-representative Training Data

- As in statistics, too small data has large sampling noise: https://en.wikipedia.org/wiki/Margin_of_error
- Large datasets are not immune: sampling bias
- (Provide scenarios)

Challenges of ML - Data Quality

- Missingness?
- Cleanliness? (Text input? *Curse you, PowerSchool*)

Challenges of ML - Feature Relevance

- *How do you know if you’ve trained on the right variables?*
- **Feature selection**: select the most useful features train on
- **Feature extraction**: combine existing features to produce a more useful one
- Gather more data!

Challenges of ML - Model Complexity

- Overfitting the training data: *I had a bad intro Chem teacher….*
    - Regularization: constrain the model to make it simpler. Can specify the degree of flexibility on parameters, via a *Hyperparameter* (Not to be confused)
- Underfitting the training data: model does not accommodate complexity
- *To the book's repo!*

Testing and Validation

- Training and Test Sets:
    - Training: what the algorithm ...trains... on (minimize training error), usually 80% of data
    - Test: what the algorithm ...tests… on (after training, inspect test error), usually 20% of data
    - *Next week: Iris Dataset*
- *If training error is low, but generalization (test) error is high, then it may be overfitting!*
- Cross Validation: *Define multiple training and test set partitions on the dataset, run over each partition, then average their resulting parameter estimates*

Questions?