AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 6: AI: Supervised ML and Trust Issues

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

7TH SEP, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 6

- Introduction Segment
  - Recap from Lecture 5
  - Project discussion
- Main Segment
  - Datasets in Weka
  - Method: Random Forest
  - Method: Neural Network
  - Comparing supervised ML methods
- Concluding Segment
  - Quiz 1
  - About next lecture – Lecture 7
  - Ask me anything

Introductory Segment

Recap of Lecture 5

- Project sharing instructions and coding guidelines
  - Reviewed rubric
- Explored UCI datasets
- Explored Weka tool
- Looked at supervised ML methods
  - linear regression
  - Decision trees
  - Noticed difference in explainability of the two methods on a sample problem

Project Discussion – Review Projects

- Information to be shared by students
  - Go to Google sheet: https://docs.google.com/spreadsheets/d/1VAX8ntb5zBQ-vOdsMHMhvEdwoaCZtuBaO4kJdkSA4eQ/edit?usp=sharing
  - Create a Google drive called "CSCE 590-1 Trusted AI (<YourName>)" and share with instructor: firstname.lastname@gmail.com
    - Put shared url in Column E
    - Put project title in column G
    - Create a folder in shared directory call project. Under it, have a Google doc called "Project Description". In it, have the following as bullets with associated details: Problem, User, AI Method, Data, Reliability: Testing, Holding Human Values, Human-AI interaction. See next slide for framework and guidance on what to put.
  - Put Github location for your code in F
    - Create one repository
    - For each quiz, project, etc, create a sub-folder

Course Project – What to Focus On

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
  4. (Data) Explore the data for a solution to work
  5. (Reliability:Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Rubric for Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

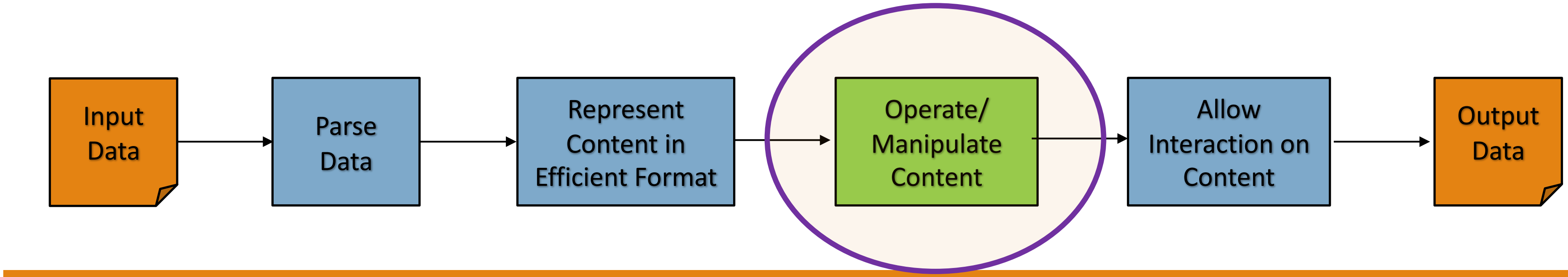
- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

Class Timing for Invited Guest

Diptikalan Saha, IBM Research
https://researcher.watson.ibm.com/researcher/view.php?person=in-diptsaha

Sep 21 (Tu)	**Invited Guest – AI - Supervised ML: External Talk/ IBM's Approach**	Move to earlier time - 11:00 am EST/ 8:30 pm IST
Sep 23 (Th)	**Invited Guest – AI - Supervised ML: External Talk/ Working Session**	Move to earlier time - 11:00 am EST/ 8:30 pm IST

Main Segment


Input Data
Parse Data
Represent Content in Efficient Format
Operate/ Manipulate Content
Allow Interaction on Content
Output Data

Datasets - Weka

- Dataset:
  - Weka: https://www.ics.uci.edu/~mlearn/MLRepository.html (e.g., download: https://prdownloads.sourceforge.net/weka/uci-20070111.tar.gz)

- German credit
  - In Weka dataset (dump above)
  - Direct link: https://github.com/Waikato/weka-3.8/blob/master/wekadocs/data/credit-g.arff
  - As part of development packages
    - like DataHub, https://datahub.io/machine-learning/credit-g#python

German Credit Data

https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

- Dataset that classifies people's credit risk based on their individual attributes such as Age, Income, Gender, etc.
  1000 rows of data, each with 20 attributes to check bias against
- Each entry represents an individual who takes credit from a bank
- Each entry is classified as *Good* or *Bad* credit risk based on their profile
  - It is **worse** to class a **customer as good when they are bad**, than it is to class a **customer as bad when they are good**.

1. Credit amount (numerical); 2. Credit duration (numerical); 3.Credit purpose (categorical); 4. Status of existing checking account(categorical); 5. Status of savings accounts and bonds (categori-cal); 6. Number of existing credits (numerical); 7. Credit history(categorical); 8. Installment plans (categorical); 9. Installment rate(numerical); 10. Property (categorical); 11. Residence (categorical);12. Period of present residency (numerical); 13. Telephone (binary);14. Employment (categorical); 15. Employment length (categorical);16. Personal status and gender (categorical); 17. Age (numerical);18. Foreign worker (binary); 19. Dependents (numerical); 20. Otherdebtors (categorical); 21. Credit score (binary)

Example Instance:
A11 6 A34 A43 1169 A65 A75 4 A93 A101 4 A121 67 A143 A152 2 A173 1 A192 A201 1

A11: less than 0 balance; 6: six months; A34: critical account; A43: has radio/ television ..

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science

VERMA, S., AND RUBIN, J. 2018. FAIRNESS DEFINITIONS EXPLAINED. IN PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS, FAIRWARE '18, 1–7. NEW YORK, NY, USA: ASSOCIATION FOR COMPUTING MACHINERY, HTTPS://WWW.ECE.UBC.CA/~MJULIA/PUBLICATIONS/FAIRNESS_DEFINITIONS_EXPLAINED_2018.PDF

Classification in German Credit

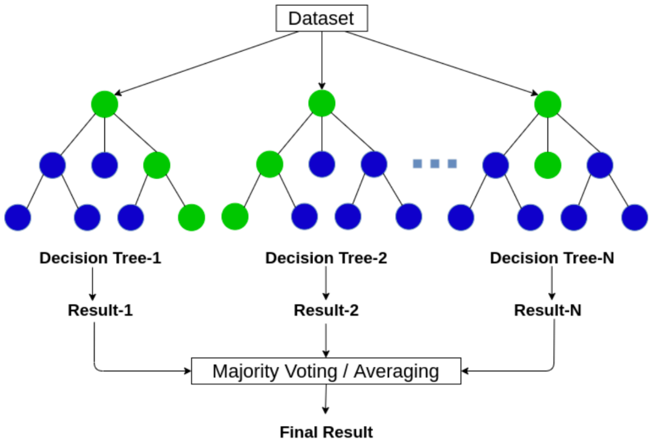
- Demonstration with Weka
  - Simple Logistic Classifier
  - Decision Tree

- We will use 2 other methods on the same dataset soon

Figure Credit:
https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Random Forest

- An ensemble method



- Credits
  - Ideas introduced by Tin Kam Ho in 1995, https://en.wikipedia.org/wiki/Tin_Kam_Ho
  - Matured by Leo Breiman and Adele Cutler at Berkeley (https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro)
  - History: Khaled Fawagreh, Mohamed Medhat Gaber & Eyad Elyan (2014) Random forests: from early developments to recent advancements, Systems Science & Control Engineering, 2:1, 602-609, DOI: 10.1080/21642583.2014.956265
  - Blog: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Slide Courtesy: Leo Breiman and Adele Cutler website

Random Forest

- Main steps (Input: data, N= number of trees)
  - If the number of cases in the training set is N, sample N cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree.
  - If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
  - Each tree is grown to the largest extent possible. There is no pruning.
- Choice of m is implementation dependent; affects correlation between trees and their accuracy

- Characteristics:
  - Fast
  - Accurate
  - Unexplainable

Slide Courtesy: Leo Breiman and Adele Cutler website

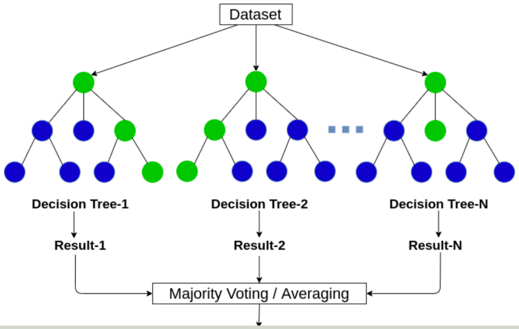
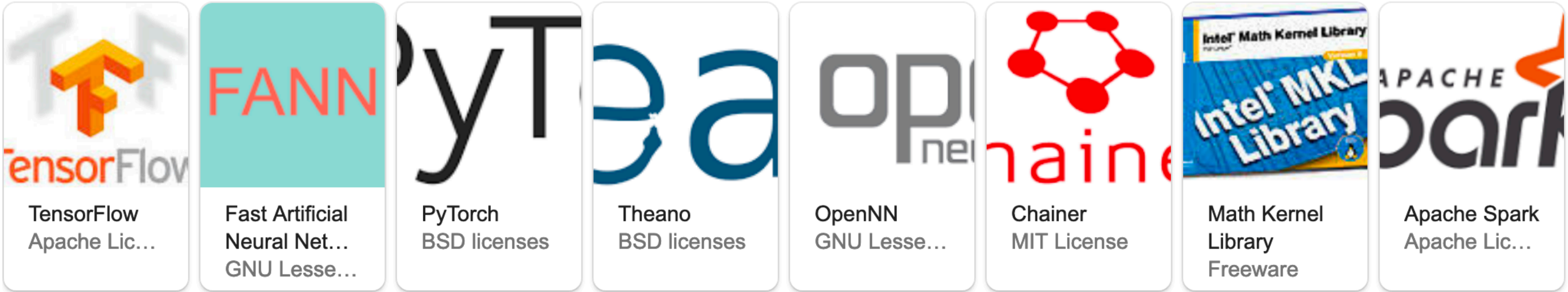


Figure Credit:
https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Neural Network Methods



TensorFlow
Apache Lic…

Fast Artificial Neural Net…
GNU Lesse…

PyTorch
BSD licenses

Theano
BSD licenses

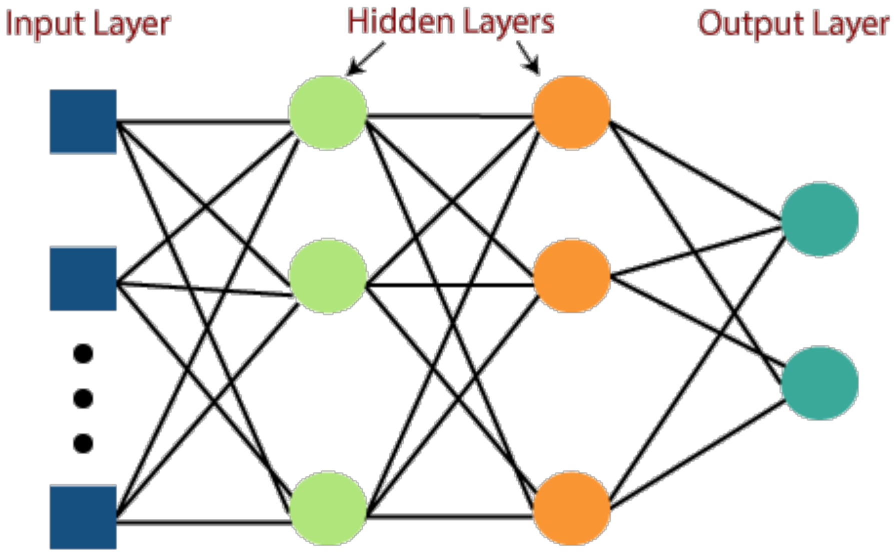
OpenNN
GNU Lesse…

Chainer
MIT License

Math Kernel Library
Freeware

Apache Spark
Apache Lic…

NN – Multi Layer Perceptron



Content and Image Courtesy:
https://github.com/Thanasis1101/MLP-from-scratch

Logistic Regression in a Slide

Function estimate (linear)
W: weight, b: bias

$$f(X_j) = X_j W + b$$

Update Weight

$$W^* = W - \eta \frac{dL}{dW}$$

Error Term (mean squared error)

$$MSE = \frac{1}{n} \sum_{j=1}^{n} \left[f(X_{j\cdot}) - y_j \right]^2$$

Common Code Pattern

```
y = tf.matmul(x, W) + b
loss = tf.reduce_mean(tf.square(y - y_label))
```

NN with Keras and TensorFlow

- By Example:
  - https://github.com/biplav-s/course-nl/blob/master/l9-ml-review/Basic%20TensorFlow%20and%20Keras.ipynb
- TensorFlow's NMIST tutorial
  - https://www.tensorflow.org/tutorials/quickstart/beginner

Classification in German Credit

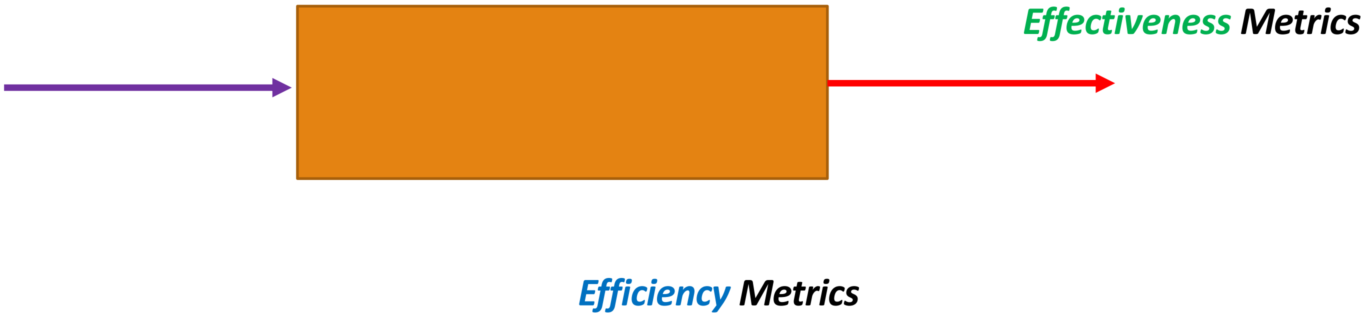
- Demonstration with Weka
  - Random Forest Classifier
  - Multi Layer Perceptron Classifier

- Read about more classifiers at:
  - https://machinelearningmastery.com/imbalanced-classification-of-good-and-bad-credit/
  - https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Comparing and Choosing Supervised ML Methods

Metric Types

- **Effectiveness**: what the **user** of a system sees, primarily cares about
- **Efficiency**: what the **executor** in a system sees, primarily cares about



Criteria for Comparing Classification Methods

Effectiveness

- Goodness of output: predictive accuracy, recall, F1
- Interpretability: providing insight
- Robustness: handling noisy data

Efficiency

- Speed
- Scalability: large volume of data

Source: Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber

Discussion: 10 Tips Paper

- Access: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3
- Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining* **10,** 35 (2017). https://doi.org/10.1186/s13040-017-0155-3

The Tips

- Tip 1: Check and arrange your input dataset properly
- Tip 2: Split your input dataset into three independent subsets (training set, validation set, test set), and use the test set only once you complete training and optimization phases
- Tip 3: Frame your biological problem into the right algorithm category
- Tip 4: Which algorithm should you choose to start? The simplest one!
- Tip 5: Take care of the imbalanced data problem
- Tip 6: Optimize each hyper-parameter
- Tip 7: Minimize overfitting
- Tip 8: Evaluate your algorithm performance with the Matthews correlation coefficient (MCC) or the Precision-Recall curve
- Tip 9: Program your software with open source code and platforms
- Tip 10: Ask for feedback and help to computer science experts, or to collaborative Q&A online communities

Machine Learning – Insights from Data

- Descriptive analysis
  - Describe a past phenomenon
  - **Methods**: classification, clustering, dimensionality reduction, anomaly detection, neural methods
- Predictive analysis
  - Predict about a new situation
  - **Methods**: time-series, neural networks
- Prescriptive analysis
  - What an agent should do
  - **Methods**: simulation, reinforcement learning, reasoning

- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

References

- Insead course
  - Description: https://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions67/ClassificationAnalysisReading.html
  - Data analytics for Business: https://inseaddataanalytics.github.io/INSEADAnalytics/

- Textbooks
  - Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber, https://hanj.cs.illinois.edu/bk3/
  - Introduction to Data Mining (Second Edition), by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, https://www-users.cse.umn.edu/~kumar001/dmbook/index.php

Concluding Segment

Quiz 1

Water Data – Water Atlas

Look at water data from Florida for WaterAtlas project and do four tasks

- WaterAtlas information
  - Website: https://orange.wateratlas.usf.edu/
  - About data collection and group: https://orange.wateratlas.usf.edu/about/
  - APIs to get data - Data download: https://dev.chnep.wateratlas.usf.edu/data-download/beta/
- Local cache with data
  - https://github.com/biplav-s/course-tai/tree/main/sample-code/common-data/water
  - Data for one lake: https://github.com/biplav-s/course-tai/blob/main/sample-code/common-data/water/WaterAtlas-OneLake.csv

Water – Use Cases - Suggestions

- “Water quality status and trends is always a big issue. There are dozens of separate agencies who collect water quality data, different labs – NELAC certified and not, and different formats and repositories of data. In order to determine the status or trend of wq for a waterbody, you have to collect a sufficiently large dataset, but also minimize the presence of bad/questionable data. This is an ongoing challenge for us. In Florida, all local governments and agencies are required to upload data to the WIN (a STORET replacement) database if they want their data used for Impaired/TMDL determinations. We have even found bad data in WIN that were actually collected by FDEP. Figuring out how to minimize “bad” data in the absence of robust data qualifiers might be a nice use-case.

- Along those lines, we implemented a WQ Trends analysis using seasonal Kendall Tau. For Sarasota County (https://www.sarasota.wateratlas.usf.edu/water-quality-trends/), we only use the County monitoring data because they have very rigorous procedures (and Mote does their monitoring and lab analysis). For CHNEP (https://chnep.wateratlas.usf.edu/water-quality-trends/), they include data from numerous sources. We built in filters to remove data based on qa codes, but I always question whether there might be bad data without qa codes. Maybe your students want to develop an AI approach to the “trend analysis” and compare the results to the stats we generate. The data and the R script used for the analysis is available on the web pages.

- People obviously want to know if it safe to go to the beach or swim in the water. Once again, we have several agencies who conduct monitoring for bacteria and red tide. We built the Recreational Water Quality Map to pull together the various monitoring results, but the map itself does not give people an answer – it only provides the data that they then have to interpret. Perhaps the students can use AI to provide a clearer answer. https://www.pinellas.wateratlas.usf.edu/maps/coastal-water-quality-map/”

Lecture 6: Concluding Comments

- We completed discussion of linear methods
- Explored decision trees and random forest
- Looked at water data

About Next Lecture – Lecture 7

Lecture 7: Supervised ML – Trust Issues

- Trust issues with supervised ML
  - Trust Issues – Reading Material - Turing Institute report on AI for Covid in UK
  - Trust Issues Issues – Reading Material – Bias in AI