

DS 100 – Intro to Data Science

Lecture 23 – Classification

04/24/2025

Adam Poliak



BRYN MAWR
COLLEGE



Announcements

Midsemester feedback form - <https://forms.gle/M4jVdGTDgQknWAwo6>

HW10 (due Friday May 2nd)

Project 3 (due Friday May 2nd)



HW08 Feedback

The Central Limit Theorem says that the probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, *regardless of the distribution of the population from which the sample is drawn.*

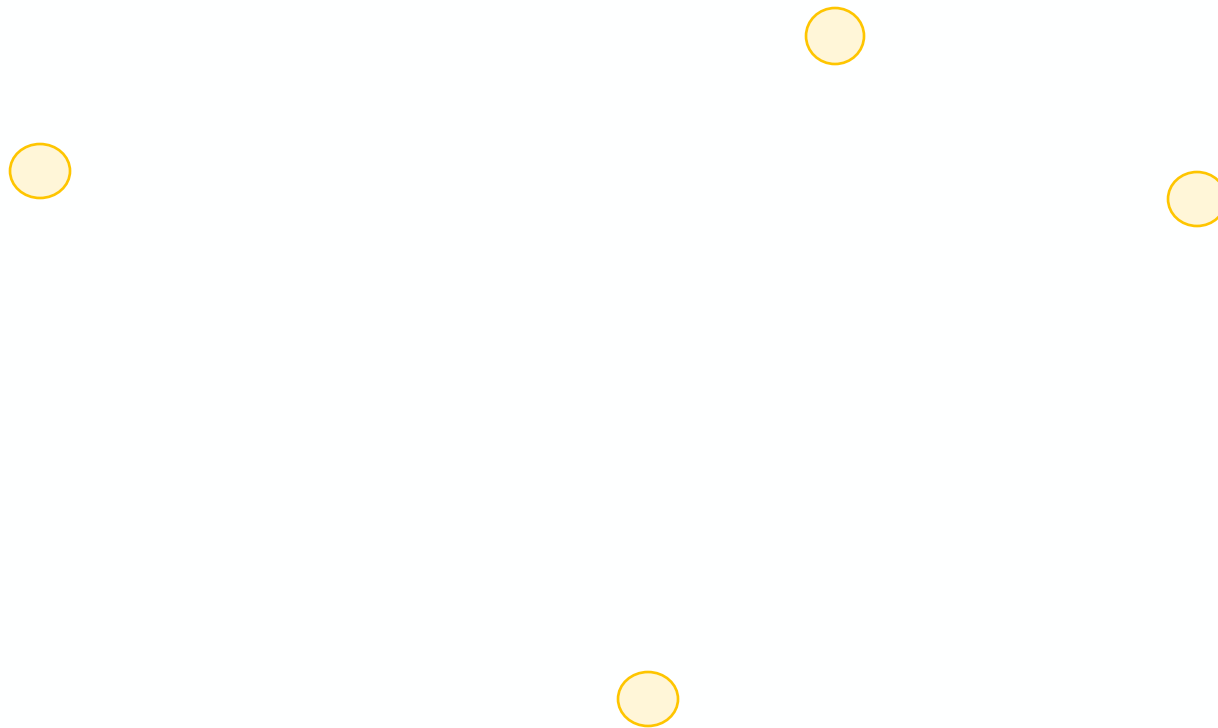
Regression Model



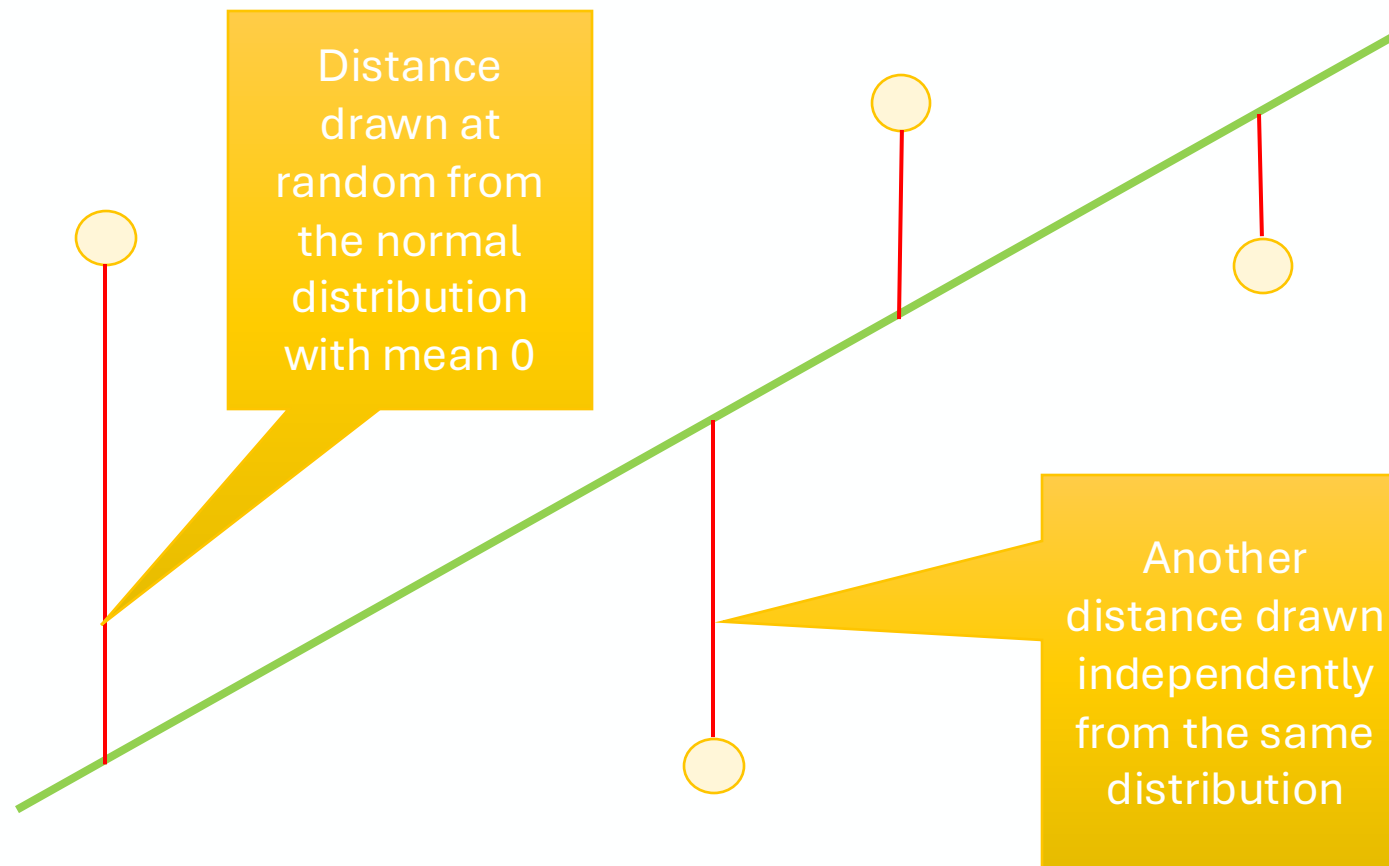
brynmawr.edu
COLLEGE



What we get to see



A “Model”: Signal + Noise



Prediction Variability



brynmawr.edu
COLLEGE



Confidence Interval for Prediction

- **Bootstrap the scatter plot**
- **Get a prediction for y using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the height of the true line at y .

Predictions at Different Values of x

Since y is correlated with x , the predicted values of y depend on the value of x .

The width of the prediction's CI also depends on x .

- Typically, intervals are wider for values of x that are further away from the mean of x .

Inference about the true slope



brynmawr.edu
COLLEGE



Confidence Interval for True Slope

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat the first two steps.
- Draw the empirical histogram of all the generated slopes.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the slope of the true line.



Test Whether There Really is a Slope

Null hypothesis: The slope of the true line is 0.

Alternative hypothesis: No, it's not.

Method:

- Construct a bootstrap confidence interval for the true slope.
- If the interval doesn't contain 0, the data are more consistent with the alternative
- If the interval does contain 0, the data are more consistent with the null

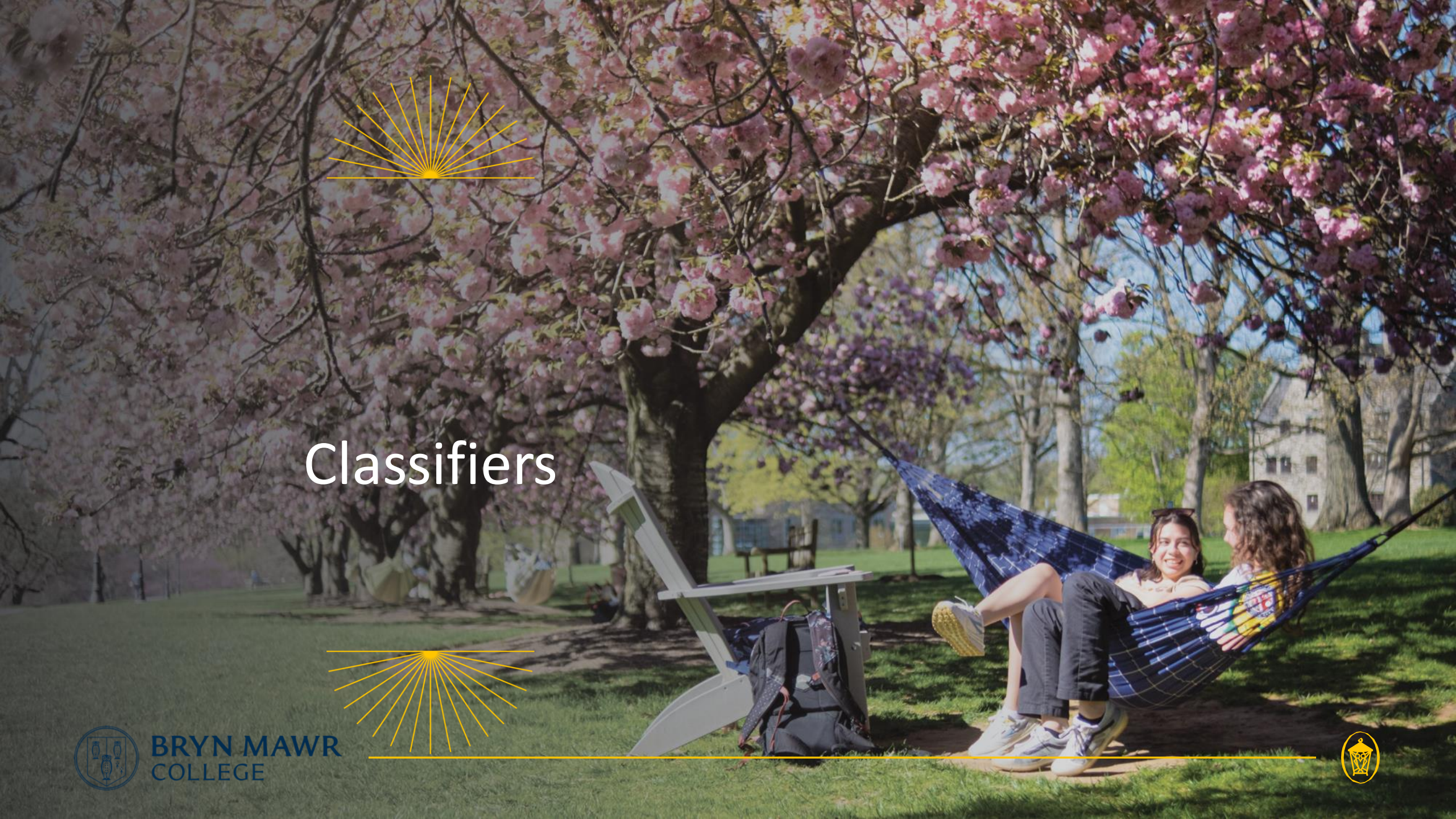


Classification



BRYN MAWR
COLLEGE





Classifiers



BRYN MAWR
COLLEGE



A Classifier





What do rows in a Table
represent?



BRYN MAWR
COLLEGE

brynmawr.edu

Rows of a Table

Each row contains all the data for one individual

t.row(i) evaluates to ith row of table **t**

t.row(i).item(j) is the value of column **j** in row **i**

If all values are numbers, then **np.array(t.row(i))** evaluates to an array of all the numbers in the row.

To consider each row individually, use

for row in t.rows:

... row.item(j) ...

t.exclude(i) evaluates to the table **t** without its ith row

Machine Learning Algorithm

- A mathematical model
- calculated based on sample data ("training data")
- that makes predictions or decisions without being explicitly programmed to perform the task

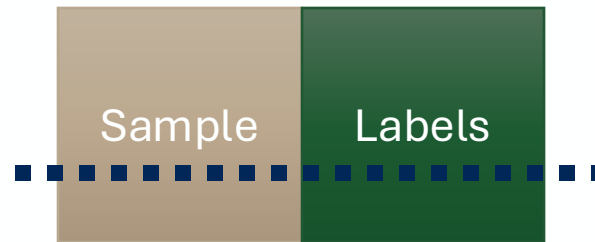


Training and Evaluating a Classifier

Attributes
(features) of
an example



Predicted
label of the
example



Model
association
between
attributes and
labels



Estimate
classifier's
accuracy



Accuracy of a Classifier

The accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly

Need to compare classifier predictions to true labels

If the labeled data set is sampled at random from a population, then we can infer accuracy on that population



Nearest Neighbor Classification



brynmawr.edu
COLLEGE



Nearest Neighbor Classifier

Attributes
(features) of
an example

NN Classifier:
Use the label of
the most similar
training example

Predicted
label of the
example

Population

Sample Labels

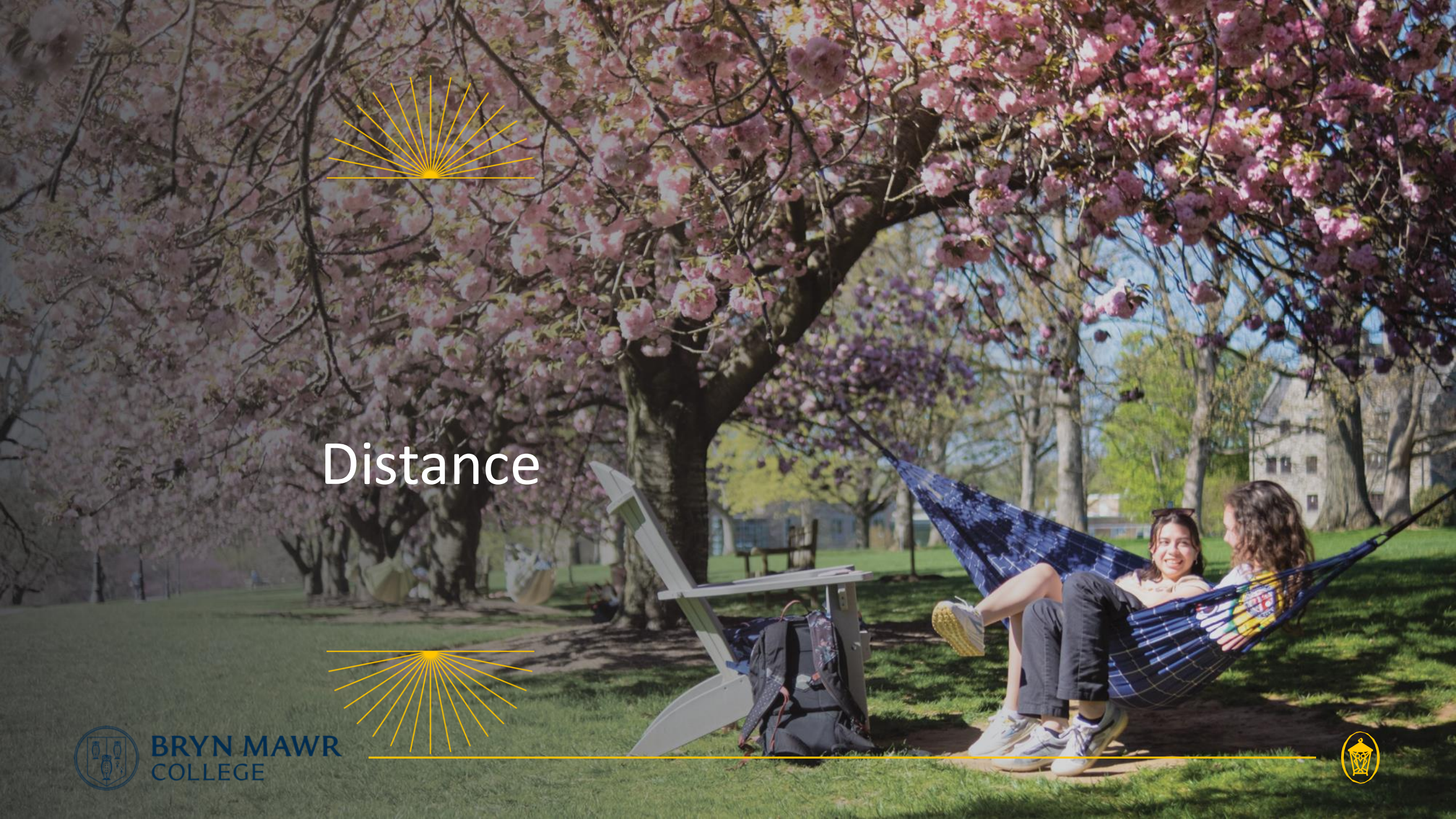
Training

Set

Test

Set





Distance

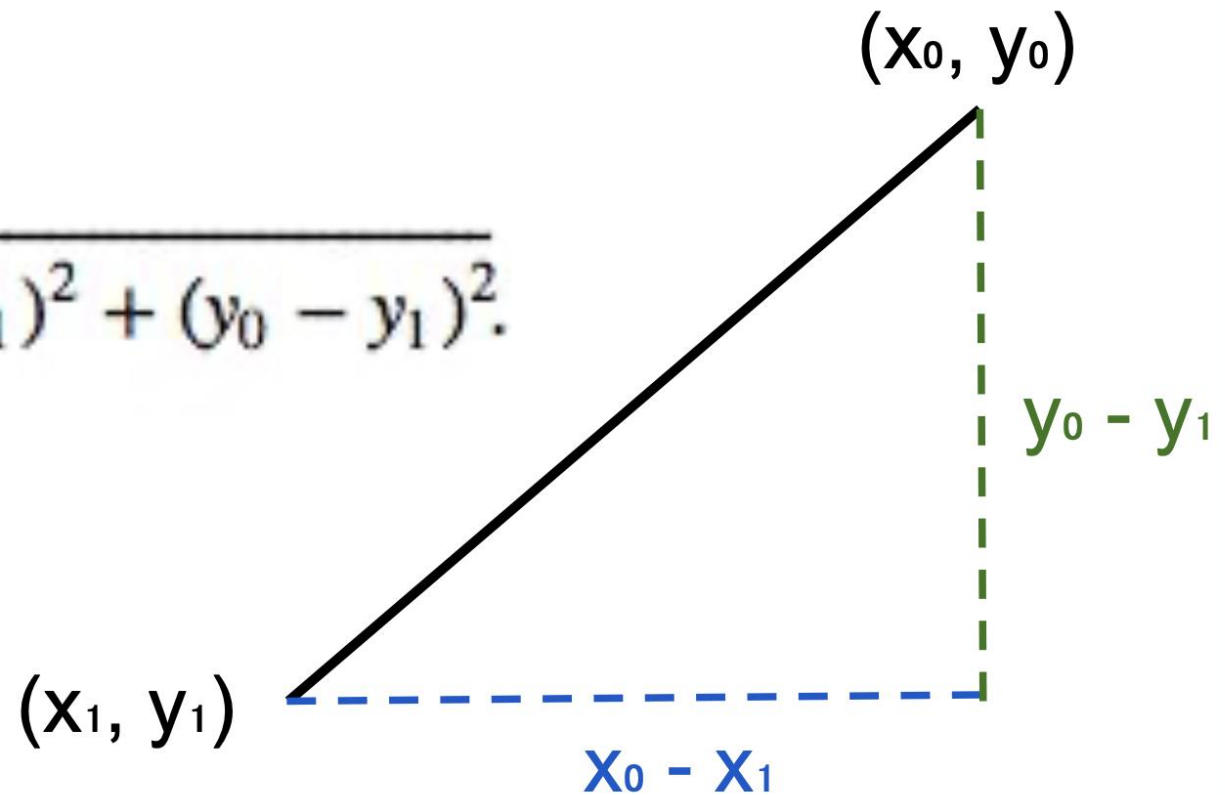


BRYN MAWR
COLLEGE



Pythagore's Formula

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



Distance Between Two Points

Two attributes x and y:

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

Three attributes x, y, and z:

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



Nearest Neighbors Classification



BRYN MAWR
COLLEGE



Finding the k nearest neighbors

1. Find the distance between the example and each example in the training set
2. Augment the training data table with a column containing all the distances
3. Sort the augmented table in increasing order of the distances
4. Take the top k rows of the sorted table



Nearest Neighbor Classifier

Attributes
(features) of
an example

NN Classifier:
Use the label of
the most similar
training example

Predicted
label of the
example

