COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 22: Classification

April 27, 2026

11	4/06	Inference		HW 6 - Confidence Intervals
	4/08		Lab 8 - Normal Distribution and Variability of Sample Means	HW 5 Due 11:59PM
12	4/13	Prediction		HW 7 - Sample Sizes and Confidence Intervals
	4/15		Lab 9 - Regression	HW 6 Due 11:59PM **Final Project Proposals** Due Friday, 4/17
13	4/20	Special Topics: Data Ethics (Computing Fellows Workshop)		HW 8 - Linear Regression
	4/22	Special Topics: Bias in AI (Guest lecture by Murad Megjhani)	Lab: Final Project Work Time	HW 7 Due 11:59PM
14	4/27	Prediction		**Final Project Progress Reports** HW 9 - Regression Inference (Optional, Due 5/4)
	4/29	(Either special topics or prediction)	Lab: Final Project Consultations	HW 8 Due 11:59PM
15	5/04	Wrap up!		HW 9 Due 11:59PM (Optional) **Final Project Report** Due Friday 5/8

Logistics

- Final Project Progress Reports due TONIGHT
  - Submit on Gradescope as a group
  - **Cannot submit late**
  - **Submit as a Python notebook (.ipynb)**
- Final Project Reports are due next Friday (May 8)
  - You will need to complete a (private) peer evaluation by May 8: https://forms.gle/bYj2UrSEEQry9Cry5
    - This is our method of checking that everyone is pulling their weight
    - Results will only be visible to teaching staff

Final Project Progress Report (Updated)

Information can be found on the course website:

https://www.eysalee.com/courses/s26/bc1016_final.html

Progress report (due Monday, April 27):

- Complete the exploratory data analysis section (graphs/tables + explanations)
- ~~Complete~~ Start the hypothesis testing section
  - State null and alternative hypotheses and test statistic
  - Start implementing the hypothesis test
    - Your implementation does *not* need to work
- Begin the prediction section
  - Anything you do not complete you should list out what remains and if there are any issues you are running into

Reminders about Final Projects

- If you want to use anything that was not explicitly covered in this class (e.g., more advanced stats, new Python libraries, …), you MUST state where you have learned this material before

- You are NOT allowed to use AI for your code or analysis

- Unless you receive explicit permission to use a different dataset, you may not use any datasets that were not provided

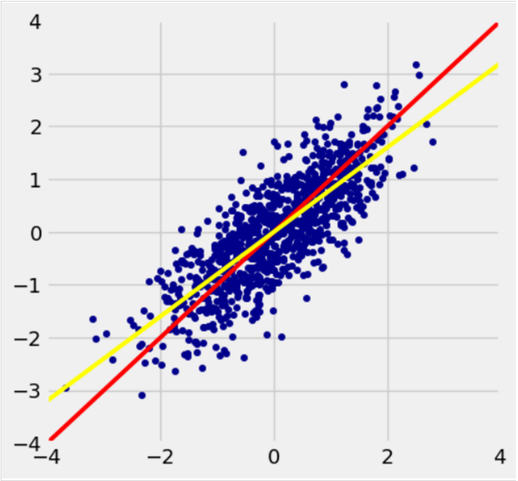
- All of this information is located on the final project page: https://www.eysalee.com/courses/s26/bc1016_final.html

- Violations may result in a failing grade for the project and the course

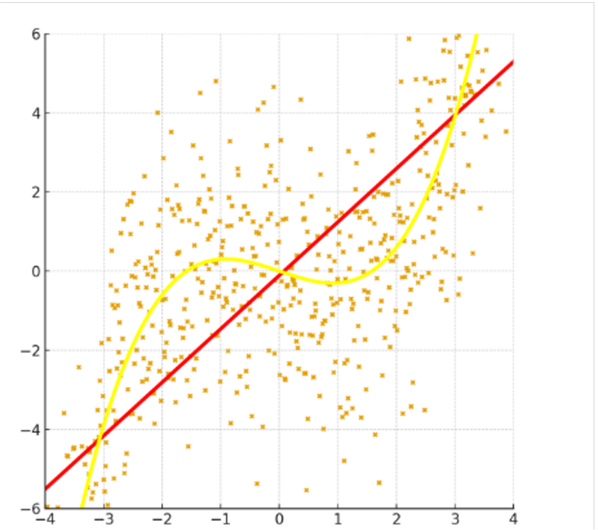
Last time, long, long ago…

Last Time: Least Squares and Residuals

How can we know we've created the best line to fit through our data (i.e., that we've minimized error)?



How can we check whether a line is appropriate (versus a non-linear model)?

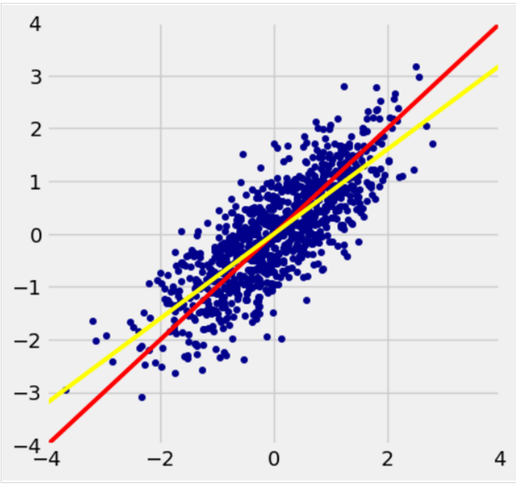


Last Time: Least Squares and Residuals

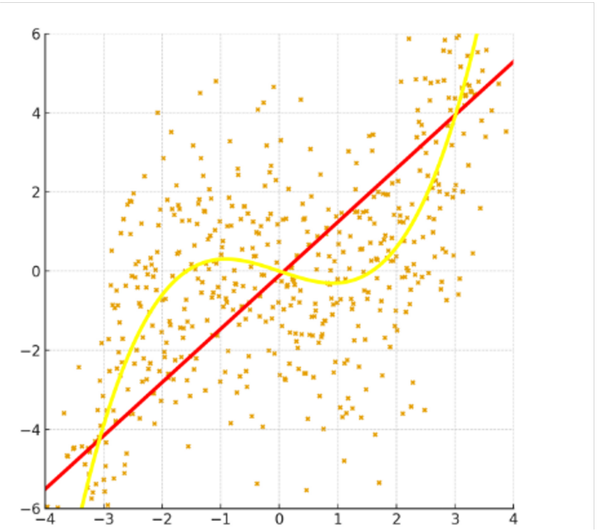
How can we know we've created the best line to fit through our data (i.e., that we've minimized error)?

Root Mean Square Error as a measure of error

Line of best fit minimizes this value



How can we check whether a line is appropriate (versus a non-linear model)?

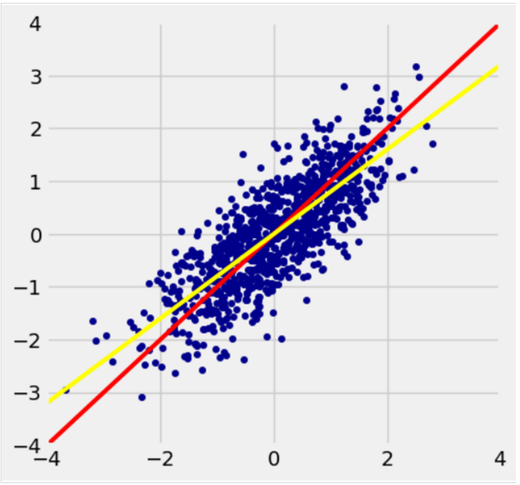


Last Time: Least Squares and Residuals

How can we know we've created the best line to fit through our data (i.e., that we've minimized error)?

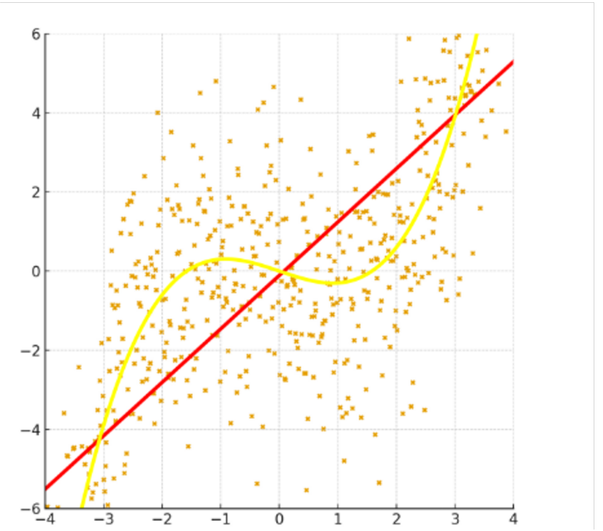
Root Mean Square Error as a measure of error

Line of best fit minimizes this value



How can we check whether a line is appropriate (versus a non-linear model)?

Visualize errors via residual plots



Last Time: Regression Inference

- **Estimate uncertainty** with a confidence interval for our **regression prediction**

    1. Uncertainty around our **predicted value** for a given x-value

    2. Uncertainty around our regression line **slope**

        - Do we think that the variables are linearly related?

Classification

Machine Learning

- **Machine learning** is a class of techniques for automatically finding patterns in data and using it to draw inferences or make predictions

- **Machine learning algorithms** are mathematical models that are calculated based on sample data to make predictions / decisions without being explicitly programmed to perform the task

Does this sound familiar?

Predicting Values

- Based on incomplete information

- One way to make predictions:

  - To predict an outcome for an individual

    Find others who are like that individual
    and whose outcomes you know

    Use those outcomes as the basis of
    your prediction

Predicting Values

- Based on incomplete information
- One way to make predictions:
  - To predict an outcome for an individual

    Find others who are like that individual and whose outcomes you know

    Use those outcomes as the basis of your prediction

<u>Two types of predictions:</u>

- **Regression**: Numeric
- **Classification**: Categorical

Spam or not?

Angell Animal Medic. Invoice 10/16/2024 - This is an automated notification. Please do not reply to this message.

Your Free $4,000 B. **>>eysa.lee!! - eysa.lee -Bonus Code: Use BETMGM**

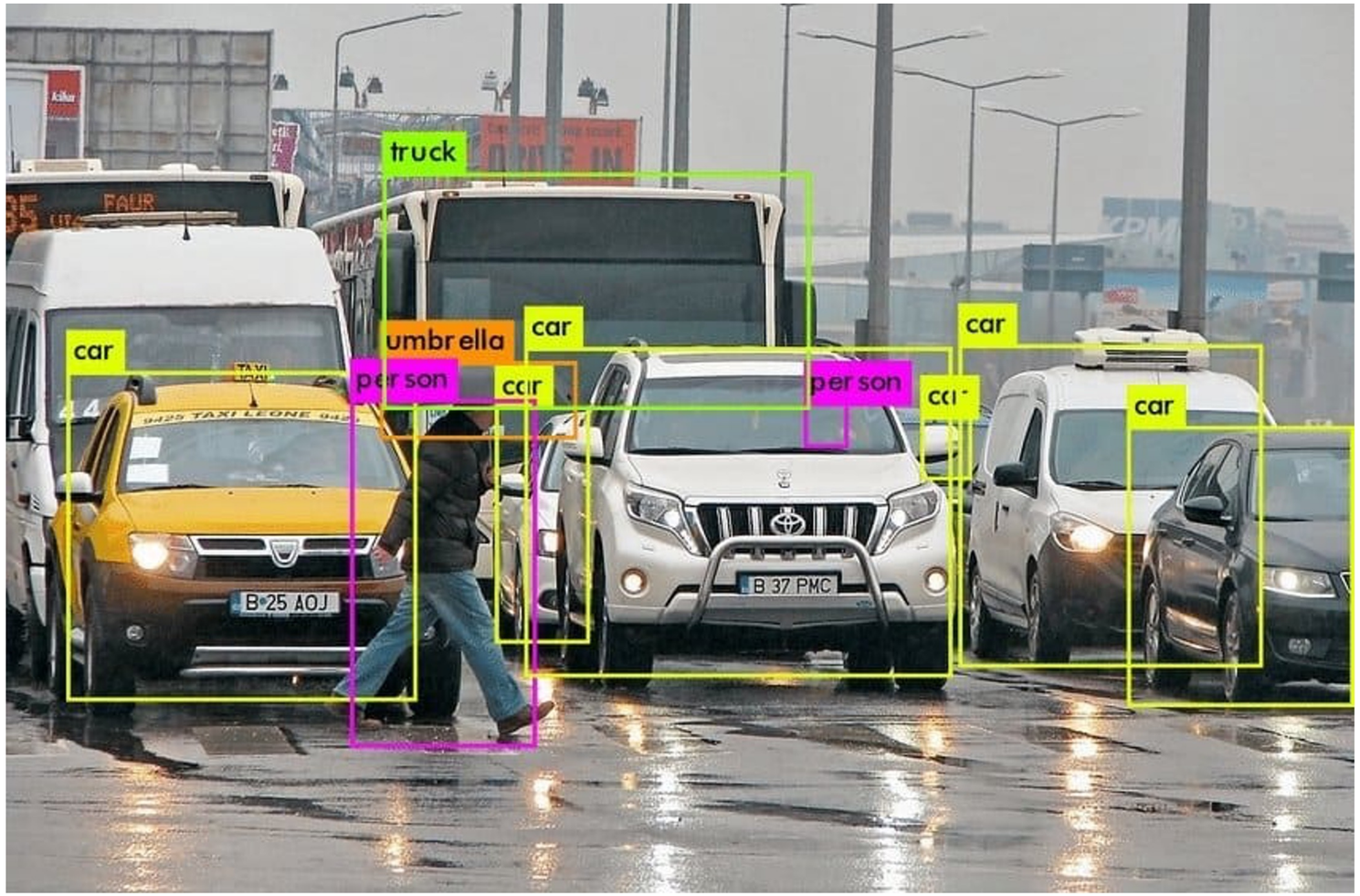
me, Riverside 3 Cat is constipated? - Good Morni...

no-reply Administrator has responded to your request for 'Internal Reports - Classified' - Good news. You now have access to 'Intern...

mailings [IACR] 2025 Election - Dear Eysa Lee, Dear IACR Member, The 2025 Election for Board positions is now open.

idontknowyou...@gmail.c... (no subject) - Hello, Please could you drop a contact to text you on, Thank you. Laura Rosenbury President Barnard College

Object Classification

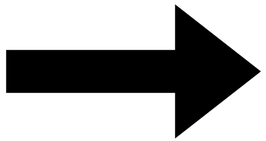


https://viso.ai/computer-vision/image-recognition/

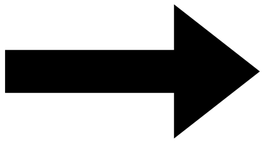
Predicting Categorical Values

Predicting Categorical Values

Attributes (features) of an example



Classifier



Predicted label of the example

Remember the cat census?

1. What do the rows represent?
2. What do the columns represent?
3. What are things we could try to predict with a classifier?

Name	Age	Weight	Coloring	Sex	Owner
Ruby	14	8	tuxedo	F	Alice
Gertrude	15	12	tuxedo	F	Alice
Hamby	8	16	tabby	M	Bob
Fig	3	7	tabby	F	Bob
Corina	6	10	tortie	F	Carol
Frito	2	8.5	tabby	M	Carol

Table Rows

- Data type of our columns are **`array`**
  - Array elements are the same type
- Data type of the rows is the **`row`** object
  - Row elements may be different types
  - If the row contains only elements of the same type, you can convert it to an array using `np.array(..)`

Name	Age	Weight	Coloring	Sex	Owner
Ruby	14	8	tuxedo	F	Alice
Gertrude	15	12	tuxedo	F	Alice
Hamby	8	16	tabby	M	Bob
Fig	3	7	tabby	F	Bob
Corina	6	10	tortie	F	Carol
Frito	2	8.5	tabby	M	Carol

Classification

- The attribute we want to predict is referred to as the class

Classification

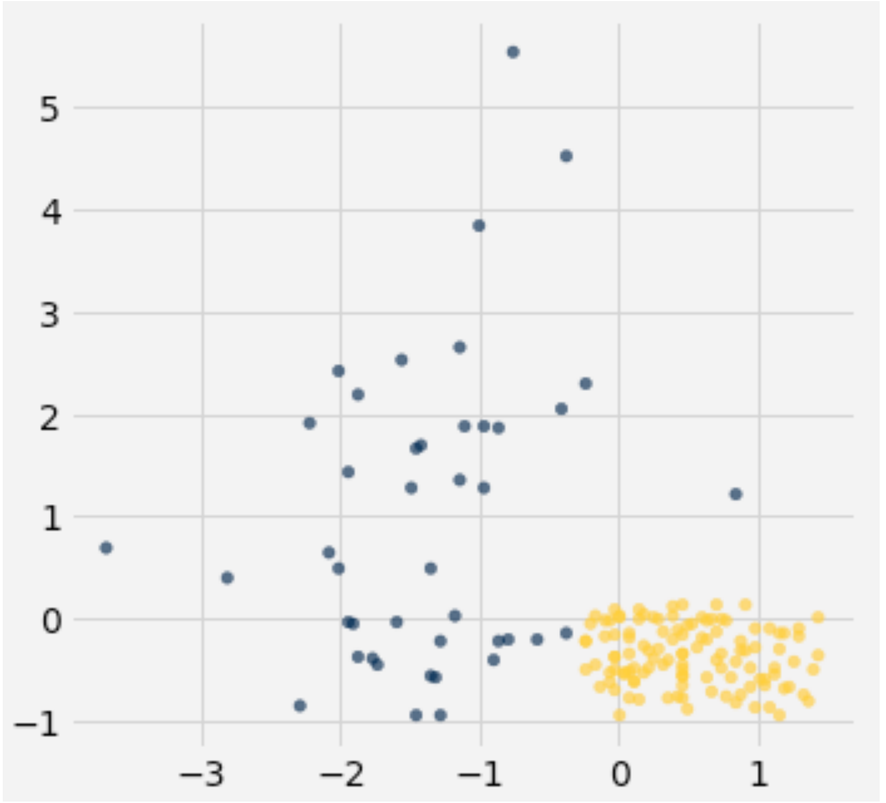
- The attribute we want to predict is referred to as the class
- To predict, we look for patterns in the existing data

Demo: Medicine Classification Example

Classifier Example

- Suppose the two colors indicate different groups (classes)

- How might we predict if new points fall within one group vs the other?



Distance between points

- Distance between two points with attributes x and y:

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$$

- Distance between points with attributes x, y, z:

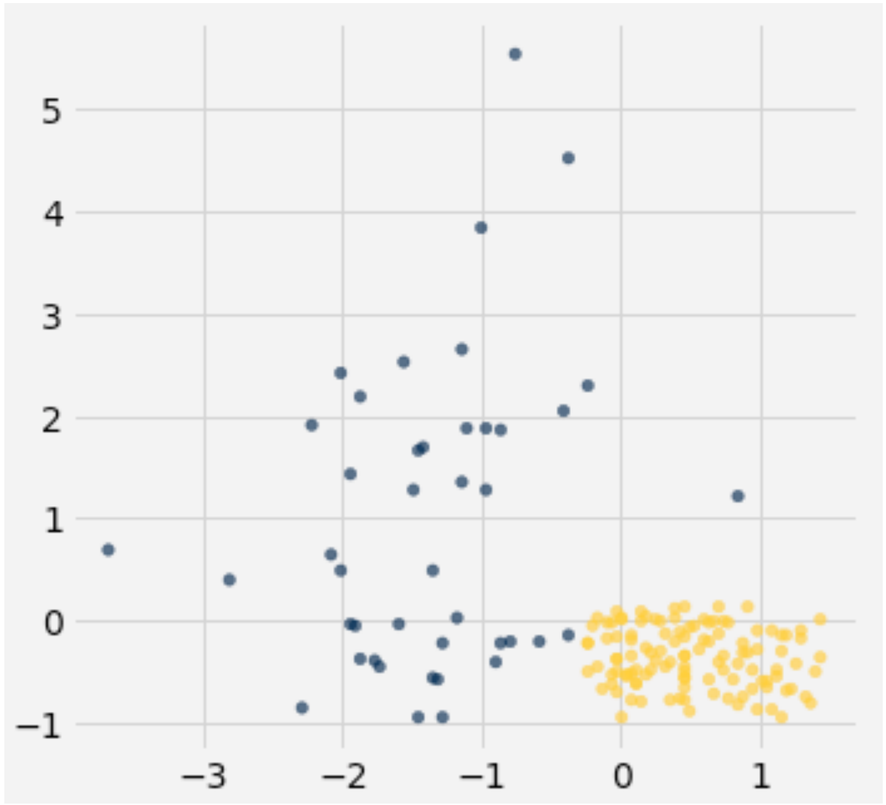
$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$$

Nearest Neighbor Classification

To classify a point:

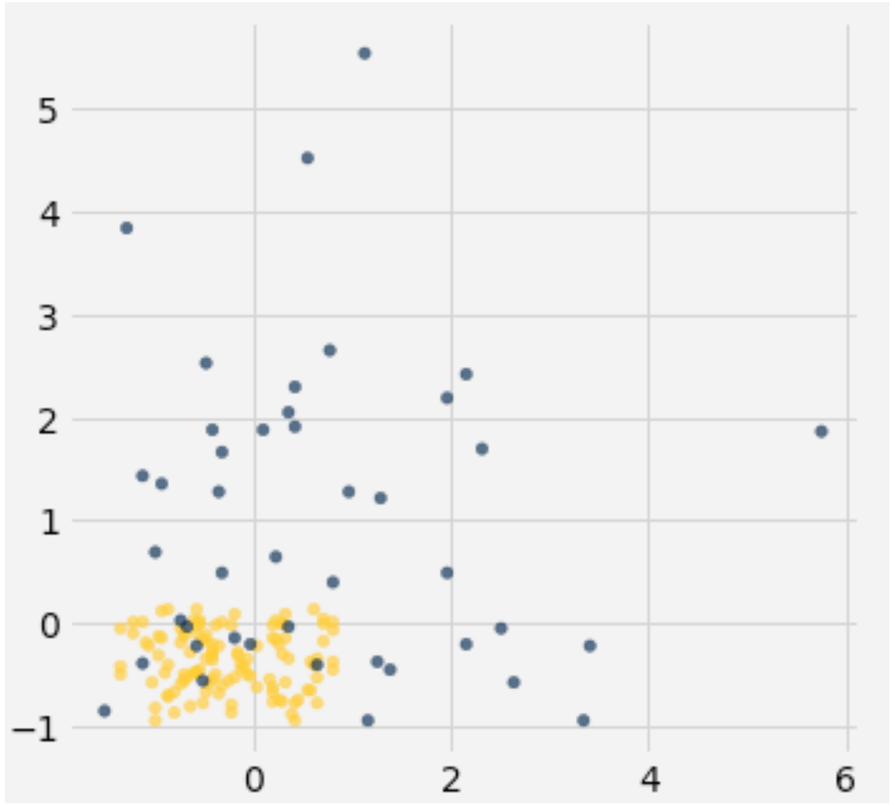
1. Find the nearest point to it (i.e., shortest distance)

2. Assign the label of the nearest point



Classifier Example

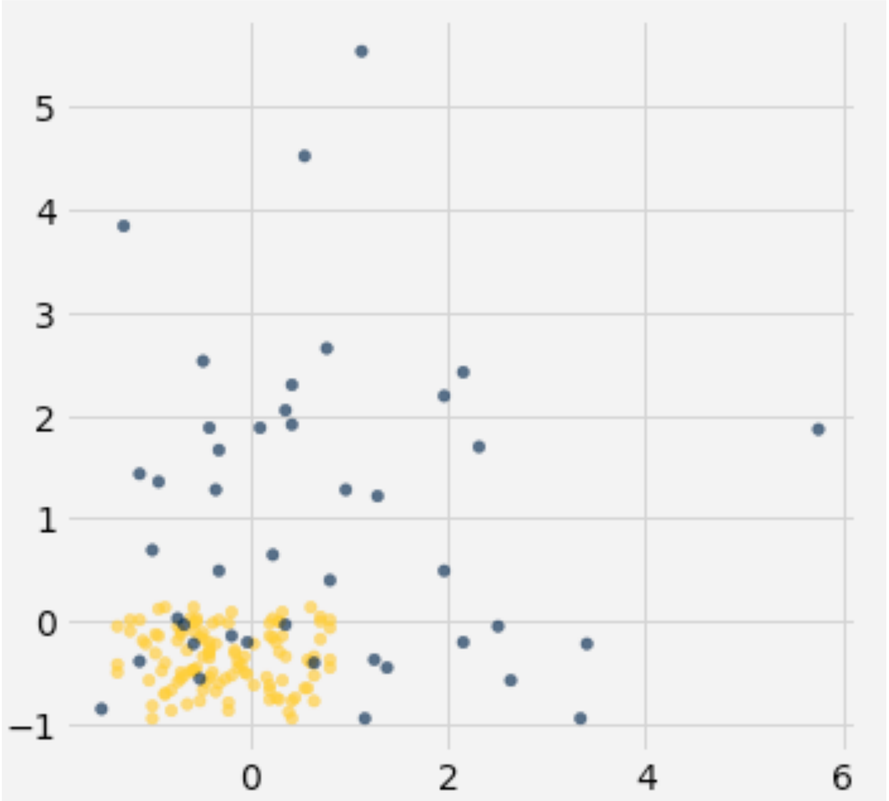
What about for this dataset?



k-Nearest Neighbor Classification

To classify a point:

1. Find the closest k neighbors
2. Find the majority label among those neighbors
3. Assign the majority label to the new point



Evaluation

Accuracy of a classifier

- How do we know our predictions are good?
  - We look at how often our predictions are correct!
- Accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly
  - Though we'd like our predictions to always be correct, it can still be useful even if it doesn't predict correctly 100% of the time

Evaluation Metrics

- Suppose a model performs 95% accuracy on a test set. Is this a good performance?

- Depends!

  - If 90% of the population is Group A, then 95% accuracy isn't that much better than labeling everyone Group A

  - If 50% of the population is Group A, then 95% accuracy is pretty good

- Evaluation depends on contextualizing the performance with the baseline

Evaluation: Confusion Matrix

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positive 😌	False Positive ☹️
Predicted Negative (0)	False Negative ☹️	True Negative 😌

Accuracy of a classifier

- How do we know our predictions are good?
  - We look at how often our predictions are correct!
- Accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly
  - Though we'd like our predictions to always be correct, it can still be useful even if it doesn't predict correctly 100% of the time

Where does this come from?

Generating Training and Testing Data

- The data we use to look for patterns (i.e. build our classifier) we call training data

- The data we use to test how good our classifier is is the testing data

- Let's say you only have a single dataset. How do you choose the testing and training data?

  - What happens if you use the entire dataset for both training and testing? (Hint: What would be the accuracy of using nearest neighbor?)

Generating Training and Testing Data

- Let’s say you only have a single dataset. How do you choose the testing and training data?
  - Create two samples out of the original data set and use one for training and one for testing
  - How to generate these two samples?
    - Select at random!
    - If original sample drawn at random from population and test data drawn at random from sample, we can infer accuracy is similar on population

Next time

- Special topics