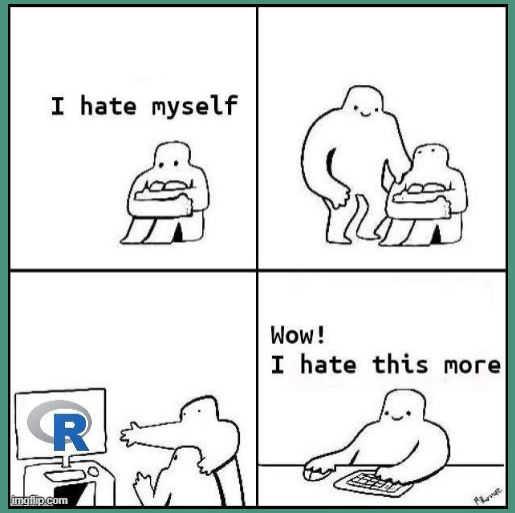
DATA 180
[Intro to Data Science]

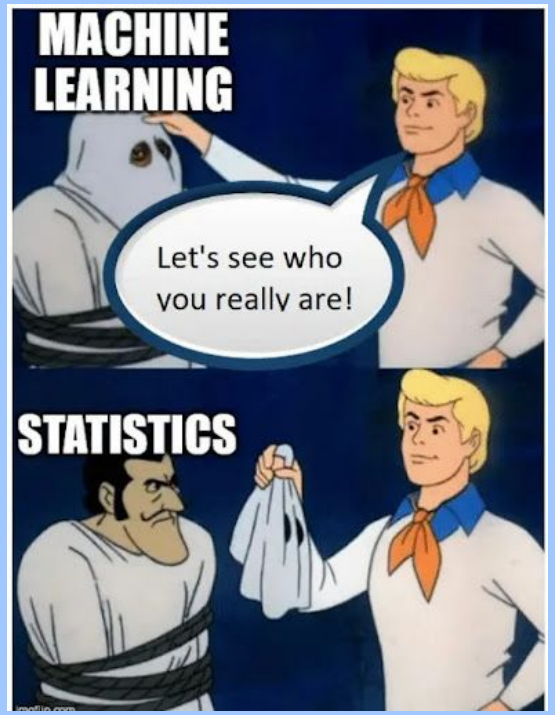


Andrew, Ben, Ellison, Kieran

Feb. 4, 2025

1. Probability Warm Up
2. dplyr Refresher (ahhhhhhhh!)
3. Distributions 101 (this one's normal...ha! get it?)
4. Supervised Learning
5. (Un)supervised Learning
6. Interactive Case Study

Overview



The Two-Child Problem

A family has two children. You know that at least one of them is a boy. What is the probability that both children are boys?

(Hint: Assume that each child is independently equally likely to be a boy or a girl.)

Problem

One in three!

1. The (equally likely) possible combinations of children are: boy-boy, boy-girl, girl-boy, and girl-girl.
2. Since we know already that *at least one* of the children is a boy, that eliminated the girl-girl possibility.
3. In only one out of the three possible cases (boy-boy), are both the children boys, so we end up with ⅓.

Answer

The Birthday Paradox

In a room of 23 people, what is the probability that at least two people share the same birthday?

(Hint: there are 365 days in a year…duh)

Problem

Just over 50%!

1. It is simpler to calculate the probability that no one shares a birthday, then subtract it from 1.
2. Then, the equation below follows:

$$P(\text{no shared birthday}) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \ldots \times \frac{343}{365}$$

3. This comes out to **P = 0.4927**, so if we subtract that from 1, we get **0.5073**

Answer

Monty Hall Problem

You are on a game show with three doors: Behind one is Prof. Bilen's car, behind the other two are tickets to a show by the Math & CS department's faculty band. You pick a door. The host, who knows what's behind the doors, opens another door revealing tickets. You are given the option to switch or stay. What should you do to maximize your chances of winning Prof. Bilen's car?

(Hint: Prof. Bilen never agreed to give away his car…)

Problem

Switch, dummy!

1. At first, the car is equally likely to be behind all the doors.
2. Let's say you choose Door 1. The host will reveal tickets behind Door 2 or Door 3.
3. Now, the choice: if the car was behind Door 1 (original choice), you win by staying. But, if the car was behind Door 2 or Door 3, you win by switching.
4. The probability that the car is behind your original choice is ⅓ and the probability that the car is behind one of the other doors is ⅔. So, you should switch!

Answer

dplyr, a review

1.

What is tidy data? Making sure we understand what tidy data is and how to work with it.

2.

How do we group data using dplyr? Grouping is one of the most important data wrangling functions.

3.

How can we manipulate data using dplyr? From filtering down by attribute to changing a column.

4.

How do we combine tables? good data scientists know how to merge data sets.

Rows

...all represent different **variables**

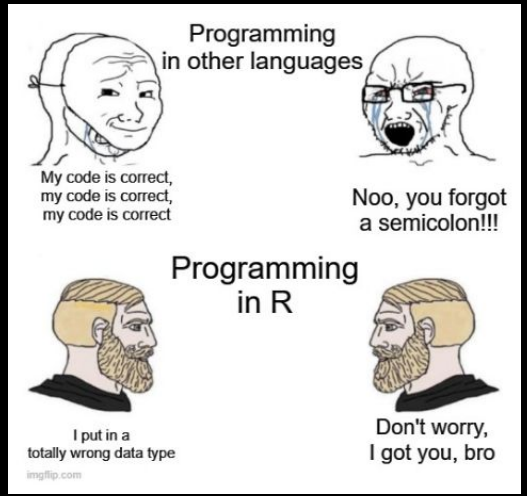
Columns

...all represent different **cases** or **observations**

Cells

...all contain a single **value**

What is tidy data?



group_by()

What it does: Create a grouped copy of a table grouped by columns

Code:

```
mtcars |>
        group_by(cyl) |>
       summarize(...)
```

Grouping observations

mutate()

What it does: Computes new columns by performing an operation on current columns

Code:

```
mtcars |>
    mutate(gpm = 1/mpg)
```

Mutating data

filter()

What it does: Extract rows that meet logical criteria, using operators like ‘<’, ‘==’, etc.

Code:

```
mtcars |> filter(mpg >20)
```

Filtering data

bind_cols(), left_join(), right_join(), inner_join(), full_join()

What it does: Join columns side-by-side (bind_cols) or join matching values in two tables (others)

Code:

```
left_join(x, y, by = "A")
```

Combining tables

Key functions

This is a collection of the most important functions to remember when using dplyr to deal with tidy data sets. If you have a good command of these tools, you’ll be wrangling all kinds of nasty data sets with ease!

Pnorm() - cumulative density function for normal distribution

- # Probability that a standard normal variable is less than or equal to 1.96
- pnorm(1.96)
- # Output: 0.9750021

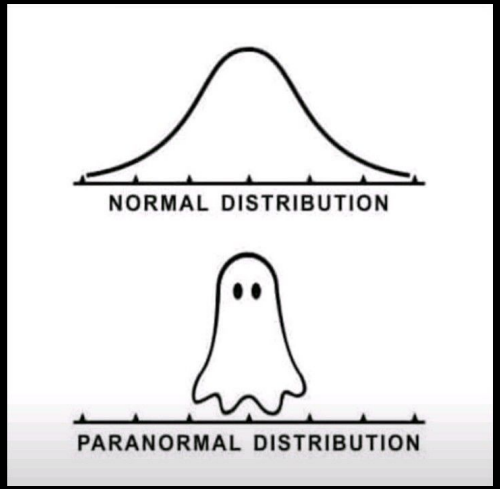
Rnorm - random numbers from a normal distribution

- # Generate 10 random numbers from a standard normal distribution
- rnorm(10)
- # Output: a vector of 10 random numbers

Qnorm - quantile function, inverse CDF

- # 95th percentile of the standard normal distribution
- qnorm(0.95)
- # Output: 1.644854

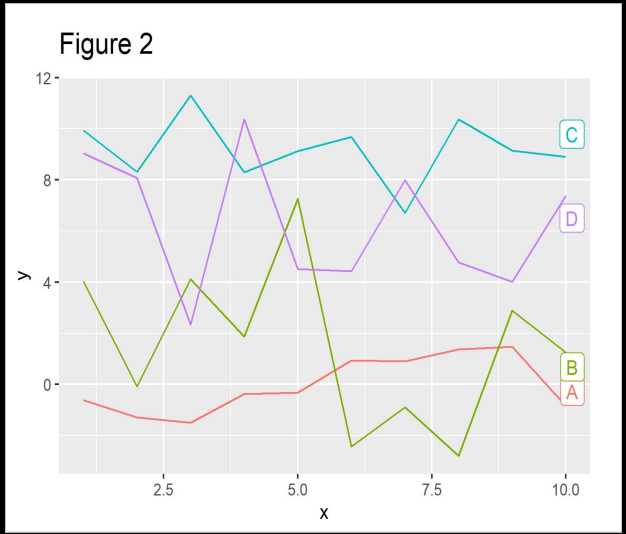
Distributions



How to display

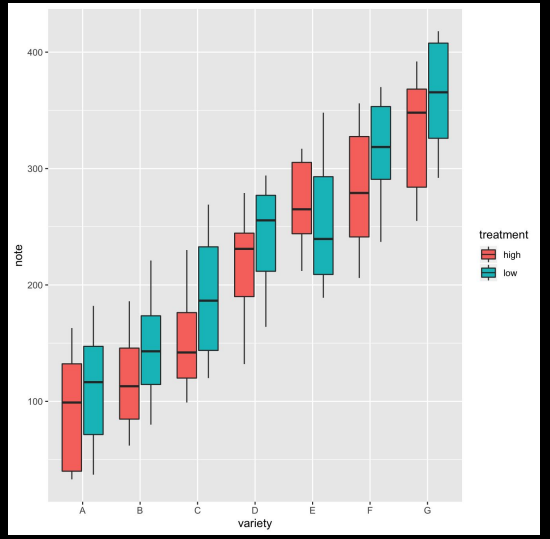
Lorem ipsum dolor sit amet, quo graecis expetenda reprehendunt et. Et has nulla intellegat. Ea vix equidem abhorreant deseruisse, eos quod suas labore ex.

Visualization library in R



Cool resources

- ggplot2 book
- ggplot2 coloring guide



Let's do a demo!

ggplot2

01

Definition: Supervised learning is a category of machine learning that uses labeled datasets to train algorithms to predict outcomes and recognize patterns.

02

Goal: To train a model to map inputs to correct outputs by learning from labeled training data.

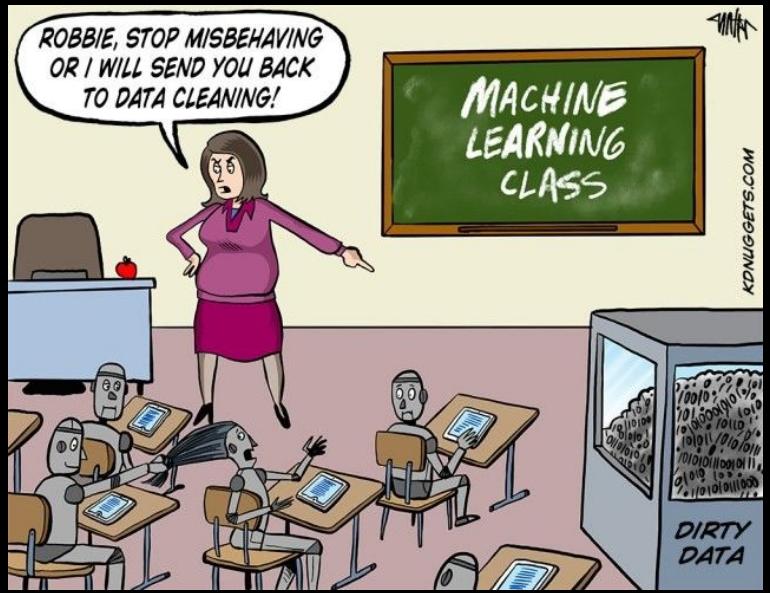
03

Common Applications: Image recognition, spam detection, fraud detection, speech recognition.

04

How is it different?: Supervised learning uses labels, while unsupervised learning finds patterns.

Supervised Learning



Supervised Learning Models

Linear Regression

Method: Models the relationship between input features and a continuous output by fitting a straight line.
Applications: Stock price prediction, sales forecasting, housing price estimation

KNN (K-Nearest Neighbors)

Method: predicts outcomes based on the majority label or average of the nearest data points.
Applications: Image classification, recommendation systems, and fraud detection.

Logistic Regression

Method: models the probability of a binary outcome using a sigmoid function.
Applications: Spam detection, disease prediction, and credit scoring.

Case Study: A city is working to improve its public transportation system. They have collected data on factors such as the number of daily riders, average wait times, bus routes, and traffic conditions.

The goal is to predict the level of demand for public transportation in different neighborhoods and adjust services accordingly.

Which ML model would you use?

Testing your knowledge

1. **Linear Regression**

OR

2. **KNN**

OR

3. **Logistic Regression**

Linear Regression

Justification:

- Predicts the continuous demand for transportation based on features like population density.
- May not be best if the data is complex

KNN (K-Nearest Neighbors)

Justification:

- Handles complex, non-linear relationships between factors like neighborhood characteristics.
- Could provide clear results if the dataset has clusters of similar neighborhoods

Logistic Regression

Justification:

- Models the probability of a neighborhood having high or low demand.
- Good for providing you with clear probabilities

Discussion

01

Definition: Machine learning where the model searches for a structure in *unlabeled data*.

02

Goal: Identify patterns, relationships, and groupings in data sets.

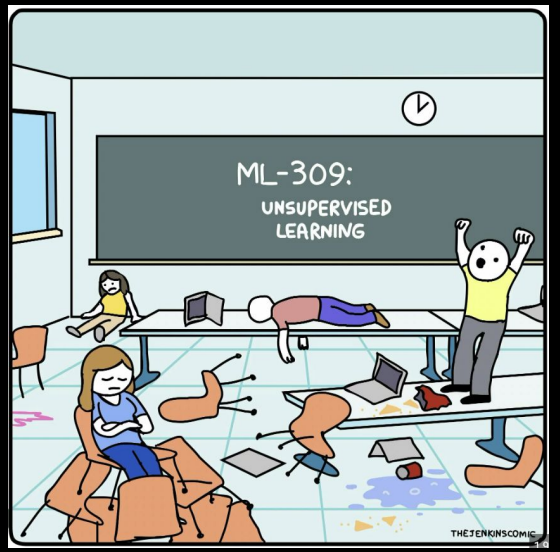
03

Common applications: customer segmentation, anomaly detection, topic modeling, geospatial analysis

04

How is it different? Unlike supervised learning, unsupervised techniques don't have an explicit target variable

Unsupervised Learning



Unsupervised ML Models

k-Means

Method: Assigns data into *k* clusters based on centroids.

Applications: fraud detection, image compression, document clustering

Hierarchical Clustering

Method: Groups observations in a hierarchical manner, making “cuts” at appropriate distances.

Applications: social network analysis, biological taxonomy

DB-SCAN*

Method: Partitions data into clusters based on their distance to other points.

Applications: environmental studies, geospatial analysis, medical image analysis, anomaly detection

*We didn’t cover this in DATA 180, but the slide template has 3 boxes

Case study: the borough of Carlisle wants to select the best location for a new park. They have data on population density, existing green spaces, accessibility, demographics, and environmental factors.

The goal is to identify a series of locations that have the **highest need** of a park.

Which ML model would you use?

Testing your knowledge

1. **K-means clustering**

OR

2. **Hierarchical clustering**

OR

3. **DB-SCAN**

k-Means?

Justification:

- Cluster based on similar needs and characteristics
- Efficient for large data sets
- Provides clear, actionable groups

Hierarchical Clustering

Justification:

- Explore different levels of clustering
- Useful with no clear idea of how many clusters are necessary
- Dendrogram helps planners visualize relationships

DB-SCAN

Justification:

- Useful if geospatial data is provided, creates clusters of unspecified shapes
- Helpful if the data contains many outliers

Discussion

Download this dataset from Kaggle

- Top 100 TikTok Accounts

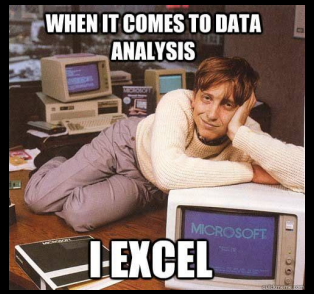
Tasks

- Exploratory data analysis with dplyr's summarise() function
- Create a graph in ggplot2
- Generate a linear regression model
- Variable selection

Goal

- Take time to perform data wrangling, visualization, and ML learning on the dataset
- We will go through it all together as a class

Let's Test Your Skills!



Thank you & hope you learned something!