Sensor, Signal and Information Processing Center

Instructional Lesson Plan

HOW SUCCESSFUL IS OUR GROUPING OF COLORS?

Ebubekir SEN, David Ramirez, Dr. Andreas Spanias

LESSON DETAILS

Subject Area(s): Computer Science

Focus Grade Level: High School Freshman

Grade Level Range: 9-12

RESEARCH BACKGROUND

Supervised and Unsupervised Machine Learning

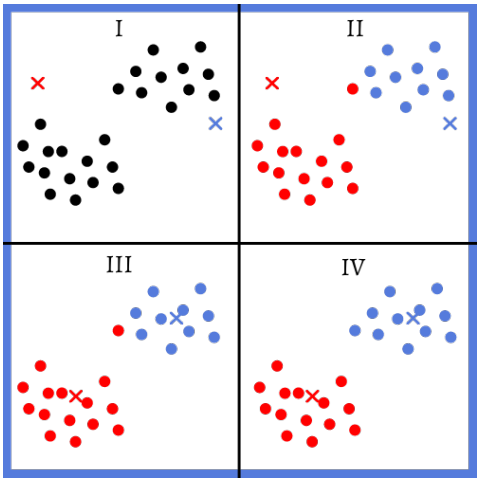
Supervised Learning: Computers are trained with images and their corresponding labels, and computers are tested with a new image they have not seen before to predict the label of the test image.

Unsupervised Machine Learning: Machines are not trained with labeled data sets. Machines try to find patterns by looking at the given dataset by using some algorithms.

Most common Unsupervised Learning algorithm: K-Means Algorithm works

In this type of approach, our model will try to find natural clusters (groups) in uncategorized data. If similarities are found, we'll have different clusters grouping related input samples. In the K-means algorithm, we first need to define the number of clusters K that we'll have. This can be done arbitrarily or using well-established methods. The main idea of this approach is that we:

- randomly initialize centroids for each cluster
- then assign each input sample to the closest centroid
- we move all the centroids to the mean of the input samples that were assigned
- finally, we assign each input sample to the closest centroid again
- We do this iteratively until we reach a defined number of iterations of if the centroids stop changing:



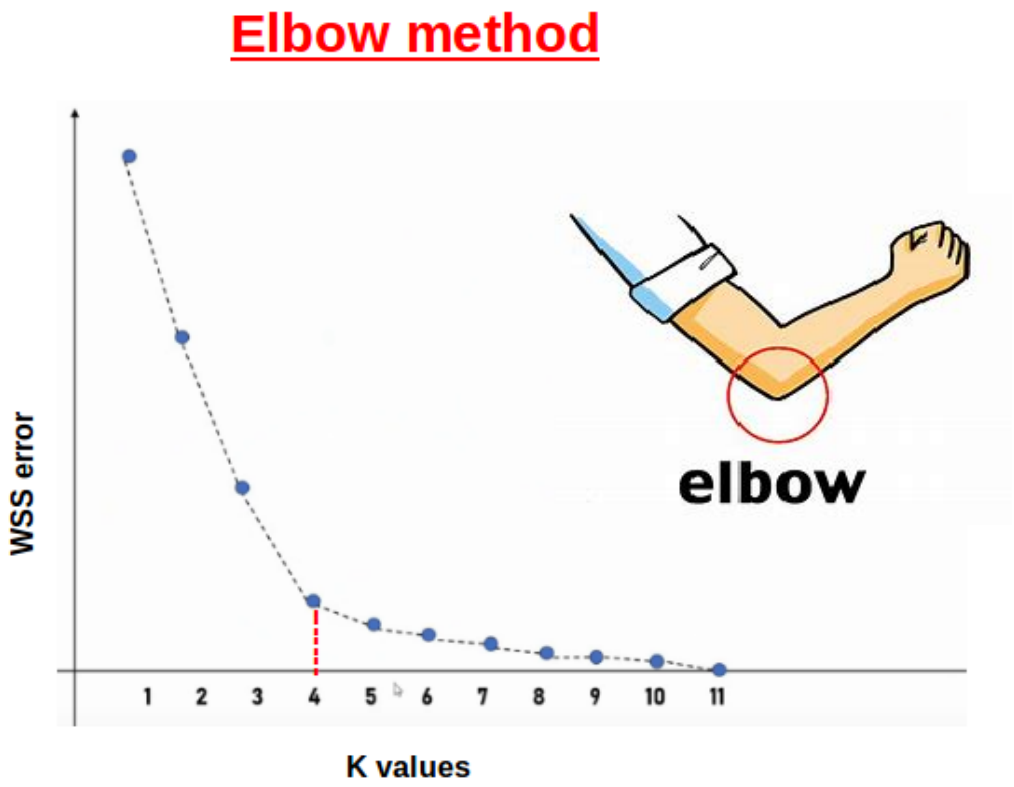
Picture I: Random centroids are assigned to each cluster

Picture II: Running K-means algorithm, first assignment of data points into two different clusters (check colors for data points)

Picture III: Centroids are moved to represent the group little bit better after second iteration

Picture IV: After multiple iterations, each cluster is seen in separate groups with different colors.

Criteria for a successful grouping is called "ELBOW METHOD". Distances of each point to the centroids are calculated and then a cost function is calculated to minimize the distances to the centroids. In the Cost function vs we will be creating, you should be looking for an elbow shape. The corner of the elbow shape will tell you the number of clusters we might have.



LESSON SUMMARY

This lesson is intended for teaching Unsupervised Learning by applying it with an example on grouping pixels. The idea is inspired from Code.org Discovery Computer Science Curriculum Unit 7 Machine Learning and Artificial Intelligence unit Lesson 2. Students are asked to group numbers which are composed of three numbers in a text box. When students check the next step, they realize that number text cells each represent the color of a pixel in which the numbers stand for R(red), G(Green) and B(Blue) respectively. However, there is an ambiguity in the number of the grouping's students should be using. The aim of this lesson is to enable students to decide the optimum number of groups they can create with unlabelled data as part of an unsupervised machine learning. Students are introduced with Google Colab notebook and some of its basic features. They apply their learning by creating code cells and modifying an array. Then students are introduced K means algorithm and evaluate the number of different clusters. Elbow method is introduced to enable students to decide optimal number of clusters in an unknown pattern data. Students research real-life applications and explain real-life applications in their own words. Formative assessments are checked through Google forms and results are displayed on teacher notebook.

MATERIALS AND EQUIPMENT

A computer connected to the internet is needed. If students are using a school provided Chromebook, a communication with the school IT department might be needed to enable Google Colab notebooks.

ATTACHMENTS

Teacher should be using Teacher Colab notebook:

https://colab.research.google.com/drive/1xKjlTY24qdKFDRdXSXaWhoa1H2MpSdc5#scrollTo=TD3ZN9C3zO-R

Students will be using Student Colab Notebook:

https://colab.research.google.com/drive/1Ncoaq7Mlm9Xpfi7yGnAXwBYRQ5V99nx7

EDUCATIONAL STANDARDS

K-12 TEACHERS

Next Generation Science Standards (NGSS):

Science Analyze data using tools, technologies, and/or models (e.g., computational, mathematical) in order to make valid and reliable scientific claims or determine and optimal design solution (Grades 9-12)

LEARNING OBJECTIVES

1. Effective use of Google COLAB by:
   - Creating a code cell
   - Running individual code cell
   - Running all code cells
   - Commenting on a code cell
2. Apply K-Means Algorithm on a clustering example
3. Plot distribution of array points on a graph with different RGB colors
4. Discuss possible number of different color clusters
5. Plot cluster distribution and match centroids with the clusters
6. Plot Cost function and decide the optimal number of clusters
7. Explain real life applications of unsupervised learning in your own words
8. List real life examples of K-means algorithm
9. Explain how unsupervised machine learning works

VOCABULARY

word	*definition*
algorithm	a process or set of rules to be followed.
array	a process or set of rules to be followed.
centroid	the center point of clusters.
cluster	a group.
cost function	a mathematical representation of the squared distance between all the points to their closest cluster center.

library	a collection of pre-written codes for specific tasks.
model	a machine learning model is a file that has been trained to recognize certain types of patterns.
optimal	best or most favorable.
RGB	Red, green, and blue refers to a system for representing the colors to be used on a computer display.
Label	Labels are the values of the target variables (what's being predicted).

LESSON PROCEDURE

Teacher should be using Teacher Colab notebook:

https://colab.research.google.com/drive/1xKjITY24qdKFDRdXSXaWhoa1H2MpSdc5#scrollTo=TD3ZN9C3zO-R

Students will be using Student Colab Notebook:

https://colab.research.google.com/drive/1Ncoaq7Mlm9Xpfi7yGnAXwBYRQ5V99nx7

All the teacher and student instructions as well as illustrative instructions are provided in detail on Colab notebooks.

INTRODUCTION/MOTIVATION

Is there anyone who knows this song? I wish there was an easier way to find out.

Shazam Demo - Does anyone know how this app works?

Remember supervised and unsupervised learning from yesterday.

Let's have a quick refresher.

Sample Scenario: Imagine that we're working for a company that sells clothes and we have data from previous customers: how much they spent, their ages and the day that they bought the product.

Our task is to find a pattern or relationship between the variables in order to provide the company with useful information so they can create marketing strategies, decide on which type of client they should focus on to maximize the profits or which customer segment they can put more effort to expand in the market.

What type of machine learning is this?

LEARNING ACTIVITIES/STRATEGIES

https://colab.research.google.com/drive/1xKjITY24qdKFDRdXSXaWhoa1H2MpSdc5#scrollTo=TD3ZN9C3zO-R

Think-pair-Share will be used to discuss students' reasoning for the number of clusters they create.

Socratic questioning techniques will be used to enhance scaffolding.

CLOSURE

Unsupervised learning occurs without training the computer. There are no labels. Computer tries to find patterns from given data. Most common unsupervised learning algorithm is the K-Means Clustering algorithm. It is widely used in credit card fraud detection, market basket analysis, crime maps and many more.

ASSESSMENT

Students' progresses are monitored throughout the lesson. There are multiple checkpoints which can be found on Teacher/student Colab notebooks. Google forms are used to collect data from students and their responses are shared with the class as instant feedback on the Teacher Colab Notebook.

FORMATIVE ASSESSMENT

Teacher Colab notebook has embedded formative assessments.

SUMMATIVE ASSESSMENT

Students will be searching for real life applications. Teachers will be collecting their responses in their own words with Google Form. A word cloud will be created to share the overall findings with the class.

CONTRIBUTORS

INDIVIDUALS

Teacher: Mr. Ebubekir SEN

Graduate Student Mentor: David Ramirez

Education Advisor: Dr. Jean Larson

Faculty Advisor: Dr. Andreas Spanias

REFERENCES

1. Sample scenario and explanation of K-means clustering is taken from https://www.baeldung.com/cs/examples-supervised-unsupervised-learning
2. Scattering of RGB values idea is taken from Code.org Discoveries Course 2021 Curriculum, Machine Learning U nit Lesson 2.
3. The code for connecting Google Drive to check Formative assessments are taken from https://gitlab.com/foohm71/mylovelysurvey/-/blob/a3795ce2492280dac4d346a1b7e5e51439684812/SurveyResults.ipynb There are some credits on this website for different parts of the code.
4. The code for K-Means Algorithm is taken from Kristen Jaskie's lecture at https://colab.research.google.com/github/kpjaskie/SenSIP21/blob/main/3_ML_Algorithms_Clustering.ipynb#scrollTo=4DxWvUVFsAxp
5. Accessing Google Sheets https://medium.com/mlearning-ai/how-to-access-google-sheets-on-google-colaboratory-8766b3a0996f
6. The Elbow Figure is taken from https://miro.medium.com/max/1340/1*RnvrdhlUxHWss3vOffHT5g.png
7. Sample real life examples can be found in the following link https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-algorithm

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.