Machine Learning for Everyone

Villanova University Machine Learning Project

Introduction to Classification and Clustering

Overview

This module introduces two important machine learning approaches: Classification and Clustering. Each approach provides a way to group things together, the key difference being whether or not the groupings to be made are decided ahead of time. While these grouping techniques are a type of Artificial Intelligence designed to be performed by a computer, they can be used on any sort of data from any discipline.

This document provides you with the background needed to use the materials included in this module to teach your students about Classification and Clustering in a way that is relevant to the topic of your course.

- Level: Beginning, Introductory
- Length: One lecture class period
- Prerequisite: Introduction to Machine Learning, Module 1, or none

Contents

- This Introduction
- Lesson Plan
- Hands-On Activity
- PowerPoint Lecture Slides
- Test Your Knowledge Quiz
- Start-up and Spot-check Surveys

Topics

- Machine Learning
- Artificial Intelligence
- Classification
- Clustering
- Supervised Learning
- Unsupervised Learning

Machine Learning

Machine learning (ML) is the ability of a computer to improve its own performance through the use of software that employs artificial intelligence techniques to mimic the ways by which humans seem to learn. It focuses on the development of computer programs that can teach themselves to change and behave in some improved way when exposed to new data.

Machine learning attempts to “learn” by looking for and detecting patterns in the data and adjusting the program’s actions accordingly, similar to how humans try to improve understanding through observation, study and repetition.

Artificial Intelligence

Artificial intelligence (AI) is the intelligence exhibited by machines or software. It is a branch of computer science that deals with helping machines find solutions to complex problems in a more human-like fashion. This generally involves borrowing characteristics from human intelligence and applying them to the design of computer programs. Although primarily thought of as a science, AI is also considered the art of creating machines that perform functions that require intelligence when performed by people.

Introduction to Classification and Clustering

Classification & Clustering

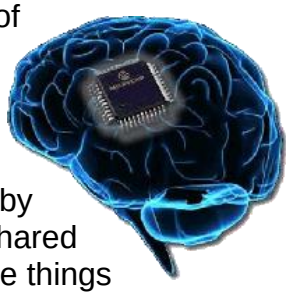
The general idea of Classification and Clustering is to group inputs into (hopefully) distinct categories. They can also then use those categories to identify group membership of new input in the future.

Classification

Classification is a form of machine learning where two or more distinct groups, or classes, of things are known ahead of time and used to group additional things. The features of members of each class are analyzed or learned, and then generalized to build an understanding of what it means to be a member of each class. The generalizations are then compared with potential new member and used to "class-ify" each into the best fitting class. The result is a classification of each of the new members into one of the pre-defined classes.

Classification is an example of **supervised learning** because classes are known.

Clustering



Clustering is a form of machine learning where groups, or classes, are not known ahead of time so groupings are created by looking at similar and shared characteristics among the things being grouped.

Given a set of potential class members, the task of clustering is to establish the existence of two or more classes or clusters in the members. These clusters are often determined through trial and error by grouping things together by looking for similar features, analyzing the results to see how good they are, and repeating this learning process until the groups are deemed acceptable.

Clustering is an example of **unsupervised learning** because classes are unknown at the start.

Lesson Plan

Description

In this lesson, students will learn about Classification and Clustering from a stand-up activity, a brief lecture, a hands-on activity and plenty of discussion.

Preparation

- ✓ Review Module: Introduction to Machine Learning (optional)
- ✓ Read this Introduction sheet
- ✓ Make copies of the Hands-On Activity and the Start-up and Spot-check Surveys as needed
- ✓ Review PowerPoint slides

Part 0: Beginning

Before presenting any information, have the students complete the quick Start-up Survey. Next, tell students they are going to start by conducting an experiment.

Part 1: Stand-up Activity

For this group activity, have all students stand up and instruct them to form themselves into two fairly equal groups any way they like. Groupings can be based on gender, eye color, handedness, birthday month, home state, number of siblings, or any other criteria students devise.

Discussion: While the students are still standing in groups, ask them to explain how they formed these groups, including anything they tried that didn't work. Ask them to make suggestions about other ways they could form groups.

Optional: An optional follow-up activity is to have them form two groups where all members of each group are the most like other members of the same group. Or, have them try forming three groups using the same technique they used before. Or, have them try to form "perfect groups" where the members in each group are most like the other members in their same group. Repeat the discussion of the approaches used, what worked, what didn't and how else they might approach the task.

Part 2: Machine Learning Lecture

Present the first few slides until you get to the first example (Example: Classification). Feel free to add more explanation if you like from the material earlier in this guide or that you find from other resources.

Describe the first example by presenting the idea of Classification, noting that in the exercise the task is to use the pre-defined classes (Group A & Group B) to classify each of the pictures on the following slide into one of those two classes.

Part 3: Hands-On Activity

Provide each student with a copy of the Hands-On Activity.

Suggest that they can use any technique they like for determining how to fit each picture into one of the classifications. They could try re-using ideas from the earlier stand-up activity or come up with new ideas of their own.

When students are done with the Classification task, **discuss** how the classification worked:

- How did you approach the task?
- What made it easy to do?
- What made it hard to do?
- How good are your final categories?

Optional: During the discussion, repeat the classification exercise by projecting the slide containing many pictures of dogs and cats on a whiteboard, drawing circles of one color around each member of Group A and circles of another color around Group B members.

Alternately, if only one colored marker is available, draw circles around Group A members and nothing around Group B.

Next, introduce the clustering example (Example: Clustering) and have students work through the exercise on the Hands-On Activity handout.

Emphasize that the groupings are entirely up to them, although they may NOT repeat a previously used classification. They should come up with a different way to divide the pictures into two distinct groups (Group 1 & Group 2).

When students are done with the Clustering task, **discuss** how the clustering worked:

- How did you approach the task?
- What made it easy to do?
- What made it hard to do?
- How does it compare with Classification?
- How good are your final categories?

Part 4: Explore Further

Finally, introduce the third example (Example: Clustering Again!) and instruct them to repeat the clustering task this time grouping the pictures into three classes.

When students are done, **discuss** how this more difficult clustering worked:

- How did you create categories this time?
- Did it get harder to do with more categories?
- What made it harder?
- How good are your results?

Tips & Ideas

Classification and Clustering can be applied to just about any subject area. Asking these questions may lead you to ways that these machine learning techniques can apply to your discipline:

- Are there any pre-existing groupings of items within your discipline?
- What are the types of data I use or gather in my research area?
- What types of distinctions are made between various items or topics or ideas in my discipline?
- What are areas of disagreement within my discipline and what ideas or items are part of each point of view?

Once you have identified a good potential use of groupings of items or ideas within your discipline, see which way of forming groups is the best fit. Knowing the groups ahead of time indicates Classification, while unknown groups suggest Clustering.

Part 5: Make it Real

Finally, discuss how these approaches might be used within your own discipline or in other areas in general:

- How could classification or clustering be used to help a college determine who they should admit?
- How could a hospital use these approaches to classify patients in the emergency room?
- How could an employer use these approaches to figure out who to hire?
- How could a weather forecaster use these approaches to better predict future weather?
- How could classification or clustering be used by somebody studying or doing research in:
  - Art History
  - Astronomy
  - Biology, Chemistry or Physics
  - Business or Economics
  - Cultural Studies
  - Communication or Psychology
  - Criminology or Sociology
  - Cultural Studies
  - Education
  - English or Literature
  - Environmental Science
  - Foreign Languages
  - Geography or History
  - Gender and Women's Studies
  - Mathematics
  - Philosophy or Theology
  - Political Science

Test Your Knowledge Quiz

At the conclusion of the lesson, have students complete the quiz which includes a repeat of the Start-up Survey questions.

Test Your Knowledge: Key (10 points total)

Circle the letter in front of the single best answer for each question.

1. When a computer program exhibits what appears to be human-like intelligence, it is probably using an approach known as:
   - A Electronic thought
   - B Digital mimicry
   - C Intelligent learning

1 point (D) Artificial intelligence

2. Specific examples of Machine Learning where a computer is able to improve its own performance over time include:
   - A Performing millions of mathematical computations per second and drawing impressive 3D gaming graphics

1 point (B) Filtering out spam email messages and converting handwriting into computer text

   - C Turning on a computer screen saver after a period of inactivity and automatically dimming a cell phone screen in low-light situations
   - D Finding the shortest route home on your GPS device and analyzing paint samples to get a perfect color match

3. Classification and clustering are which of these types of machine learning?

1 point (A) Supervised and unsupervised learning

   - B Categorized and de-categorized learning
   - C Repetitive and experiential learning
   - D Filtered and identified learning

4. Which one of the following everyday situations is the most like Classification?
   - A Forming teams when the captains don't know any of the players
   - B Figuring out where to sit during lunchtime in a high school cafeteria

1 point (C) Deciding whether to pay using cash or credit

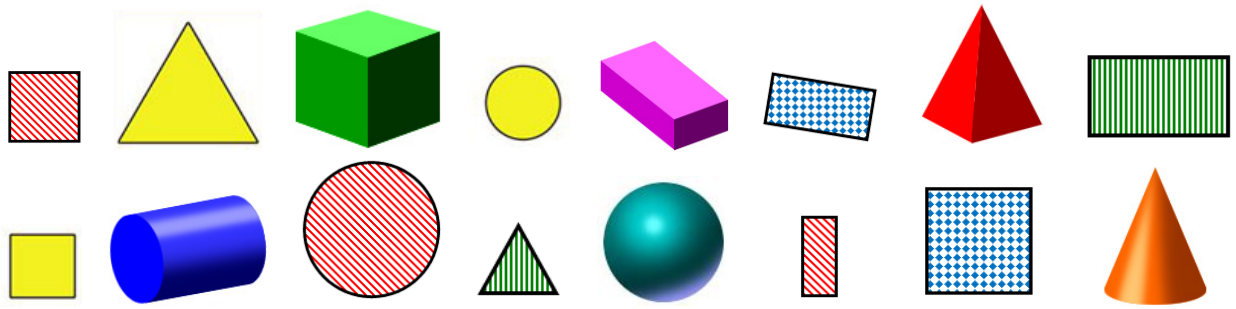
   - D Reorganizing an accidentally dumped-out box of 64 crayons

5. What makes Clustering more difficult than Classification?
   - A Not knowing the class labels ahead of time
   - B Doing lots of comparisons until you finally find the best clusters
   - C Dealing with items that don't seem to fit well into any cluster

1 point (D) All of the above

Answer the following questions.

6. Describe the steps needed to put these shapes into two equal-sized groups?



2 points - presents a clear procedure that involves identifying similar features, using those features to define classes, putting items into groups, and possibly repeating to improve the groupings.

1 point - procedure is on the right track but misses some steps or details or presumes pre-existing classes.

0 points - incorrect answer.

7. What is the difference between the terms Classification and Clustering?

1 points - mentions that classification groups into pre-defined classes while clustering has no pre-defined classes and groups items based on features or characteristics of the items being grouped

0 points - incorrect answer

8. Explain whether Classification or Clustering is harder and why.

2 points - makes a case that classification is harder because fitting items into pre-defined classes might be difficult if there is no “good” fit, or that clustering is harder because you don’t know the classes ahead of time and it is time-consuming and possibly imprecise to create classes from the items on hand.

1 point - attempts to explain which is harder and is generally correct but misses important details.

0 points - incorrect answer