Math 107 Lecture 23

Data Science and Least Squares

by Dr. Kurianski
on December 2, 2024

» Today's Announcements and Objectives

Announcements

- HW13 Due Friday 12/6 (last homework of the semester!)
- Skill Check 7 is on December 9
- Fill out your SOQs!
- End-of-Semester Survey still available
- Final Exam Review Topic Survey still available
- Final exam will be
  Wednesday, December 18 from 3:00 PM to 4:50 PM

Objectives

- Explore examples of linear algebra applications
- Describe random walks using matrices
- Use Google's PageRank algorithm to rank a given network of websites

Student Opinion Questionnaires

Examples of Data

» New York Times COVID-19 Data

One of the most prevalent examples of data in our lives today is COVID-19 data. The New York Times has extraordinary visualizations of data available: NYT COVID Data.

Below you can see COVID cases per 100,000 people in the United States from 2021 (left) and 2024 (right).

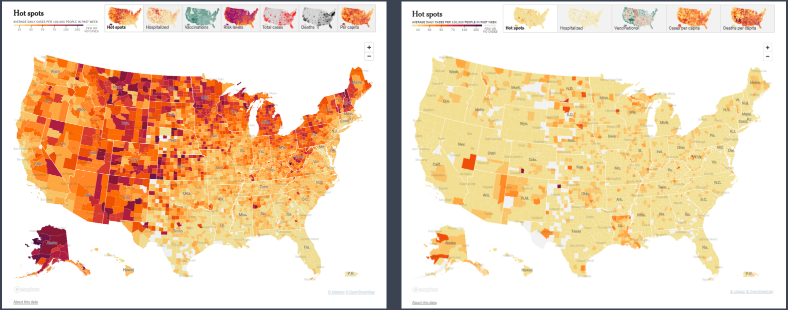

Image from https://www.nytimes.com/interactive/2021/us/covid-cases.html

» Center for Disease Control COVID-19 Forecasts

There are dozens of statistical models developed by various scientific groups using different assumptions and techniques. These models attempt to predict the state of hospitalizations in the coming weeks.

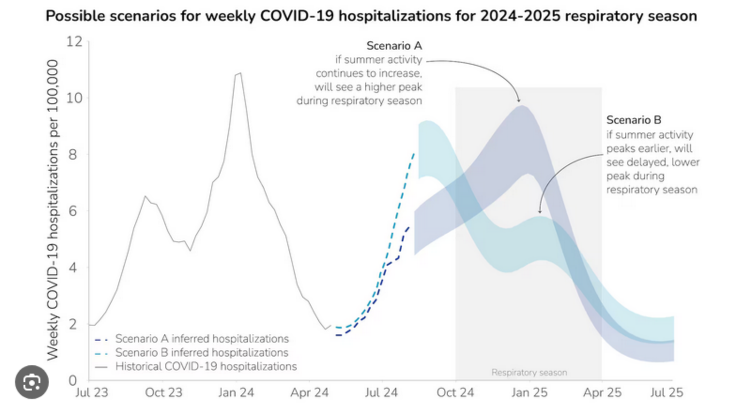

Image from www.cdc.gov

Dataset Exploration

» Math Scores Dataset

What’s in the dataset?

- Math scores from standardized test
- Years 2011, 2012, 2013
- Six schools labeled A, B, C, D, E, F, G, H
- Gender of each student (female or male)

» Math Scores Dataset

What’s in the dataset?

- Math scores from standardized test
- Years 2011, 2012, 2013
- Six schools labeled A, B, C, D, E, F, G, H
- Gender of each student (female or male)

What’s not in the dataset?

- The fact that a new pedagogical strategy was attempted between years 2012 and 2013 in an attempt to improve scores
- External factors that might impact scores

» Math Scores Dataset

What’s in the dataset?

- Math scores from standardized test
- Years 2011, 2012, 2013
- Six schools labeled A, B, C, D, E, F, G, H
- Gender of each student (female or male)

What’s not in the dataset?

- The fact that a new pedagogical strategy was attempted between years 2012 and 2013 in an attempt to improve scores
- External factors that might impact scores
- What can you think of that’s not accounted for in the dataset?: **menti.com** **Use code: 9828 2638**

Fitting Data with a Line

» Main Idea

Data is often not linear (it does not fall along a straight line). But sometimes, we can try to capture the trend of a data set using a "best fit" line. This allows us to make predictions about what might happen to the data in the future.

Main Idea: We want to find the "best" numbers m and b so that

$$y = mx + b$$

describes the data well.

» Residual vector

We want to find a line that is "closest" to all the points. But what do we mean by "closest"?

Say we picked a value for m and a value for b to represent our data by the line $y = mx + b$. Our predicted values for the data would be

$$A \begin{bmatrix} m \\ b \end{bmatrix} = \vec{v}_{predict}$$

To see how good our line is at modeling the data, we would compare $\vec{v}_{predict}$ to the true values on the right-hand side $\vec{v}_{true}$.

Definition: The **residual vector** measures how far off our predicted values are from the true values

$$\vec{r} = \vec{v}_{true} - \vec{v}_{predict}.$$

» Least Squares error

Definition: A good choice for the line $y = mx + b$ is the one that produces the shortest residual vector. So we want to compute

$$||\vec{r}||_2^2 = (v_{true,1} - v_{predict,1})^2 + \cdots + (v_{true,n} - v_{predict,n})^2$$

which is called the **least squares error**.

Least Squares Fit

» Least Squares Method

We want to find m and b so that

$$A \begin{bmatrix} m \\ b \end{bmatrix} = \vec{v}_{true}$$

produces the shortest residual vector.

It turns out that the way to do this is to multiply both sides of the equation by A^T on the left:

$$A^T A \begin{bmatrix} m \\ b \end{bmatrix} = A^T \vec{v}_{true}$$

and then solve this system for m and b.
This means putting the following augmented matrix in reduced row echelon form:

$$\left[A^T A \quad | \quad A^T \vec{v}_{true}\right] \to \text{rref} \to \left[\begin{array}{cc|c} 1 & 0 & m \\ 0 & 1 & b \end{array}\right]$$