SenSIP

Sensor, Signal and Information Processing Center

Instructional Lesson Plan

Instructional Lesson Title	KMeans Clustering with Customer Data		
Subject Area	Computer Science / Data Analysis / Problem Solving / Science and Technology		
Keywords (4-10 words)	Problem-based Learning, machine learning, data science, KMeans algorithm		
Unit Duration (in min.)	350 minutes (seven 50-minute lessons)		
Focus Grade Level	12th	**Grade Level Range**	11 to 12

Instructional Unit Summary

Students explore basic principles of machine learning in solving complex problems. Students develop a process to understand and apply the KMeans learning algorithm. They begin by developing a way to find the centroid of a single cluster of points. Next, they develop a way to find two centroids in the same data set. They then represent their processes with diagrams and mathematical equations. Students then build a working model using Excel (or Google Sheets) to tests their protocol. After developing the KMeans method 'manually' students learn to use the Python KMeans algorithm. The KMeans algorithm instantly finds k numbers of centroids for a data set. Students apply the algorithm to a creative scenario involving real world complex problems.

Engineering Connection	**Engineering Category**
Engineers apply math to solve complex real-world problems. Modern computer systems and hardware produce endless streams of data which needs to be analyzed to better understand a phenomenon. But processing tremendous amounts of data can be overwhelming. Using machine learning to analyze data allows engineers to thoroughly process data and investigate phenomenon at an in-depth level. This enables them to uncover hidden patterns once not available for detection.	Engineering with some science/math

Materials & Equipment

- Data set
- Spreadsheet package (Excel/Google)
- Google Colab and dataset

Attachments

1. Data set
2. Presentation-Teacher version
3. Presentation-Student version
4. Restaurant Data Clusters Colab
5. Where to Build_Tchr
6. Where to Build_Std

Prerequisite Student Knowledge

- Basic algebra skills, such as finding midpoint of two ordered pairs
- Spreadsheet programming skills, such as calculating midpoint of two ordered pairs
- Basic Python programming skills for reading code

Educational Standards

Arizona State Standards

STANDARD 3.0 APPLY MATHEMATICAL LAWS AND PRINCIPLES RELEVANT TO ENGINEERING AND TECHNOLOGY

STANDARD 5.0 APPLY ENGINEERING TECHNOLOGY AND TOOLS

Next Generation Science Standards (NGSS)

HS-ETS1-4. 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.

International Technology and Engineering Educators Association (ITEEA) Standards

P. Use computers and calculators to access, retrieve, organize, process, maintain, interpret, and evaluate data and information in order to communicate. (Grades 9 – 12)

Learning Objectives

Students will be able to:

- Create and represent a procedure to find the mathematical center (centroid) of a cluster
- Create and represent a procedure to mathematically place two centroids in a cluster
- Describe the steps followed in a KMeans clustering algorithm
- Modify existing Python code to create a new cluster of data with new centroids
- Apply KMeans algorithm to a new situation (cellular towers) and make informed recommendations
- Share recommendations in a report

Vocabulary	Definitions
centroid	Middle point of a cluster
algorithm	Process or set of steps to be followed in completing a calculation
kmeans clustering	Computer algorithm for placing k number of centroids in a cluster
iterate	Repeat steps to improve results
iterative process	Repetition of a process

Lesson Procedure

Introduction/Motivation

How to engineers use data? How do they sort large sets of data into useful segments? How do they find patterns in data? Machine Learning is a relatively new tools available to engineers (and others) which enables them to quickly handle and make sense of large amounts of data.

Problem Presentation

A local, but popular restaurant has developed a big customer base across the metro area. As business has grown the restaurant owner has decided to look at building a bigger restaurant in a new location. She decides to use customer address data to help her to decide the best location for the new restaurant. She also wants to consider the idea of opening more than one restaurant in other parts of town in order to bring her restaurants closer to her customers. As a college student working in the restaurant, she has tasked you to use the data to suggest the optimal location or locations in which to build.

Lesson Background/Teacher Concepts

Basic algebra- finding midpoints: Given a small data set (~15 data points) students will find the averages (mean values) of the x and y coordinates to find the center of a cluster.

> To place two centroids in the same cluster students need to 1) place two dots randomly 2) measure distance to each dot 3) pair cluster points into a group with closest centroid dot 4) find center of new group 5) move centroids to new center spots 6) repeat process until no more movement (iterative process)

Spreadsheet skills- teacher will help students build the manual mid-point process on a spreadsheet.

Understanding of KMeans algorithm

Google Colab experience- students will run a simple KMeans algorithm using a Google Colab notebook

Learning Activities/Strategies

Lesson 1- Manual development of simple centroid algorithm

Problem: A local, but popular restaurant has developed a big customer base across the metro area. As business has grown the restaurant owner has decided to look at building a bigger restaurant in a new location. She decides to use customer address data to help her to decide the best location for the new restaurant. She also wants to consider the idea of opening more than one restaurant in other parts of town in order to bring her restaurants closer to her customers. As a college student working in the restaurant, she has tasked you to use the data to suggest the optimal location or locations in which to build.

Teacher Task: Present spreadsheet with single cluster of dots (representing the loyal customer addresses). Have students place dot in the center of the dots.

Student Tasks:

Place an X in the 'center' of the dot. This represents the new restaurant location.

Q: How do you know it's the center?

Q: How would you find the mathematical center?

Task: Create a process to find the 'center'

Task: Create a symbolic diagram (flow chart/equations) to show the process

Task: Build and test the process in an Excel model

Student Products:

- Symbolic model of process (flowchart/equations)
- Excel model

Lesson 2- Expansion of centroid algorithm

Teacher Task: Go to 2nd tab on same spreadsheet. Instruct student to do again using two center points.

Q: How would you do this with two centers? (2 optimal restaurant locations)

Q: How would you mathematically locate the centers?

Student Tasks:

Task: Flow chart of process two find two centers

Task: Build this process on a spreadsheet so the spreadsheet does this efficiently.

Student Products:

- Expanded Symbolic model
- Expanded Excel model

Lesson 3- Introduction of Python KMeans algorithm

Teacher: Increase the complexity by asking students to recreate the centering process for multiple centers. Ask them to consider how they would build a computer program to do this.

Student Tasks:

Q: How would you do this for multiple centers?

Task: Flowchart a potential coding process (using Python)

Teacher Task: Share Colab for KMeans example. Show students how to run the program on Colab.

https://colab.research.google.com/drive/15pZXzQx0l81aP9L3Q4fbTUaZteTjgDm_?usp=sharing

Products:

- Symbolic model of process (flowchart)
- Colab notebook

Lesson 4- Application of Python KMeans algorithm (Assessment)

Task: Use KMeans algorithm to solve a new problem.

A mobile phone company needs to set up new cell towers in a big city. They have maps and data sets of where their existing customers live. Based on the maps showing the cluster of customers they must find a way to put the cell tower right in the center of all their customers. Use the KMeans algorithm to find the optimal locations for new cell towers. Create a report explaining your recommendations for the optimal number of towers to be placed and their locations based on your machine learning analysis. Include relevant data and findings in the report as evidence that justifies your recommendations.

Scaffolding Strategies

Teacher will:

- Monitor and facilitate, making suggestions, as students will develop mathematical procedure to find centroid
- Help students find effective ways to represent their procedure (using flow chart diagram and math symbols)
- Assist students with creating spreadsheet models
- Assist students with understanding and modifying Python KMeans algorithm
- Guide students to come up with a new applicable scenario to apply KMeans algorithms to

Collaboration Strategies

Students will work in small groups (3-4) of their choosing to build procedures for finding the center of a cluster. They will share out their work in shorts presentations to other groups for feedback. The team will then co-develop a spreadsheet model for finding centroids. Lastly, they will work together to use the KMeans algorithm to solve a 'real world' problem and produce a report.

Closure

Machine learning is useful in several contexts. We have explored one technique of machine learning - the KMeans clustering algorithm. This can be useful for sorting data and creating bins or categories of data sets. This type of work is done frequently in digital image compression, voice recognition, population binning and fake news detection.

Assessment

Pre-Assessment

The initial activities of having students manually create algorithms/procedures will serve as a pre-assessment opportunity. Observing student approaches to the problem will provide insight into prior knowledge into math, problem solving and design abilities.

Formative Assessment

Students present their initial procedures and symbolic models to peers in peer review sessions.

Students share progress of Excel models and get teacher feedback.

Teacher meets daily with each group to monitor progress.

Summative Assessment

Students apply KMeans algoritm to a data set of cell phone company customers in order to decide the optimal location for new cell towers. They will also share process and data results in a final report.

Homework

Students are given independent time to work on spreadsheet models. They will also need to research certain spreadsheet functions in order to complete the model.

Students will be given independent time to create final storyline, solution and justification.

Optional Aspects

Lesson Extension Activities

Student are invited to create a copy of Kmeans Colab notebook and modify the code in order to produce other results. This is an exploration without a final objective.

Technology Integration

Spreadsheet software, such as Google sheets or MS Excel

Google Colab website

Contributors

Individuals

Teacher: **Mr. Milton Johnson**

Graduate Student Mentor: Kristen Jaskie

Education Advisor: Dr. Jean Larson

Faculty Advisor: Dr. Andreas Spanias

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.