Introduction to Statistical Modeling MSDS 598

Introduction to this course

Michael Ruddy

Overview

- Subject of this course
- How this course will operate
- EDA Review
- Probability and Statistics Review

Why Statistical Modeling?

- Introduction to basic concepts of modeling
  - Statistics
  - Machine Learning

- Practice doing data science in Python with notebooks
  - Pandas, Matplotlib, Numpy, Statsmodels, Sci-kit learn, etc.

- Sometimes Linear/Logistic Regression is *best*
  - More interpretable
  - Computationally cheap
  - Rich, rigorous statistical theory

About Me

- Michael Ruddy (he/him)
  - PhD in Mathematics from North Carolina State, 2019
  - Max Planck Institute Math in the Sciences (Leipzig, Germany)
  - USF Data Science

- Research in geometry, deep learning, and its intersection

- New to California!

- Bouldering, Gardening, Cooking

Today

- Exploratory Data Analysis Review
  - Why explore data before modeling?
  - Why visualize data?

- Popular libraries
  - Pandas
  - Matplotlib

- Statistics and Probability Review
  - Probability, Random variables, Confidence intervals etc.

Tentative Schedule

- Available on GitHub

- 1/31: Linear Regression
- 2/7: Basics of Modeling Concepts
- 2/14: Feature Engineering and Importance
- 2/21: No Class
- 2/28: Logistic Regression
- 3/7: Regularization and Bias/Variance

Modality

- Hybrid
  - In-person class
  - Recorded lectures available to everyone

- Anyone can take the remote option
  - Must be synchronous without prior authorization
  - EXCEPT anyone showing symptoms associated with COVID-19

- Class will be fully remote if I show symptoms associated with COVID-19
  - (for the weeks I show symptoms, test positive, etc.)

Course Structure

- Lectures
  - Slides/Notebooks/Lab
  - Breaks!
  - Quizzes OR HW Review

- Assessment
  - Timed Quizzes
  - Homework Assignments
  - Final Case Study

Assessment

- Labs
  - Worked on in-class, due that Sunday, completion grade
- Quizzes
  - On Canvas
  - Beginning of class
  - **Not** open book
- Homework
  - Due after one week
  - Start class going over
- Case Study (Final), Due March 10th
  - Real dataset
  - Groups possible (Feb 19th)
  - Notebook report (nice format!)

Grade Outline

- Participation 20%
  - Attendance/Remote as dictated
  - Don’t come if you are sick!
  - In-class Labs (completion graded)
- HW Assignments 30%
  - Submit as Notebook
  - Appropriate Formatting! (Name, Q#’s, etc.)
  - Late Policy 10%/20%/50% for < 1 day/2 day/1 week
- Quizzes 20%
- Final 30%
  - Long-Form HW

What you'll need

- Laptop required (bring it to class)
  - Labs and Quizzes

- Some familiarity with basic Python
  - Kaggle Mini-course

- Ability to use Python Notebooks
  - Jupyter Notebook (Anaconda suggested)
  - Visual Code Studio
  - Google Colab (quick and easy option)

Resources

- Office Hours: Wednesdays from 12pm - 1pm Virtually

- Questions outside of Office Hours?
  - Use the Canvas discussion board
  - Email only for matters concerning just you
  - Helps everyone learn
  - Get your answers faster, your classmates can/should answer!

- Google

Resolving Errors in Code

- Google

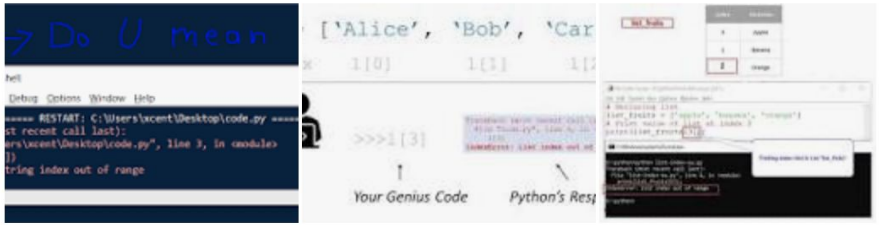
```
---------------------------------------------------------------------------
IndexError                                Traceback (most recent call last)
<ipython-input-3-e4d12b5190dc> in <module>
     12     for file in os.listdir(str(mask_path)):
     13         split_file = file.split('_')
---> 14         if (split_file[0] == patient_no) & (split_file[1] in masks):
     15             patient_masks.add(split_file[1])
     16     if patient_masks == set(masks):

IndexError: list index out of range
```

IndexError: list index out of range

All Videos Books News Images More Settings Tools

About 551,000 results (0.61 seconds)



View all

The error "**list index out of range**" arises if you access invalid indices in your Python **list**. For example, if you try to access the **list** element with **index** 100 but your lists consist only of three elements, Python will throw an **IndexError** telling you that the **list index** is **out of range**.

blog.finxter.com › python-indexerror-list-index-out-of-ra...

Python IndexError: List Index Out of Range (How to Fix This ...

About featured snippets • Feedback

stackoverflow.com › questions › indexerror-list-index-...

IndexError: list index out of range and python - Stack Overflow

Academic Honesty

- **Never** copy code without crediting the source
  - In real-life
  - In this course
  - Academic honesty violation

- Do not replicate/use code you do not understand in this course

- Disclose all outside help on assignments (classmates, stack exchange posts, etc.)