Popular Machine Learning Methods: Idea, Practice and Math

Part 1: Introduction to Machine Learning

Yuxiao Huang

Data Science, Columbian College of Arts & Sciences
George Washington University

Fall 2024

Reference

- This set of slices was largely built on the following 7 wonderful books and a wide range of fabulous papers:

HML	Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd Edition)
PML	Python Machine Learning (3rd Edition)
ESL	The Elements of Statistical Learning (2nd Edition)
PRML	Pattern Recognition and Machine Learning
NND	Neural Network Design (2nd Edition)
LFD	Learning From Data
RL	Reinforcement Learning: An Introduction (2nd Edition)

- For most materials covered in the slides, we will specify their corresponding books and papers for further reference.

Google Colaboratory Instruction

- See Google Colaboratory instruction in github repository: /instructions/google_colab_instruction

Table of Contents

Learning Objectives: Expectation

- It is expected to understand
  - the definition of machine learning
  - the definition of feature and target
  - the difference between:
    - Regression
    - Classification
  - the difference between:
    - Supervised Learning
    - Unsupervised Learning
    - Semisupervised Learning
    - Reinforcement Learning
  - the difference between:
    - Shallow Learning
    - Deep Learning
  - how to use Google Colaboratory

Learning Objectives: Recommendation

- It is recommended to understand
  - the difference between:
    - Batch Learning
    - Online Learning
  - the difference between:
    - Instance-based Learning
    - Model-based Learning

What is Machine Learning?

- "*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 with experience E*" [Mitchell et al., 1997].
- Take Machine Translation for example:
  - E: parallel text translation data between two languages
  - T: machine translation between the two languages
  - P: translation accuracy (e.g., Bilingual Evaluation Understudy, a.k.a., BLEU, see /p3_c2_s4_recurrent_neural_networks)

Data Preprocessing

- The experience E where we learn machine learning systems is the data.
- However, it is essential to perform data preprocessing before feeding the data to the systems.
- This is mainly due to the fact that, without data preprocessing the systems either do not work at all or, at least, do not work well.
- Anecdotally machine learning scientists sometimes spend 80% of their time on data preprocessing.
- We will discuss data preprocessing in:
  - /p2_c1_data_preprocessing
  - /p3_c1_data_preprocessing

Types of Machine Learning

- There are different criterion to divide machine learning into different categories:
  - if the criteria is whether learning needs human supervision, then we have:
    - Supervised Learning: needs human supervision
    - Unsupervised Learning: does not need human supervision
    - Semisupervised Learning: needs human supervision
    - Reinforcement Learning: needs human supervision
  - if the criteria is whether learning can be done incrementally on the fly, then we have:
    - Batch Learning: cannot learn incrementally
    - Online Learning: can learn incrementally
  - if the criteria is whether learning needs training a model, then we have:
    - Instance-based Learning: does not need training a model
    - Model-based Learning: needs training a model
  - if the criteria is whether learning needs training Deep Neural Networks (DNNs), then we have:
    - Shallow Learning: does not need training DNNs
    - Deep Learning: needs training DNNs
- In this set of slides, we will focus on the first and last criterion (human supervision and DNNs) and their corresponding categories.
- See details of the second and third criterion in HML: Chap 1.

Feature and Target

- Before discussing the first criteria (human supervision), let us introduce two kinds of variables in the data, which are closely related to this criteria.
- The first kind of variable is called the *Target*:
  - this is the variable whose value is the most interesting to us
  - we may have any number of target in the data (including zero but usually just one)
- The second kind of variable is called the *Feature*:
  - any variable that is not a target is a feature
  - we may have any number of feature in the data (including zero but usually many)
- See two examples of feature and target on pages 11 and 12.

Kaggle Competition: Predicting House Price

Figure 1: Kaggle competition: predicting house price. Picture courtesy of Kaggle.

- House Prices (Advanced Regression Techniques) dataset:
  - features: 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa
  - target: the sale price of homes

Table 1: The first 7 features and target (SalePrice) of House Prices dataset.

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	SalePrice
1	60	RL	65.0	8450	Pave	NaN	208500
2	20	RL	80.0	9600	Pave	NaN	181500
3	60	RL	68.0	11250	Pave	NaN	223500
4	70	RL	60.0	9550	Pave	NaN	140000
5	60	RL	84.0	14260	Pave	NaN	250000

Kaggle Competition: Predicting Breast Cancer

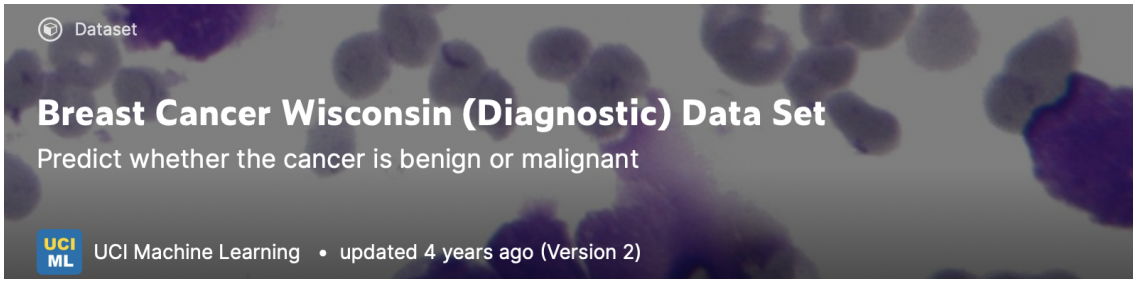

Figure 2: Kaggle competition: predicting breast cancer. Picture courtesy of Kaggle.

- Breast Cancer Wisconsin (Diagnostic) dataset:
  - features: ID number + 30 variables computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, describing characteristics of the cell nuclei present in the image
  - target: the diagnosis of breast cancer, Benign (B) or Malignant (M)

Table 2: The first 5 features and target (diagnosis) of Breast Cancer Wisconsin dataset.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
0	842302	M	17.99	10.38	122.80	1001.0
1	842517	M	20.57	17.77	132.90	1326.0
2	84300903	M	19.69	21.25	130.00	1203.0
3	84348301	M	11.42	20.38	77.58	386.1
4	84358402	M	20.29	14.34	135.10	1297.0

Supervised Learning

- We can perform *Supervised Learning* when:
  - there is target in the data
  - and for most data the target has values
- That is, the two kaggle competitions on pages 11 and 12 belong to supervised learning.
- For predicting house price (page 11):
  - since the target, SalePrice, can take infinite number of values, it is a *Continuous* variable
  - we call this kind of prediction (where the target is continuous) *Regression*
- For predicting breast cancer (page 12):
  - since the target, diagnosis, can take finite number of values (two in this case, benign and malignant), it is a *Discrete* variable
  - we call this kind of prediction (where the target is discrete) *Classification*
- We will discuss supervised learning in:
  - /p2_c2_supervised_learning
  - /p3_c2_supervised_learning

Unsupervised Learning

- We can perform *Unsupervised Learning* when:
  - there is no target in the data
  - or there is target in the data but for most data the target has no values
- Here are some of the most important areas in unsupervised learning:
  - Clustering: dividing the data into clusters
  - Anomaly Detection or Novelty Detection: detecting the outlier that is different from most data (as the case for anomaly detection) or different from all data (as the case for novelty detection)
  - Dimensionality Reduction: transforming high-dimensional data into low-dimensional data
  - Association Rule Learning: detecting the correlation between features
- We will discuss unsupervised learning in:
  - /p2_c3_unsupervised_learning
  - /p3_c3_unsupervised_learning

Semisupervised Learning

- We can perform semisupervised learning when:
  - there is target in the data
  - and the size of data where the target has values is:
    - too small to be used for supervised learning
    - too large to be ignored for unsupervised learning
- While semisupervised learning originated from shallow learning, it becomes more and more popular in deep learning.
- We will discuss semisupervised learning in /p3_c3_s1_deep_generative_models.

Reinforcement Learning

- *Reinforcement Learning*, which is another type of machine learning, is quite different from the ones mentioned earlier (supervised / unsupervised / semisupervised learning).
- In reinforcement learning, an *Agent* (the model) learns the *Policy* (an action) in a *State* (a snapshot of the environment), so that the agent can maximize the *Reward*.
- Example: A mouse in a maze (e.g., matrix)
  - agent: the mice
  - policy: move $\leftarrow$, $\rightarrow$, $\uparrow$, $\downarrow$
  - state: an entry in the maze
  - reward: escape from the maze
- We will discuss reinforcement learning in:
  - /p2_c4_reinforcement_learning
  - /p3_c4_reinforcement_learning

Shallow Learning

- We usually apply shallow learning to non-perceptual data, where perceptual data often refer to:
  - image
  - text
- For non-perceptual data, we usually use shallow models, that is, models other than deep neural networks.
- We will discuss shallow learning in /p2_shallow_learning, including:
  - shallow data preprocessing
  - shallow supervised learning
  - shallow unsupervised learning
  - shallow reinforcement learning

Deep Learning

- Unlike shallow learning, we usually apply deep learning to perceptual data (e.g., image or text).
- For perceptual data, we usually use deep neural networks.
- We will discuss deep learning in /p3_deep_learning, including:
  - deep data preprocessing
  - deep supervised learning
  - deep unsupervised learning
  - deep reinforcement learning

Jupyter Notebook

- *Jupyter Notebook* (notebook hereafter) is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
- We will be using notebook to write python code exclusively throughout.
- We recommend to install notebook via Anacoda:
  - go to: https://docs.anaconda.com/anaconda/install/ (or google install anaconda when broken link)
  - go to the "System requirements" section, choose your operating system and follow the corresponding instructions

Google Colaboratory

- *Google Colaboratory* (Colab hereafter) is an executable document that allows you to write, run and share python code on your Google Drive, with (up to 12 hours) free access to GPU and TPU.
- We will be using Colab to run notebook exclusively throughout.
- See Colab instruction in github repository: /instructions/google_colab_instruction

Bibliography I

Mitchell, T. M. et al. (1997).

Machine learning. 1997.

Burr Ridge, IL: McGraw Hill, 45(37):870–877.