BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 23 – Residuals & Regression Inference

Announcements

- No lab this week
- Homework 9 - Regression Inference
  - Due Monday 04/25
- Course Evaluations:
  -
- Project 3:
  - Due Monday 05/02

Rubrics

- Rubric 1:
  - Projects (not final project): 45%
  - Homeworks: 25%
  - Participation: 5%
  - Project 3 required
- Rubric 2:
  - Projects (not final project): 30%
  - Homeworks: 35%
  - Participation 10%
  - Project 3 optional

We will compute scores for both Rubrics and then use whichever is best for each student

Linear Regression

Finding the best-fit line

- Compute correlation coefficient *(r)*
  - Prediction in standard units

- Find slope and intercept of the data
  - Prediction in original units
  - slope = r * sd(y) / sd(x)
  - intercept = mean(y) – slope * mean(x)

- Numerical Optimization:
  - Use a compute to find slope and intercept to minimize y

y = slope * x + intercept

Residuals

Residiuals

- Error in regression estimate

- One residual corresponding to each point (*x*, *y*)

- **residual**
  = observed *y* - regression estimate of *y*
  = observed y - height of regression line at *x*
  = vertical distance between the point and line

Regression Diagnostics

Residual Plot

A scatter diagram of residuals

- For linear relations, plotted residuals should look like an unassociated blob
- For non-linear relations, the plot will show patterns
- Used to check whether linear regression is appropriate
- Look for curves, trends, changes in spread, outliers, or any other patterns

Properties of residuals

- The mean of residuals is always 0
- Variance is standard deviation squared
- (Variance of residuals) / (Variance of y) = $1 - r^2$
- (Variance of fitted values) / (Variance of y) = r^2
- Variance of y =
  (Variance of fitted values) + (Variance of residuals)

SD of Fitted (Predicted) Values

- We just said
  - (Variance of fitted values) / (Variance of y) = r^2
  - variance is standard deviations squared,

- So:
  - $\frac{SD\ of\ fitted\ values}{SD\ of\ y} = |r|$
  - $SD\ of\ fitted\ values\ = |r|\ * (SD\ of\ y)$
  - $\frac{Variance\ of\ fitted\ values}{Variance\ of\ y} = r^2$

Variance of Fitted (Predicted) Values

- Variance = Square of the SD
  = Mean Square of the Deviations

- Variance has weird units, but good math properties

- $\frac{Variance\ of\ fitted\ values}{Variance\ of\ y} = r^2$

A Variance Decomposition

By definition,

y = fitted values + residuals

Var(y) = Var(fitted values) + Var(residuals)

A Variance Decomposition

Var(y) = Var(fitted values) + Var(residuals)

- $\frac{Variance\ of\ fitted\ values}{Variace\ of\ y} = r^2$
- $\frac{Variance\ of\ residuals}{Variace\ of\ y} = 1 - r^2$

A Variance Decomposition

Var(y) = Var(fitted values) + Var(residuals)

- $\frac{SD\ of\ fitted\ values}{Variace\ of\ y} = |r|$

- $\frac{SD\ of\ residuals}{Variace\ of\ y} = \sqrt{}(1\ -\ r^2)$

Residual Average and SD

- The average of residuals is always 0
- $\frac{Variance\ of\ residuals}{Variace\ of\ y} = 1 - r^2$
- SD of residuals = SD of y, not $\sqrt{(1 - r^2)}$

Question 1

Midterm: Average 70, SD 10

Final: Average 60, SD 15

$r = 0.6$

The SD of the residuals is _____.

Question 2

Midterm: Average 70, SD 10

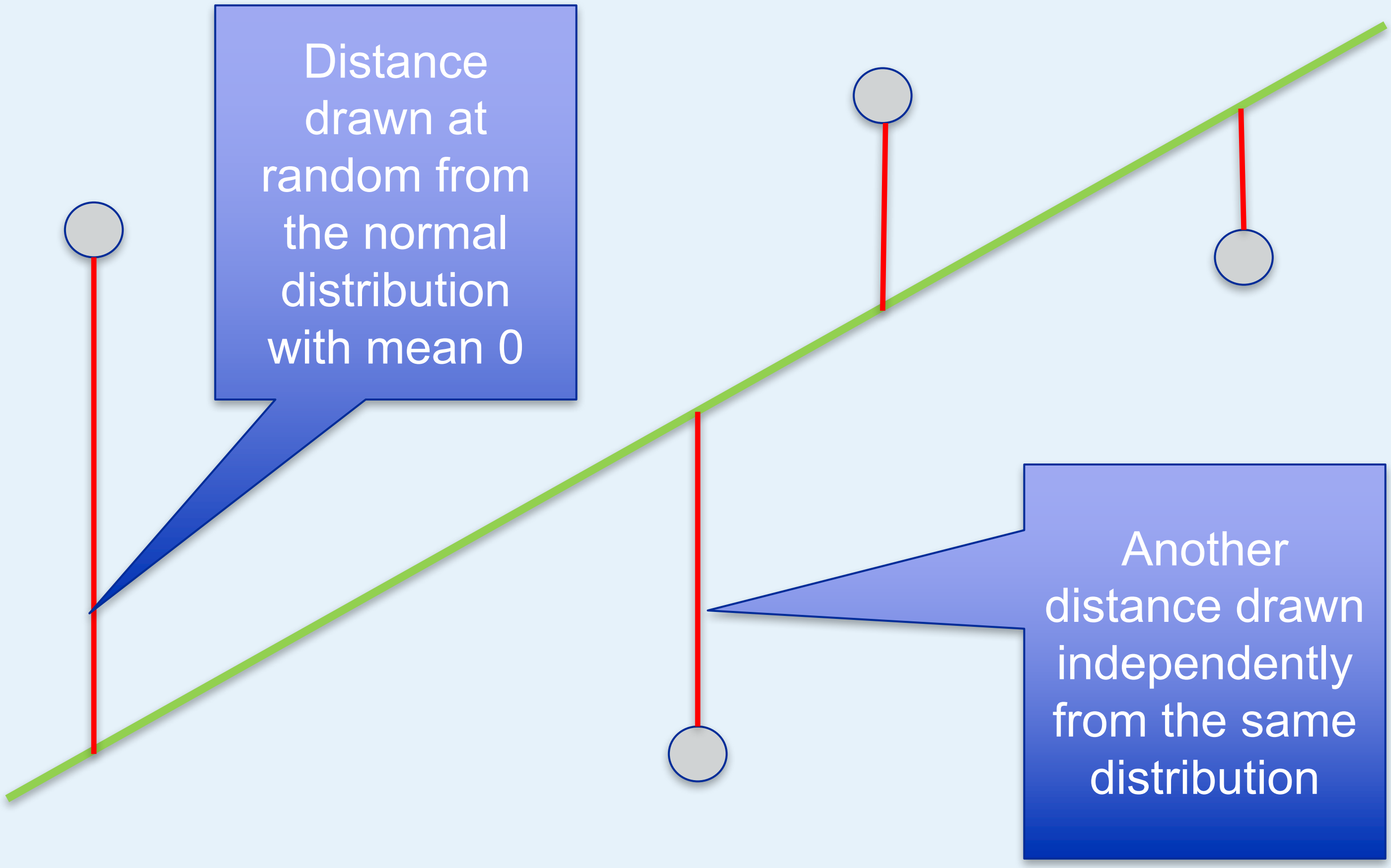
Final: Average 60, SD 15

$r = 0.6$

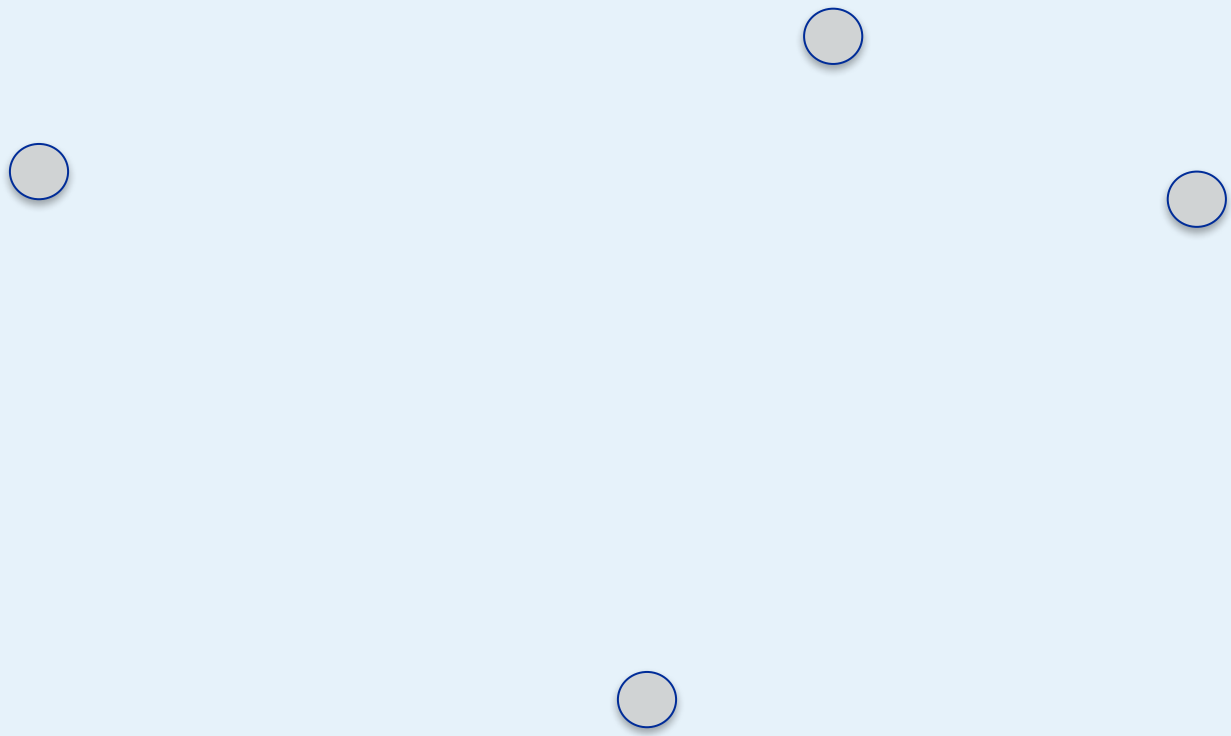
For at least 75% of the students, the regression estimate of final score based on midterm score will be correct to within ______________ points.

Regression Model

A “Model”: Signal + Noise



What we get to see



Prediction Variability

Regression Prediction

- If the data come from the regression model,
- And if the sample is large, then:

- The regression line is close to the true line
- Given a new value of x, predict y by finding the point on the regression line at that x

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

Confidence Interval for True Slope

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat.
- 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

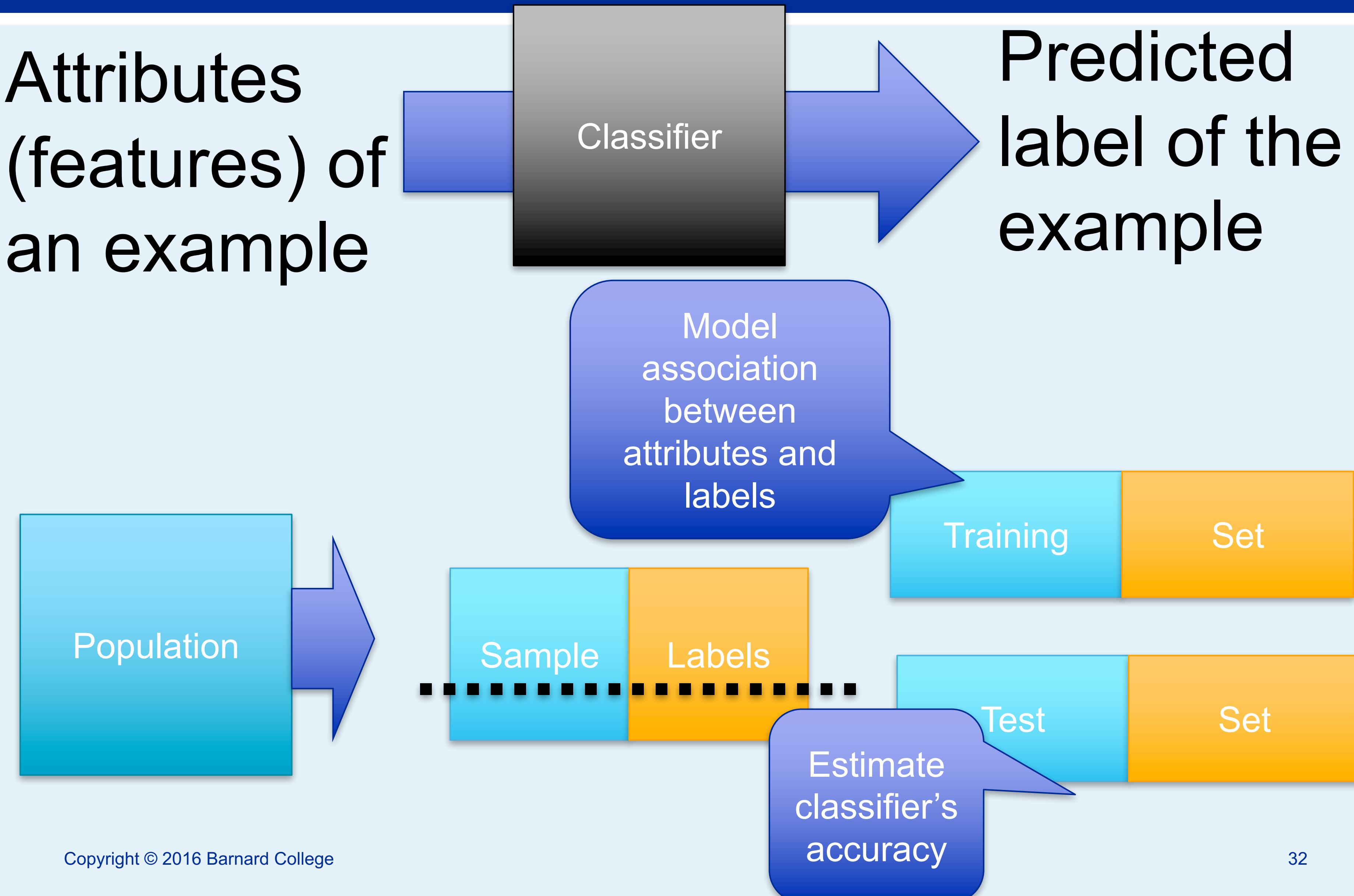
Advanced Regression

- minimize() works no matter what*!
- Define a function that computes the prediction you want, then the error you want, for example:
  - Nonlinear functions of *x*
  - Multiple columns of the table for *x*
  - Other kinds of error instead of RMSE
- Nonlinear functions can get complicated, fast!

Classification

Classifiers

Training a Classifier

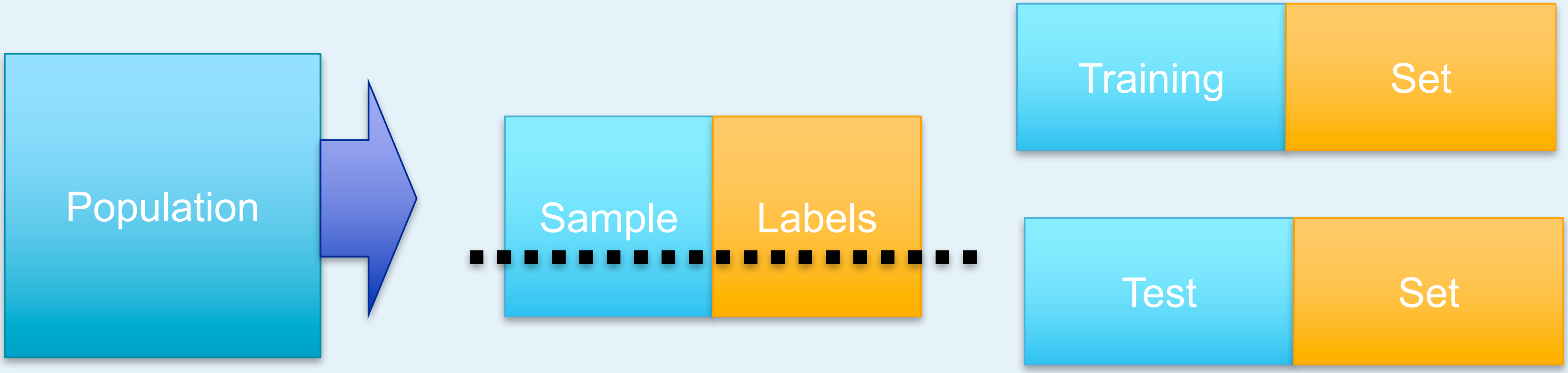


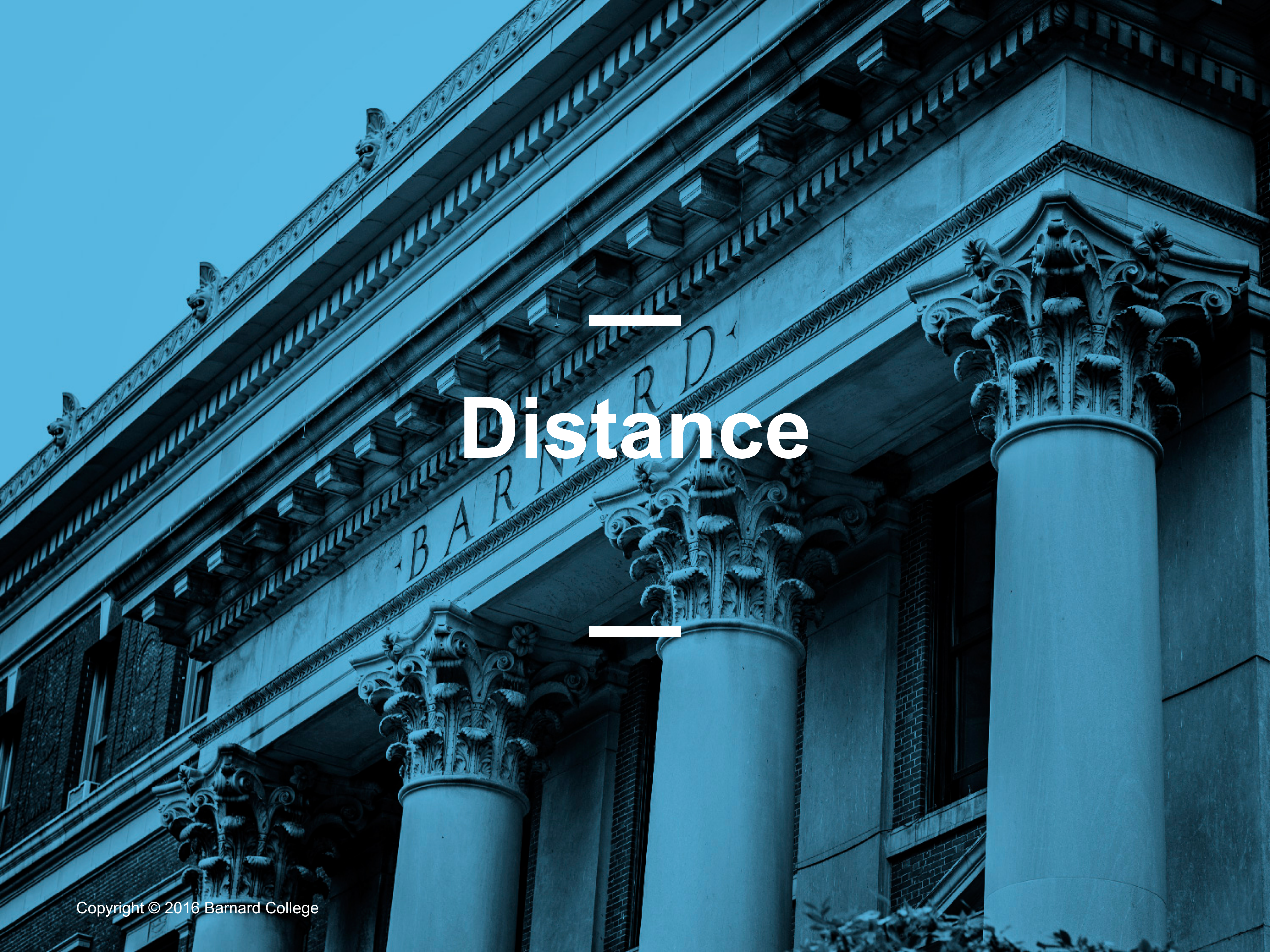
Nearest Neighbor Classifier

Attributes (features) of an example

<u>NN Classifier</u>: Use the label of the most similar training example

Predicted label of the example

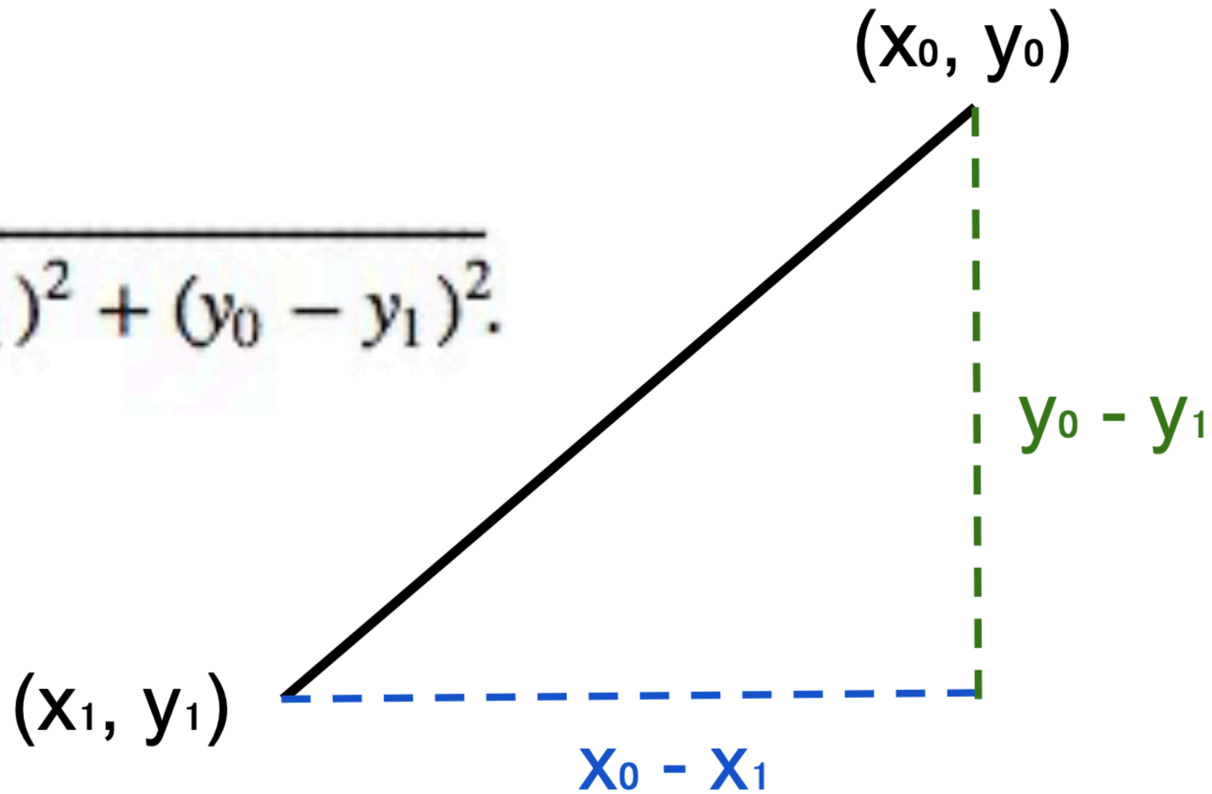




Distance

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