Instructional Lesson Plan

An Introduction to Machine Learning via Reflective Thinking through the Iris Data Set

Sharon Stefan, Dr. Rosemary Renaut, Dr. Wolfgang Stefan

LESSON DETAILS

Subject Area(s): Mathematics, Computer Science, (MAT 220: Calculus 1)
Focus Grade Level: Community College
Grade Level Range: 12 to 14 (Post High School)

RESEARCH BACKGROUND

As of June 5, 2022 [1], there have been over 529 million confirmed cases of COVID-19 and over 6 million deaths according to the World Health Organization (WHO). Some cases have shown that after contracting COVID-19, lung abnormalities are developed similar to ones present after contracting SARS and MERS. CT scans show parenchymal destruction of lungs expressed as ground-glass opacities and consolidation. [2-5] As WHO declared a global pandemic in March 2020, the number of affected people overwhelmed hospitals and healthcare workers. In particular, digital health care systems could not keep up with the demand for data analysis. Machine Learning (ML) can play a vital role to help radiologists assess properly the severity of the case and ease the burden of cases on physicians. [6]

Kadry, Seifedine, et al. [7] published a comparison of five ML algorithms to classify whether a lung CT was normal or showed signs of COVID infection. The investigation concluded that the choice of feature vectors impacted the accuracy of classification more than the choice of ML algorithm itself. The highest level of accuracy was attained when a tri-level threshold was included in the feature vector.

In [8, 9], Jones describes the importance of including ML algorithms in high school and undergraduate curricula to help increase the number of students in STEM fields. For example, the use of research in medical imaging introduces students the role of ML in health diagnosis. [17-23] This research project serves as a means for students to think about careers in data science. There are few opportunities for students to be exposed to Machine Learning and Artificial Intelligence in a typical mathematics community college curricula. Early exposure to a growing field like data science would benefit groups such as first-generation college students, students of color, non-traditional students, and other marginalized groups.

LESSON SUMMARY

The lesson begins with a pre-assessment for the instructor on students' perception of Learning in general. Students will discuss the differences between Human Learning and Machine Learning. This will lead into the agreed definition of Machine Learning in the classroom.

Based on students' working knowledge of "learning", using website, https://www.whichfaceisreal.com/index.php, students will guess which an image is real or artificially generated. Discussions on how to use features in images to help classify if real or not. Vocabulary will be introduced and we will have discussions on Supervised Learning, Labeled Data, Training and Testing sets.

Using a simple Python code, K-Nearest Neighbor (KNN) is introduced with the focus on the concept of minimizing the distance between a training set to a point in a testing set using 2 features: Petal length and Petal Width.

The next set of activities will use data on Iris Classification [13], a modified lesson activity from [8,9] and Python code from [12]. The data set is larger and uses 4 features as opposed to 2 features in previous activity. The activity will lead into exploring the influence on the choice of K, the number of features selected when K was not defined, justifying the choice of K to improve choice of the classification method. In addition to learning KNN, the activity has students compare accuracy levels using Confusion Matrices between Support Vector Machine (SVM), Bernoulli Naive Bayes, Gaussian Naive Bayes, and KNN with K=5.

MATERIALS AND EQUIPMENT

Computations will be on Jupyter Notebook via Anaconda.

Student all have a Maricopa Gmail account which gives them access to GoogleColab

Canvas Learning Management System (LMS)

Access to reliable internet and a computer

ATTACHMENTS

Handouts:

PreAssessment Lab_1.pdf, Activity 1.pdf, Activity 2.pdf, Post Assessment Activity

Python Code:

Iris_KNN.ipynb, ML_Iris_Oak_Tree_Example.ipynb (Python Code to be run on Anaconda)

Websites:

https://www.whichfaceisreal.com/index.php

https://oak-tree.tech/blog/intro-machine-learning-classification

https://drive.google.com/drive/folders/1pkg0t-qoNvjXCasrKZbhqMia58_FhH3w?usp=sharing. *— Shares the Python code in Google Colab*

EDUCATIONAL STANDARDS

NATIONAL STANDARDS

Next Generation Science Standards (NGSS):
https://www.nextgenscience.org/sites/default/files/HS.ED%204.29%2013With%20Footer.pdf

Grade: High School (9-12)

Discipline: Engineering, Technology, and Applications of Science

1. **HS-ETS1-1** Engineering Design- Analyze a major global challenge to specify qualitative and quantitative criteria and constraints for solutions that account for societal needs and wants.

2. **HS-ETS1-2** Engineering Design- Design a solution to a complex real-world problem by breaking it down into smaller, more manageable problems that can be solved through engineering.

3. **HS-ETS1-3** Engineering Design-Evaluate a solution to a complex real-world problem based on prioritized criteria and trade-offs that account for a range of constraints, including cost, safety, reliability, and aesthetics as well as possible social, cultural, and environmental impacts.

4. **HS-ETS1-4** Engineering Design- Use a computer simulation to model the impact of proposed solutions to a complex real-world problem with numerous criteria and constraints on interactions within and between systems relevant to the problem.

COMMUNITY COLLEGE COURSE COMPETENCIES

AZ Maricopa Community College District Course Competencies: Math 220

1. Read and interpret quantitative information when presented numerically, analytically or graphically. (I, II, III)

2. Compare alternate solution strategies, including technology. (I, II, III)

3. Justify and interpret solutions to application problems. (I, II, III)

4. Communicate process and results in written and verbal formats. (I, II, III)

LEARNING OBJECTIVES

Estrella Mountain Community College Course Level Learning Outcomes (CLO) for Math 220:

1. Students will be able to clearly communicate their work in verbal, written, and visual formats.
2. Students will be able to interpret and draw inferences from information presented graphically, analytically, and verbally.
3. Students will be able to determine appropriate tools/techniques to solve a particular problem.

VOCABULARY

vocab word/phrase (lower case)	*Definition punctuated like a complete sentence even if it's only a phrase.*
Machine learning	Subfield of Artificial Intelligence that focuses on how data is used to imitate the way humans learn.
Artificial Intelligence	The science of making machines perform problem solving tasks or simulation of intelligent behavior in computers
Classification	A type of supervised learning algorithm that recognizes and separates different values or observations into categories. The machine is doing "pattern recognition" as a human would from observations from data.
Model	The set of weights or decision points learned by the machine learning system. Given an unknown example, the algorithm will make the predications of the outcome given input values.
Algorithm	A series of steps to create a model to predict classes from the features as a result of the training examples.
Labeled Data Set	Set of data that is correct in identification and correct in distinguishing between features.
Training Set	The machine going through the learning process using the labeled data. Then it updates on the set of weights and decision points of model by an iteration process until no need for more improvement is achieved. First set in training process.
Validation Set	Another set to improve the model after using the training set. Second set in training process.
Testing	Third set of data. First set will be the iterated training set, then use a second set of training set if machine picks up more information. Once the model is determined to be good enough, try real data or testing data. This will yield a confidence level.

Weights	Each input feature is multiplied by a weight. During training phase, the weights are changed until best model is acquired. The role of the weights could act as a probability for that feature to occur.
K-Nearest Neigbor	Supervised Learning Classification algorithm that uses a set of nearby points to predict outcomes. This is based on minimum distances from K neighboring data points. A consensus is reached by neighbors to classify object.
Supervised Learning	A model learns by examples from labeled or classified attributes of an object. From input values, the machine decides what is the correct label of an object.

LESSON PROCEDURE

INTRODUCTION/MOTIVATION

In order to better understand students' perception of their meaning of Human Learning and Machine Learning, we begin with an individual and small group activity. Start with the pre-assessment handout which asks questions on their interpretations of Learning in general, Human Learning and Machine Learning. Give students 5 minutes to jot down answers to the following four questions:

1. Give examples from your life, where you were/are Learning something? It does not have to be academic, in any way. How do you know you are Learning?
2. When you hear the word "Learning", in the context of how Humans learn, what does this mean? Give examples or a definition.
3. When you hear the phrase "Machine Learning", what does this mean? Give examples or a definition.
4. Do you think there are any commonalities in "Human Learning" and "Machine Learning"? Are there any differences?

This will also help gain insight to how students define learning in general.

Next, in small groups of size 2-3, have students white board their responses as a group to create a working definition of Human Learning and Machine Learning. Bring Classroom together to create a classroom definition of Machine Learning and compare it to the definition most Data Scientists use.

LEARNING ACTIVITIES/STRATEGIES

To introduce the idea of how a machine learns to classify objects, use Activity #1. The process of machine learning will be replicated by students' experience of how they are able to distinguish between a real and artificially generated image.

Give the background information to students of the **Calling Bullshit project**. [12]

Project on screen the following website: https://www.whichfaceisreal.com/index.php. Tell the students that in this activity they will "experience" how a machine learns given some data where you know what the classification is. In this case, the 2 images presented are either a real image or artificially generated. Their job is to choose the real image.

First, have the students complete this activity as individuals. Students are given a total of 9 images in three phases. Give them about 10 seconds for each image and they tally how many in Phase 1 they got correct. Have them think of how they can improve their choices or pay attention to why they got them correct. In Phase 2, show 3 more images with a possible longer time period. Ask if they changed their strategies to improve accuracy. In Phase 3, have them compare to previous stages if they improved and what can they do to improve. In small groups, have students discuss their strategies to improve accuracy in choosing the correct real image. Then as whole group, discuss the importance of feature selection, features that were irrelevant, and the impact of having a larger number of images will improve chances for accuracy. This opens up discussion of following vocabulary: Labeled Data, Training Data, and Testing Data.

Prior to next class meeting, have students read the summary of ideas and vocabulary in Machine Learning in Activity 2 handout.

Activity 2 will introduce the background and history of Iris Data set. Students will be introduced to KNN as a supervised learning model. The mathematics to be explored is how to use the distance formula as an optimization problem for classification of an Iris.

The next part of the work will be done outside of class in GoogleColab. Students will be given code to run, called Iris KNN, within the code are instructions on how to run code. Within the code will be comments explaining the algorithm and what the algorithm is computing. No prior knowledge of coding is necessary.

Students will be assigned into groups of 3-4 to work on Python Code where they will be given a labeled training set of 20 measurements and a labeled testing set of 10 measurements.

Part of this activity is to explore if they feel like they have created an algorithm that has high accuracy. They also explore what affect changes in number of neighbors has on the accuracy of KNN. They will be given measurements of an Iris that are not labeled and asked to classify Iris.

The final activity will introduce how to work with a larger data with more features and compare across ML algorithms for accuracy.

Next, they will be given a Python code called ML Iris Oak Tree that will perform classification of Iris with 4 features. They will explore the importance of feature selection. The Iris Data set contains three classifications with 4 features. Before exploring classification of 4 features, they will compare scatter plots of 2 different pairs of features at a time. In previous activity with Iris KNN code, Petal Length and Width were used in classification. Have the students discuss and compare which pairs of features could give better classification accuracy: Petal Plots or Sepal Plots.

Students will learn how to split a labeled data set into training and testing data via Cross Validation. We will also introduce other Classification algorithms: Support Vector Machines, Bernoulli Naive Bayes, and Gaussian Naive Bayes. Cross validation and confusion matrices are used to interpret how accuracy is visualized with True Positives, True Negatives, False Positives and False Negatives.

CLOSURE

As a closure activity, give the post assessment handout and share personal research experience, if applicable. Share any insight from personal experience.

In my case, share my research experience with ML and health diagnostics. Share my research on the classification of CT scans of COVID patients who may or may not have lung abnormalities using the algorithm they learned in the activities, KNN. This will be done in class as an overall summary of activity.

Post Assessment handout will ask what kind of knowledge did they gain about ML, what they found interesting, did their definition of learning change and would they consider exploring Data Science as part of their STEM discipline. This will be an assignment to be turned in and completed outside of class.

ASSESSMENT

FORMATIVE ASSESSMENT

Describe how you will check for understanding during the process of learning. Include detailed sample items and/or list the name of the actual assessment that you will be attaching.

- *Pre-Assessment Activity*
  - *Listen to student discussion on ideas of learning to help frame whole group discussion. Students will have an opportunity to contribute the discussion from various points of view independent of math background*
  - *Evaluate their understanding of vocabulary by listening to group discussions*
  - *Tie in the whole class definition of Machine Learning and leverage with the common definition of ML that data science community uses.*
- *Activity 1*
  - *Students will be able to compare their strategies and reflect on their learning process mimics the process of how a machine using Training Set to begin the classification process and Testing Set to improve their learning on how to classify objects.*
  - *Evaluate how formal definitions are captured in their experience of being a machine*
- *Activity 2*
  - *Check how students interpret scatter plot data from Iris data; use answers to check how they justify choosing one pair of features over another.*
  - *Check how they interpret Confusion Matrices as a visual representation of model accuracy*

SUMMATIVE ASSESSMENT

Describe the final check for understanding after learning is complete. Include detailed sample items and/or list the name of the actual assessment that you will be attaching.

Use the post assessment handout to have students share their updated version of learning and to describe what was learned overall. This assignment gives the students to reflect overall the lesson and what was learned in the process.

CONTRIBUTORS

INDIVIDUALS

Wolfgang Stefan (Mayo Clinic), Rosemary Renaut (ASU), Jean Larson (ASU), Andreas Spanias (ASU), Jennifer Blain Christen (ASU), Daniel Gulick (ASU), Deep Pugara (AU), Raquel Diaz (Trevor Browne High School), Karl Ernsberger (Lumos Arts Academy), Ebubekir Sen (Sonoran Science Academy), Steven Clemens (Dysart iSchool), Abdullah Mamum (Estrella Mountain Community College), Deanna Alcala (Estrella Mountain Community College), Crystal Herrera (Estrella Mountain Community College), Jon Carlo Santos (Estrella Mountain Community College)

REFERENCES

List citation information for any graphics or copyright material used in the development of this lesson.

1. Novel Coronavirus (2019-nCoV). World Health Organization. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---8-june-2022, Published June 8, 2022. Google Scholar
2. Li, Wei, et al. "Chest computed tomography in children with COVID-19 respiratory infection." Pediatric radiology 50.6 (2020): 796-799.
3. Vijayakumar, Bavithra et al. “CT Lung Abnormalities after COVID-19 at 3 Months and 1 Year after Hospital Discharge.” Radiology vol. 303,2 (2022): 444-454. doi:10.1148/radiol.2021211746
4. Jacobi, Adam, et al. "Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review." Clinical imaging 64 (2020): 35-42.
5. Chung, Michael, et al. "CT imaging features of 2019 novel coronavirus (2019-nCoV)." Radiology (2020).
6. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. Radiographics. 2017;37(2):505-515. doi:10.1148/rg.2017160130
7. Kadry, Seifedine, et al. "Development of a machine-learning system to classify lung CT scan images into normal/COVID-19 class." arXiv preprint arXiv:2004.13122 (2020).
8. Jones, Joshua. "Integrating Machine Learning in Mathematics Classrooms." Mathematics Teacher: Learning and Teaching PK-12 114.8 (2021): 624-628.
9. Jones, Joshua. "Integrating Machine Learning in Secondary Geometry." Mathematics Teacher: Learning and Teaching PK-12 114.4 (2021): 325-329.
10. https://www.kaggle.com/datasets/andrewmvd/covid19-ct-scans
11. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks
12. https://www.oak-tree.tech/blog/ml-models
13. 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.
14. https://oak-tree.tech/blog/intro-machine-learning-classification

15. https://www.whichfaceisreal.com/index.php

16. https://www.callingbullshit.org/

17. Wang, Jing, et al. "Review of Machine Learning in Lung Ultrasound in COVID-19 Pandemic." Journal of Imaging 8.3 (2022): 65.

18. Cai, Wenli, et al. "CT quantification and machine-learning models for assessment of disease severity and prognosis of COVID-19 patients." Academic radiology 27.12 (2020): 1665-1678.

19. Javor, D., et al. "Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography." European journal of radiology 133 (2020): 109402.

20. Mesanovic, Nihad, et al. "Automatic CT image segmentation of the lungs with region growing algorithm." 18th international conference on systems, signals and image processing-IWSSIP. 2011.

21. Zhang, Fengjun. “Application of machine learning in CT images and X-rays of COVID-19 pneumonia.” Medicine vol. 100,36 (2021): e26855. doi:10.1097/MD.0000000000026855

22. Hussain, Lal, et al. "Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection." BioMedical Engineering OnLine 19.1 (2020): 1-18.

23. Dou, Qi, et al. "Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study." NPJ digital medicine 4.1 (2021): 1-11.

SUPPORTING PROGRAM

RET Site: Sensor, Signal and Information Processing Algorithms and Software
Sensor, Signal and Information Processing Center (SenSIP), in partnership with Arizona State University and the National Science Foundation.

FUNDING ACKNOWLEDGEMENTS

This project is funded by the National Science Foundation (NSF) Award 1953745. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the NSF.