Lesson 02

CSC357: Advanced Topics: Machine Learning

13 January 2020

1 Questions

- Give an example of a medical application of machine learning.
- Give an example of a use of machine learning in manufacturing.
- How might machine learning change the practice of journalism?

2 Notes

- 1. The Machine Learning Landscape
  - What is Machine Learning?
  - Why Use Machine Learning?
  - Examples of Applications
    - Analyzing images of products on a production line to automatically classify them
    - Detecting tumors in brain scans
    - Automatically classifying news articles
    - Automatically flagging offensive comments on discussion forums
  - Types of Machine Learning Systems
  - Main Challenges of Machine Learning
  - Testing and Validating

3 Notes

- I. The Fundamentals of Machine Learning
  - 1. The Machine Learning Landscape
    - What is Machine Learning?
    - Why Use Machine Learning?
    - Examples of Applications
      - Analyzing images of products on a production line to automatically classify them
      - Detecting tumors in brain scans
      - Automatically classifying news articles
      - Automatically flagging offensive comments on discussion forums
      - Summarizing long documents automatically
      - Creating a chatbot or a personal assistant
      - Forecasting your company's revenue next year, based on many performance metrics
      - Making your app react to voice commands
      - Detecting credit card fraud
      - Segmenting clients based on their purchases so that you can design a different marketing strategy for each segment
      - Representing a complex, high-dimensional dataset in a clear and insightful diagram
      - Recommending a product that a client may be interested in, based on past purchases
      - Building an intelligent bot for a game
    - Types of Machine Learning Systems
      - Supervised/Unsupervised Learning
      - Batch and Online Learning
      - Instance-Based Versus Model-Based Learning
    - Main Challenges of Machine Learning
      - Insufficient Quantity of Training Data
      - Nonrepresentative Training Data
      - Poor-Quality Data
      - Irrelevant Features
      - Overfitting the Training Data
      - Underfitting the Training Data
      - Stepping Back
    - Testing and Validating
      - Hyperparameter Tuning and Model Selection
      - Data Mismatch

4 Questions

- What is Géron talking about when he refers to the dimensions of a dataset?
- How might a bank use machine learning to control losses?
- What do you suppose the difference betweeen a parameter and a hyperparameter might be?
- The director of a chain of bookstores wants to segment the business' customers. Draw upon your own experience. What are some segments that the director might find?
- Think of some of the binary ways in which we can categorize machine learning algorithms.

  What kind of algorithm will be appropriate for a problem with these characteristics:

  - We expect to receive a continuous stream of new data that we can use to improve our model.
  - Our clients require quick responses.

- Think of some of the binary ways in which we can categorize machine learning algorithms.

  What kind of algorithm will be appropriate for a problem with these characteristics:

  - We have records that describe many voters.
  - Each record includes the values of several or many characteristics of a voter (perhaps, for example, age, sex, address, occupation, and others)
  - Although few voters possess identical characteristics, we suspect that there are several types of voters. By 'type,' we mean voters who are likely to share common concerns and respond positively to similar appeals.
  - We begin our project not knowing how many types of voters are in our population or what defines each type—these are the things that we want to discover.

5 Notes

- B. Machine Learning Project Checklist
  - Frame the Problem and Look at the Big Picture
  - Get the Data
  - Explore the Data
  - Prepare the Data
  - Shortlist Promising Models
  - Fine-Tune the System
  - Present Your Solution
  - Launch!

6 A Checklist for Machine Learning Projects

- frame the problem and look at the big picture
- get the data
- explore the data to gain insights
- prepare the data to better expose the underlying data patterns to Machine Learning algorithms
- explore many different models and shortlist the best ones
- fine-tune your models and combine them into a great solution
- present your solution
- launch, monitor, and maintain your system

7 Questions

The first item on our checklist tells us to "frame the problem and look at the big picture."

Suppose that Cornell College has hired you to write software that will help us predict which of the high school students who express interest in the college will accept an offer of admission and matriculate.

How will you "frame the problem and look at the big picture?"

8 Notes

- what is machine learning
  - statistics, artificial intelligence, computer science
  - predictive analytics
  - statistical learning
- solving a problem by writing decision rules (if/then)
  - requires expert's knowledge of problem and methods for its solution
  - algorithm works for only the one problem
  - any change in the nature of the problem may require very extensive rewriting of program
- cannot write a program to recognize faces
  - represention of images in computer's memory (array of pixels) very different from representation in human mind
  - work began in 1960s
  - solved in 2001?
  - rapid advance after 2012
  - large databases of faces
  - better cameras
  - more powerful computers
  - improved algorithms
- supervised learning
  - recognize handwritten zipcodes on envelopes at Post Office
    - no special training required to label data
    - cheap, easy (but slow)
  - distinguish benign from malignant tumors in medical images
    - experts needed to label data (expensive)
    - complex, expensive instruments needed to produce images
    - need to protect patients' privacy (respect laws, ethical principles)
  - detect fraudalent credit card transactions
    - company has records of customers' transactions already
    - customers report fraud
    - data collection is easy!
  - unsupervised algorithms
    - categorize blog posts (identify common themes)

        - categorize customers (segment market)
        - spot patterns of unusual behavior on website

- data often stored in tables
    - row → sample / data point
    - column → feature
- feature extraction / feature engineering
- things to keep in mind
    - understand data
    - how data relates to task (goal / objective)
- questions to ask
    - what kinds of answers are we seeking?
    - how to phrase questions as machine learning problem?
    - what kind of data / features are needed?
    - what are the measures of success?
    - how does the machine learning component connect with other pieces of a product or service?

9 Questions

- What are some of the synonyms (or near-synonyms) for machine learning? What are some related fields?
- Facial recognition is a recent achievement. What are some of the kinds of innnovation that made facial recognition practical?
- We will find many data sets take the form of tables. In these tables, what do find in each row? in each column?

  Have you seen this kind of organization elsewhere in your study of computer science?

10 Notes on Chapter 1

- machine learning is decades old
- OCR (optical character recognition was early application
- spam filters first mainstream application (1990s)
- kinds of ML
  - supervised vs. unsupervised
  - online vs. batch
  - instance-based vs. model-based
- what we will learn
  - workflow of typical ML project
  - principal challenges
  - how to evaluate ML system
  - how to fine-tune ML system
- ML is...
  - art and science
  - programming computer to learn from data
  - giving computers ability to learn without being explicitly programmed
  - learn from experience E, with respect to task T, and performance measure P—and P improves with E
- spam filtering without ML
  - study problem
  - write explicit rules
  - many rules, complex rules → hard to understand
  - hard to maintain
- spam filtering with ML
  - more concise program
  - easier to maintain program (updates itself!)
  - more accurate
- ML may be only method of solving some kinds of problems
  - speech recognition?

- speech recognition may reveal previously unseen patterns in data (e.g., in spam filters, data mining)
- choose ML when...
  - complexity and number of rules is very great
  - other approaches yield no results or poor results
  - characteristics of the problem are changing rapidly (e.g., new kinds of spam every day)
  - want understanding of large datasets
- ML can help people learn—by revealing patterns in data (e.g., what kinds of messages are spam)
- applications of ML
  - image classification (recognizing defective products) (CNN)
  - detecting tumors (CNN)
  - identifying subjects of news articles (NLP)
  - excluding rude comments from online fora (NLP)
  - summarizing long articles (NLP)
  - chatbots, personal assistants (automatically answer questions)
  - forecasting sales, revenues, admissions, etc.
  - voice recognition (respond to commands)
  - spot likely credit card fraud (anomaly detection)
  - recommend products to subscribers/clients/consumers
  - game playing bots
- kinds of ML systems
  - trained with or without human supervision
  - learn incrementally "on the fly" or not
  - compare new data points to know data points or build a predictive model by detecting patterns in training data
- supervised vs. unsupervised learning
  - supervised learning
    - k-Nearest Neighbors
    - Linear Regression
    - Logistic Regression
    - Support Vector Machines (SVMs)
    - Decision Trees and Random Forests

    - Neural networks
  - unsupervised learning
    - clustering
      - K-Means
      - DBSCAN
      - Hierarchical Cluster Analysis (HCA)
    - anomaly detection and novelty detection
      - one-class SVM
      - isolation forest
    - 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
  - semisupervised learning
    - some data labeled, some not (cost of labeling all data high!)
    - combination of supervised and unsupervised algorithms
    - example: Google Photos—unsupervised learning finds groups of photos (same person pictured in a group), then label one person in each group
  - reinforcement learning
    - agent (the learning system) rewarded for some actions, penalized for other actions
    - learns best strategy (policy)—maximizes rewards
    - examples: robots learning to walk, learning to play Go (computer analyzed many games, played against itself)

- batch and online learning
  - batch learning
    - when system cannot learn incrementally
    - train using all available data
    - amount of time (and other resources required) calls for off-line training
    - deploy only after learning is complete
    - no learning during operation
    - to retrain, need new and old data

    - retraining can be automated
    - might not be practical when computing resources are limited (smart phones, Mars rover)
  - online learning
    - appropriate when data arrives continuously
    - appropriate when computing resources are not great
    - appropriate when whole data set will not fit in memory (out-of-core learning)
    - more accurately called incremental learning (learning may take place off-line)
    - feed system data one at a time or in mini-batches
    - has a learning rate—if too high system might be too sensitive to noise (quickly forgets old data), if too low might have too much inertia (learns/adapts slowly)
    - risk of loss of performance with bad (misleading) data
    - may need to monitor performance (possibly with anomaly detection), turn off learning

- instance-based vs. model-based learning
  - most ML about prediction
  - prediction → generalize from experience
  - good fit to training set insufficient
  - 2 main approaches to generalization: instance- and model-based
  - instance-based learning
    - measure of similarity
  - model-based learning
    - noisy data is data that is partly random
    - example: are wealthier people happier people?

- main challenges of ML
  - insufficient quantity of training data
    - machines need more examples than people need!
    - unreasonable effectiveness of data—with enough data, all algorithms give similar results → quantity of data is more important than choice of algorithm
  - non-representative training data
  - poor quality data
    - outliers
    - missing data

        * noise
    - irrelevant features
    - overfitting the training data (e.g., modeling the system with degree n polynomial, where n is too large)
        * use model with fewer parameters (a way of simplifying model)
        * use fewer features (another way of simplifying model)
        * gather more training data
        * reduce noise (eliminate outliers, correct errors)
    - underfitting the training data
        * use more parameters (more powerful model)
        * use better features (feature engineering)
        * reduce constraints on model (reduce regularization hyperparameter)

- testing and validating
    - probably best not to test in production!
    - divide available data into training set and test set

- NFL → No Free Lunch theorem tells us, absent assumptions, no reason to prefer one algorithm over another
    - cannot test all algorithms with our problem!
    - must make some assumptions!

11 Glossary

11.1 Words

anomaly detection e.g., unusual credit card transactions, manufacturing defects, outliers in dataset

association rule learning discover correlations (interesting relations among attributes)

attribute a data type (e.g., "mileage") (sometimes used interchangably with feature)

clustering algorithm automatically (unsupervised learning) identify groups of related data

cross-validation selecting model with use of multiple validation sets

dimensionality reduction simplify data without losing too much info

feature attribute plus value (e.g., "mileage = 15,000") (several meanings, sometimes used interchangably with attribute)

feature engineering selecting the features for training

feature extraction can be a result of dimensionality reduction (when strong correlation between features is found, features are merged)—reduces time and space required to solve problem

feature selection a step in feature engineering

generalization error error rate on test data

hierarchical clustering divide groups into smaller sub-groups

labels desired solutions in a training set for supervised learning

logistic regression can be used for classification

novelty detection find data unlike any in training set (requires very clean data set)

overfitting over-generalizing

predictors a set of features used to predict a target numerical value

regression prediction of a target numerical value, given predictors (features)

regularization constraining and simplifying model to reduce risk of overfitting

sampling bias a result of flawed method of data collection

sampling noise a consequence of randomness

validation set part of training set, used in selection of model

visualization (unsupervised algorithms) produce 2D or 3D images to allow easier understanding

11.2 Initialisms

CNN Convolutional Neural Network

DBM Deep Belief Network

HCA Hierarchical Cluster Analysis

LLE Locally Linear Embedding

NLP Natural Language Processing

NLU Natural Language Understanding

PCA Principal Component Analysis

RBM restricted Boltzmann machine

RL Reinforcement Learning

RNN Recurrent Neural Network

SVM Support Vector Machines

t-SNE t-Distributed Stochastic Neighbor Embedding