In Class Activity - Machine Learning Revisited

In this activity, we will revisit the 3 steps of machine learning. Recall that the three steps are 1) data collection, 2) data cleaning, and 3) model training.

We will use survey data to build a model that will help an instructor determine if students will be receptive to a non-traditional form of instruction.

We will also see that:

1. A Decision Tree model can help identify the characteristics that best describe the students who do (or do not) want the non-traditional instruction.
2. A Correlation Matrix can help a person quickly spot pairs of correlated variables.
3. Natural groupings that exist in data can be discovered with a Cluster model. Such a model used in a system to identify expected and unexpected groupings of data.

# Step 1 - Data Collection

The survey data is real. It was collected over three semesters here at Temple. The survey was constructed to investigate student attitudes towards flipped-classroom instruction of a STEM topic, post COVID.

The excel file flipped-classroom-survey.xlsx and all the scripts related to this activity are to be found in ica-ml01.zip

Download ica-ml01.zip and unzip the ica-ml01 folder into your mis3536workspace

# Step 2 - Data Cleaning

## Instructions

1. Start by examining flipped-classroom-survey.xlsx in Excel.
2. As we edit the data, we will document our changes in the Notes tab of the xlsx file.
3. Be sure to document your changes as you go! Follow along with your instructor to:
4. Removed unneeded header rows
5. Rename the “Q” columns as i1, i2, i3 … i10 **(this is very important!)**
6. Add a numeric row id column
7. Deleted record 1 - Test record - Status 1
8. Delete other incomplete records.  **(Be sure your row count ends up matching the rest of the class.)**
9. Add an "approve" column - derived from i10 . If i10 >= 4 then approve is 1 , otherwise it is 0
10. Add a "semester" column - derived startdate - semesters are coded as 1, 2 and 3
11. Upload your cleansed xlsx file to today’s activity on Canvas.

|  |
| --- |
| **NOTE:** Technically, we are done with data cleanup here. The data has been cleaned up. But, in the steps that follow we are going to explore the relationships that exist in the data a little bit more using something called a Correlation Matrix.  A person might want to do this kind of exploration before moving straight into Machine Learning.  The Correlation Matrix is a systematic way to compare many variables. This can lend itself to automation as much as it would to a human performing analysis.  Finally, a correlation matrix could be a tool for detecting unwanted bias, ***after*** a model is deployed and in production.  While a correlation matrix isn’t an ML model, it is a sort of a Swiss-army knife technique, useful in lots of situations. |

1. Now run the correlation-matrix.ipynb against the cleansed data sheet.

|  |
| --- |
| **DISCUSS**: What do we see here? What variables are correlated? Be sure to look for positive an negatively correlated variables. Pick one and explain in plain English what the correlation means in terms of the survey questions. |

# Step 3 - Model Training

Now we are going to use a decision tree model to predict a ***categorical*** outcome variable. We want to know which survey question(s) predict a student’s approval of this instructional technique.

## Instructions

1. Using Anaconda Navigator, open the folder and the Jupyter Notebook named decision\_tree.ipynb.
2. Your instructor will review the contents of this script with the class.

**DISCUSS**: Why aren’t we considering question 10 to predict approval?

1. Run this script to generate a decision tree model with a max depth of one.
   1. What survey question is the most powerful predictor of approval?
   2. Take note of the model’s estimated accuracy reported as the “Correct Classification Rate” at the bottom of the script.

|  |  |
| --- | --- |
| *What do all these numbers mean?*  A rectangular object with numbers and symbols  AI-generated content may be incorrect. | *Well…*   * This is node #1. * 17.6% of the data in my training set ended up in node 1. * 2/3 of these “node 1” people did ***not*** approve. They were coded 0. * 1/3 of these “node 1” people did approve. They were coded 1. * That is to say… this is a “class 0” bunch of people… most of them were coded 0 because they did not approve. |

1. Upload decisiontree.pdf (with a max depth of 1) to the corresponding activity for today.
2. Now rerun the script with a max depth of 3, and again with a max depth of 5, and again with a max depth of 7. With each model, take not of the correct classification rate.

|  |
| --- |
| **DISCUSS:** What do we observe about the Correct Classification rate as we increase the depth of the tree? |

Here we have a ***scaling law.***

***Our first scaling law is*** – the quality of the ML model improves with more training data.

***Our second scaling law is*** – the quality of the ML model ***tends to*** improve with increased model complexity. The more complex the model, the greater the accuracy.

***Careful!*** This scaling law is not true all the time. At a certain point of complexity, all models can suffer from **overfitting**.

**Overfitting** happens when a model has high accuracy on training data but low accuracy on new, unseen data. Their predictive power improves with complexity up to a point and then starts to worsen as the model becomes overly complex.

1. Here is a recap of what we have done so far

* We worked through the 3 steps of Machine Learning.
* You practiced cleaning up data with Excel.
* You saw a correlation matrix as a mechanism to quickly assess multiple correlations.
* We revisited our first two scaling laws.
* You learned the term “**overfitting**” and observed overfitting start to take place.
* We learned how to read a decision tree

1. A decision tree is an example of **Supervised Learning**. It is called “supervised” because the possible classifications (“approve” or “disapprove”) are known in advance.

# ANOTHER KIND OF MODEL

1. There is another kind of machine learning called **Unsupervised** **Learning**. This name applies to any learning mechanism that seeks to discover groups or patterns that exist in the data. ***The groups are not known in advance.*** The goal of the Unsupervised Learning model is to discover the groups or patterns!

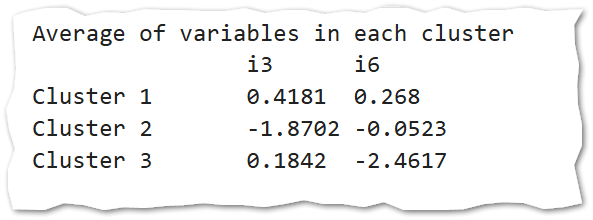
(NOTE: To be clear … the word “Unsupervised” does not imply anything about automation or unattended continuous improvement.)

1. Open the Jupyter Notebook named clustering.ipynb.
2. Your instructor will review the contents of this script with the class.
3. Run this script, generating three clusters.
4. Edit the script, taking the number of variables from 10 down to 4. (You do this by commenting / uncommenting the variable COLUMNS\_FOR\_ANALYSIS.)
5. Rerun the script, generating three clusters.
6. DISCUSS: How do we read the following table?

A number with black text

AI-generated content may be incorrect.

1. Edit the script, taking the number of variables from 4 down to 2. (You do this by commenting / uncommenting the variable COLUMNS\_FOR\_ANALYSIS.)
2. Rerun the script, generating three clusters.
3. **DISCUSS**: How do we read**, and interpret,** the following table? To **interpret** this output, you must relate it back to the survey questions. Use words that a person not familiar with the survey questions can understand.



|  |
| --- |
| Next time, we will take another look at this unsupervised model and envision how such a model might be used in an Information System that incrementally approves over time. |

# Survey Questions

|  |  |  |  |
| --- | --- | --- | --- |
| i1 | Choose the answer that BEST describes you. When do you watch the video lectures? 1 - Within approximately 1 day of being made available. 2 - Within approximately 1 week of being made available. 3 - Roughly 24 hours before the relevant quiz or exam. 4 - I don't watch the video lectures. | 1 thru 4 | Outside of class |
| i2 | Choose the answer that BEST describes you. When I watch the video lectures … 1 - I am writing code / following along with the instructor. I might go back later and take some notes. 2 - I am taking detailed notes. 3 - I am multitasking. I am watching other media, doing other tasks. 4 - I am skimming for the quiz. I don’t watch the full video. 5 - I don’t watch the video lectures. | 1 thru 5 | Outside of class |
| i3 | Indicate if you agree with the following statement: When the video lecture has a practical demonstration of a concept (for example, typing a command, or writing some code) I always stop the video and try it myself. | 1 thru 5 | Outside of class |
| i4 | Indicate if you agree with the following statement: I feel like I spend more time on the video lectures than I should have to. | 1 thru 5 | Outside of class |
| i5 | Indicate if you agree with the following statement: This "flipped classroom” approach has forced me to spend more time thinking about MIS3502 topics outside of class. | 1 thru 5 | Outside of class |
| i6 | Indicate if you agree with the following statement: When I come to class I am prepared to work. | 1 thru 5 | During class |
| i7 | Indicate if you agree with the following statement: When I leave class I feel as though I was able to follow along.. | 1 thru 5 | During class |
| i8 | Indicate if you agree with the following statement: I make a strong effort to follow along with the instructor during class time. | 1 thru 5 | During class |
| i9 | Indicate if you agree with the following statement: I spend more time writing code (inside and outside of class) because of the "flipped classroom" approach. | 1 thru 5 | Whole course |
| i10 | Indicate if you agree with the following statement: Future versions of this course should use the "flipped classroom" approach. | 1 thru 5 | Whole course |