Weeks 11 & 12 Overview

You have made it to the last two weeks of the course! These two weeks were combined to give you some extra time to focus on your final project and two large exercises. This week we will get to practice doing some machine learning – which most of what you have already done is a form of machine learning. We will be learning about K-Means Clustering and Nearest Neighbors, along with some basic terminology.

Don't forget that when two weeks are combined, you have to do 10 discussion/participation posts **each** week, for a total of 20 in Weeks 11 & 12. As always, there are topics to aide you in discussion with your peers. Your projects are a great topic of discussion, as well as providing some feedback about the course and what you have learned.

Best of luck to all of you in your future courses in the program!

I wanted to make a note about the future. One thing we didn't cover in this course is a form of statistics called, "Bayesian Inference." In the readings section, you will find a course on understanding the basics of Bayesian. You do not have to read it but eventually, you'll probably need to know what it is. It is in your best interest to go and check it out. However, maybe it's better for a bit later? After some rest? You earned it!

Contents of the Week

Objectives

After completing this week, you should be able to:

- Present a set of well-defined research question(s)
- Utilize appropriate visualizations to tell the story of the data
- Write a coherent narrative that tells a story with data
- Summarize a problem statement
- Describe the data and methodology used for the research
- Summarize interesting insights of the analysis conducted
- Summarize the ethical implications of your analysis to a selected target audience
- Discuss the limitations of your analysis
- Understand high level machine learning concepts
- Use the nearest neighbors algorithm to fit a model to datasets
- Use unsupervised algorithms to extract structure from datasets

Weekly Resources

Comprehensive R Archive Network

R Studio

Sage Publications. (2021). Discovering Statistics Using R

RStudio, PBC. (2021). RStudio Documentation

RStudio, PBC. (2021). R Studio Cheatsheets

Weeks 11 & 12 Readings, Assignments, and Tasks

Here are your tasks for this week:

Read the following:

- *R for Everyone:* Chapters 23-25
- Machine Learning Fundamentals
  - Bernard Marr. (2016). Supervised V Unsupervised Machine Learning – What's The Difference?
  - Bernard Marr. (2016). What Is The Difference Between Artificial Intelligence And Machine Learning?
  - Bernard Marr. (2016). What Is The Difference Between Deep Learning, Machine Learning and AI?
- K-Means Clustering
  - Sejal Jaiswal. (2018). K-Means Clustering in R Tutorial
- Nearest Neighbors Classification
  - Kevin Zakka. (2016). A Complete Guide to K-Nearest-Neighbors with Applications in Python and R
  - Scikit Learn. Nearest Neighbors Classification

Complete the following:

- 11.1 Discussion/Participation
- 12.1 Discussion/Participation
- 11.2 Exercise
- 11.3 Final Project Step 3

Helpful Sources

Clark, Michael. (2018). Bayesian Basics.

11.1 Discussion/Participation

Here are optional topics for discussion via Teams this week.

Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is machine learning?
2. What is the difference between machine learning, artificial intelligence and data science? Why do these terms all tend to overlap?
3. What is supervised learning vs unsupervised?
4. What is deep learning? Is it the same as artificial intelligence?
5. What are nonlinear model examples?
6. What is nonlinear least squares?
7. What is a spline?
8. What are generalized additive models (GAMS)?
9. What are decision trees and when are they used?

12.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is boosting?
2. What is Random Forests? When is it used?
3. What is clustering? What is K-Means? What are the pros/cons or uses for each?
4. What is k-nearest neighbors?
5. What is K-Medoids?
6. What is hierarchical clustering and when is it needed?
7. Why is it so important to understand your data and the type of data it is when creating a model?
8. What have you learned in this course? What are some gaps you have discovered in your learning? What do you wish had been covered more?

Discussion for the final week of class is due by **Saturday of Week 12**, 11:59 p.m. CT

11.2 Exercise

Complete the following 2 exercises in RMarkdown and submit as a PDF:

1. Introduction to Machine Learning
    a. These assignments are here to provide you with an introduction to the "Data Science" use for these tools. This is your future. It may seem confusing and weird right now but it hopefully seems far less so than earlier in the semester. Attempt these homework assignments. **You will not be graded on your answer but on your approach**. This should be a, "Where am I on learning this stuff" check. If you can't get it done, please explain why.
    b. Include all of your answers in a R Markdown report.
    c. *Regression* algorithms are used to predict numeric quantity while classification algorithms predict categorical outcomes. A spam filter is an example use case for a classification algorithm. The input dataset is emails labeled as either spam (i.e. junk emails) or ham (i.e. good emails). The classification algorithm uses features extracted from the emails to learn which emails fall into which category.
    d. In this problem, you will use the nearest neighbors algorithm to fit a model on two simplified datasets. The first dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables (You worked with this dataset last week!). The second dataset (found in trinary-classifier-data.csv) is similar to the first dataset except that the label variable can be 0, 1, or 2.
    e. Note that in real-world datasets, your labels are usually not numbers, but text-based descriptions of the categories (e.g. spam or ham). In practice, you will encode categorical variables into numeric values.
        i. Plot the data from each dataset using a scatter plot.
        ii. The k nearest neighbors algorithm categorizes an input value by looking at

Submission Instructions

For all assignments in this course, you must export the script or Markdown file to PDF. You are welcome to submit your URL to GitHub in addition, but all submissions must include a PDF (no zip files will be accepted either).

The assignment is due by **Saturday of Week 12**, 11:59 p.m. CT.

the labels for the k nearest points and assigning a category based on the most common label. In this problem, you will determine which points are nearest by calculating the Euclidean distance between two points. As a refresher, the Euclidean distance between two points:

$$p_1 = (x_1, y_1)$$

and

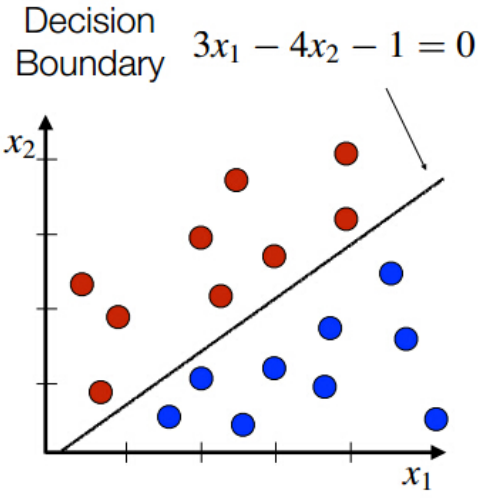
$$p_2 = (x_2, y_2)$$

is

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

i. Fitting a model is when you use the input data to create a predictive model. There are various metrics you can use to determine how well your model fits the data. For this problem, you will focus on a single metric, accuracy. Accuracy is simply the percentage of how often the model predicts the correct result. If the model always predicts the correct result, it is 100% accurate. If the model always predicts the incorrect result, it is 0% accurate.

ii. Fit a *k* nearest neighbors' model for each dataset for *k=3, k=5, k=10, k=15, k=20,* and *k=25*. Compute the accuracy of the resulting models for each value of k. Plot the results in a graph where the x-axis is the different values of k and the y-axis is the accuracy of the model.



i. Looking back at the plots of the data, do you think a linear classifier would work well on these datasets?

ii. How does the accuracy of your logistic regression classifier from last week compare? Why is the accuracy different between these two methods?

2. Clustering
   a. These assignments are here to provide you with an introduction to the "Data Science" use for these tools. This is your future. It may seem confusing and weird right now but it hopefully seems far less so than earlier in the semester. Attempt these homework assignments. You will not be graded on your answer but on your approach. This should be a, "Where am I on learning this stuff" check. If you can't get it done, please explain why.
   b. Remember to submit this assignment in an R Markdown report.
   c. Labeled data is not always available. For these types of datasets, you can use unsupervised algorithms to extract structure. The k-means clustering algorithm and the k nearest neighbor algorithm both use the Euclidean distance between points to group data points. The difference is the k-means clustering algorithm does not use labeled data.
   d. In this problem, you will use the k-means clustering algorithm to look for patterns in an unlabeled dataset. The dataset for this problem is found at data/clustering-data.csv.
      i. Plot the dataset using a scatter plot.
      ii. Fit the dataset using the k-means algorithm from *k=2* to *k=12*. Create a scatter plot of the resultant clusters for each value of k.
      iii. As k-means is an unsupervised algorithm, you cannot compute the accuracy as there are no correct values to compare the output to. Instead, you will use the average distance from the center of each cluster as a measure of how well the model fits the data. To calculate this metric, simply compute the distance of each data point to the center of the cluster it is assigned to and take the average value of all of those distances.
   e. Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.
   f. One way of determining the "right" number of clusters is to look at the graph of k versus average distance and finding the "elbow point". Looking at the graph you generated in the previous example, what is the elbow point for this dataset?

11.3 Final Project Step 3

You are now on to the final
research paper. While this
require you build a model,
welcome to do so if you fee
time. Instead, you need to

recommendation for the approach you would take and what the remaining steps would be using the information you have learned in this course to take this project from simply being an analysis exercise to proposed implementation of a solution.

- Overall, write a coherent narrative that tells a story with the data as you complete this section.
- Summarize the problem statement you addressed.
- Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented).
- Summarize the interesting insights that your analysis provided.
- Summarize the implications to the consumer (target audience) of your analysis.
- Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

You can use the following template for Step 3:

- A story / narrative that emerged from your data. Follow this structure.
    - Introduction.
    - The problem statement you addressed.
    - How you addressed this problem statement
    - Analysis.
    - Implications.
    - Limitations.
    - Concluding Remarks