Lectures on machine learning I

Yu Zhang

IHEP, Beijing

March 30, 2020

Several words before start

- Xiaohu gave a set ot lectures about statistic several years ago (Link) before he left. Now it is my turn to present ML!
- Machine learning is not a one-day lesson.
- This tutorial could be weekly tutorial before I leave for a new position.
- This set to tutorial will include algorithm, code of implementation and application in various area.

Outline

1. What is machine learning
2. Typical method of machine learning
   - Top 10 algorithms in data mining and TMVA
   - Deep learning
3. Examples
   - kNN
   - Decision Tree
   - Neural Network
4. Tools and packages
   - Python package and platform
5. Applications
6. Remarks

What is machine learning(twiki)

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

The type of machine learning:

- Supervised learning
- Un-supervised learnning
- Re-inforcement learning

How ML is used in HEP

- Classification:
  - Particle identification
  - Flavor tagging
  - Event classification
- Regression:
  - Energy calibration
  - Reconstruction : pattern recognition

Typical method of machine learning

- Top 10 algorithms in data mining(Link)

- C4.5
- K-means
- Suppport Vector Machine(SVM)
- Apriori
- EM

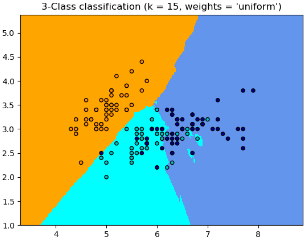
- PageRank
- AdaBoost
- kNN
- Naive Bayes
- Classfication and Regression Tree(CART)

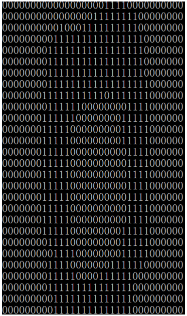
- What is in TMVA?
  - Cut, Likelihood, kNN, Linear Discrminant
  - FDA(Function Discriminant Analysis)
  - Neural Network, SVM, BDT
- Deep learning : RNN, CNN etc

k-nearest neighboring

Algorithm

- Given the labeled dataset.
- Calculte the distance between the test point and the labeled event.
- Select the k-nearest event.
- Calculate the frequent label and it is assigned to the test point.
- Parameters : **K**



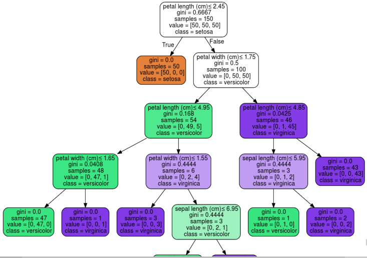


```
00000000000000000011110000000000
00000000000000001111111100000000
00000000001000111111111100000000
00000000011111111111111110000000
00000000111111111111111110000000
00000000111111111111111110000000
00000000111111111111111110000000
00000000111111111111111111000000
00000001111111111101111111000000
00000000111111000000001111000000
00000001111110000000011111000000
00000001111100000000011111000000
00000001111100000000011111000000
00000001111100000000001111000000
00000001111100000000001111000000
00000001111100000000001111000000
00000001111100000000001111000000
00000001111100000000001111000000
00000001111100000000001111000000
00000001111100000000001111000000
00000001111100000000011111000000
00000000111100000000011111000000
00000000011110000000011111000000
00000000111100000001111110000000
00000000111110000111111000000000
00000000111111111111111000000000
00000000011111111111111000000000
00000000011111111111111000000000
```

Decision Tree

Algorithm

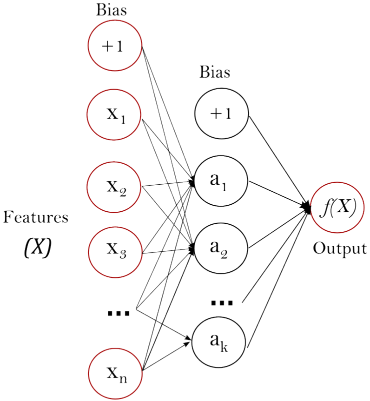
- Give a dataset with $\vec{\mathbf{X}}$
- From each node, seperate the sample to two parts by cutting on the variables, to maximize the significance(entropy, gini-index)
- ending node :
  - Number(Fraction) of events is smaller enough
  - The improvement of significance is smaller enough
  - Maximum number of nodes or depth
- Check the over-training
- Parameters : type of figure of merit, criteria of ending



Neural Network—Multi-Layer Perceptron

Algorithm

- Give a dataset with character $\vec{\mathbf{X}}$(n-dimentional vector)
- Hidden layer : $\vec{\mathbf{a}} = f(\mathbf{W}_1\vec{\mathbf{X}})$, $f(x)$ is propagation function
- Output : $y = \mathbf{W}_2\vec{\mathbf{a}}$
- Parameters : Weight Matrix, number of hidden layers, f(x)



Tools and packages

- Python packages
  - Scikit-learn(Link), Keras(Link), Theano
- Platform
  - Tensorflow(Google)(Link)
  - PyTorch(Facebook)(Link)
- Example codes from Wisconsin
  - DNN : $https : //github.com/laserkaplan/ttHyyML$
  - XGBoost : $https : //gitlab.cern.ch/wisc_atlas/ttHyyML$

Applications

- Speech recognition
- Handwriting recognition
- Natural Language Processing(NLP)
- Computer vision : face recognition, medical diagnosis
- Online advertising : I also know HEP PhD is doing electronic business

Remarks

- How to learn machine learning
  - Do you want to be a “man of parameter-tunning”

- Want to go to industry?
  - I talked to a HEP PhD from USTC working in HuaWei and he already recruited two of my undergraduate classmates who are also HEP PhD.
  - You need to be familiar with:
    - the detail of tranditional ML algorithms
    - related python packages
    - deep learning
    - one muture platform Tensorflow or PyTorch

To be continued

- Let's start with BDT next week.
- It will includes:
  - BDTA, BDTG, BDTB, BDTD, BDTF, XGBoost BDT
  - algorithms, implementation and comparison
- See you next time!