

BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 23 – Classification

BARNARD COLLEGE OF COLUMBIA UNIVERSITY

Announcements



- Homework 9 - Regression Inference
 - Due Thursday 12/10
- Homework 10 – Classification
 - Due Monday 12/14
- Course Evaluations:
 - Due 12/16
- Project 3:
 - Due Monday 12/14



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



Classification

Guessing the Value of an Attribute



- Based on incomplete information
- One way of making 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.
- Two Types of Prediction
 - Classification = Categorical; Regression = Numeric

Prediction Example: Hot dog or not Hot dog?



Prediction Example: Spam or Not Spam?



David, Adam 6

Tennis this week? - in playing tennis on Tuesday. It >>> will b...

Citi Alerts

Your Citibank account statement is available online - com to y...

Humane Rescue Allia.

Your HRA E-Newsletter - Read news and events updates from ...

SLEEP NUMBER

Check out these limited-time Weekend Specials - PLUS get fre...

aishagaddafi11119

Inquiry for Investment. - Inquiry for Investment. Assalamu Alai...

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

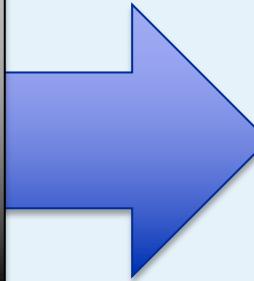


Classifiers

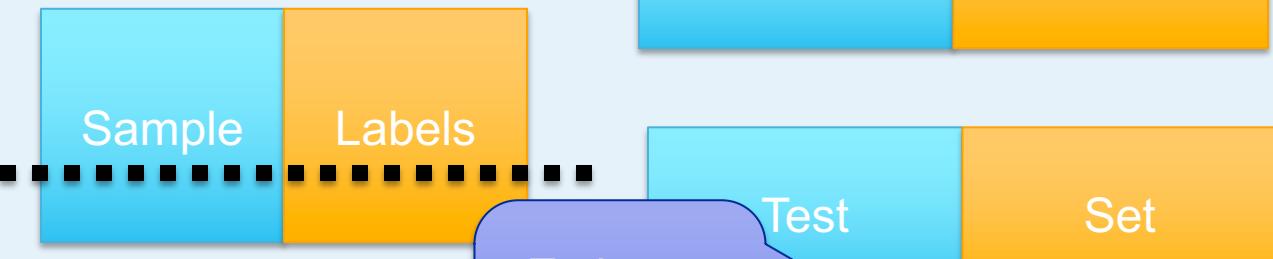


Training a Classifier

Attributes
(features) of
an example



Predicted
label of the
example



Model
association
between
attributes and
labels



Estimate
classifier's
accuracy

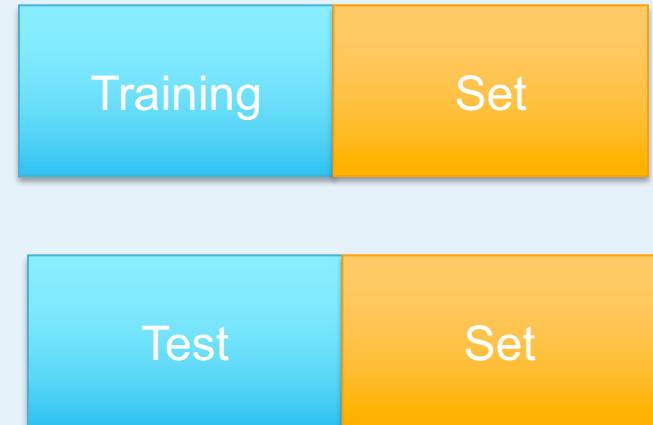
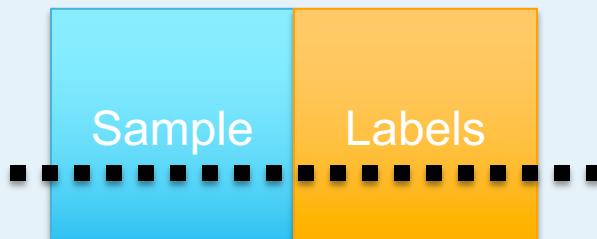
Nearest Neighbor Classifier

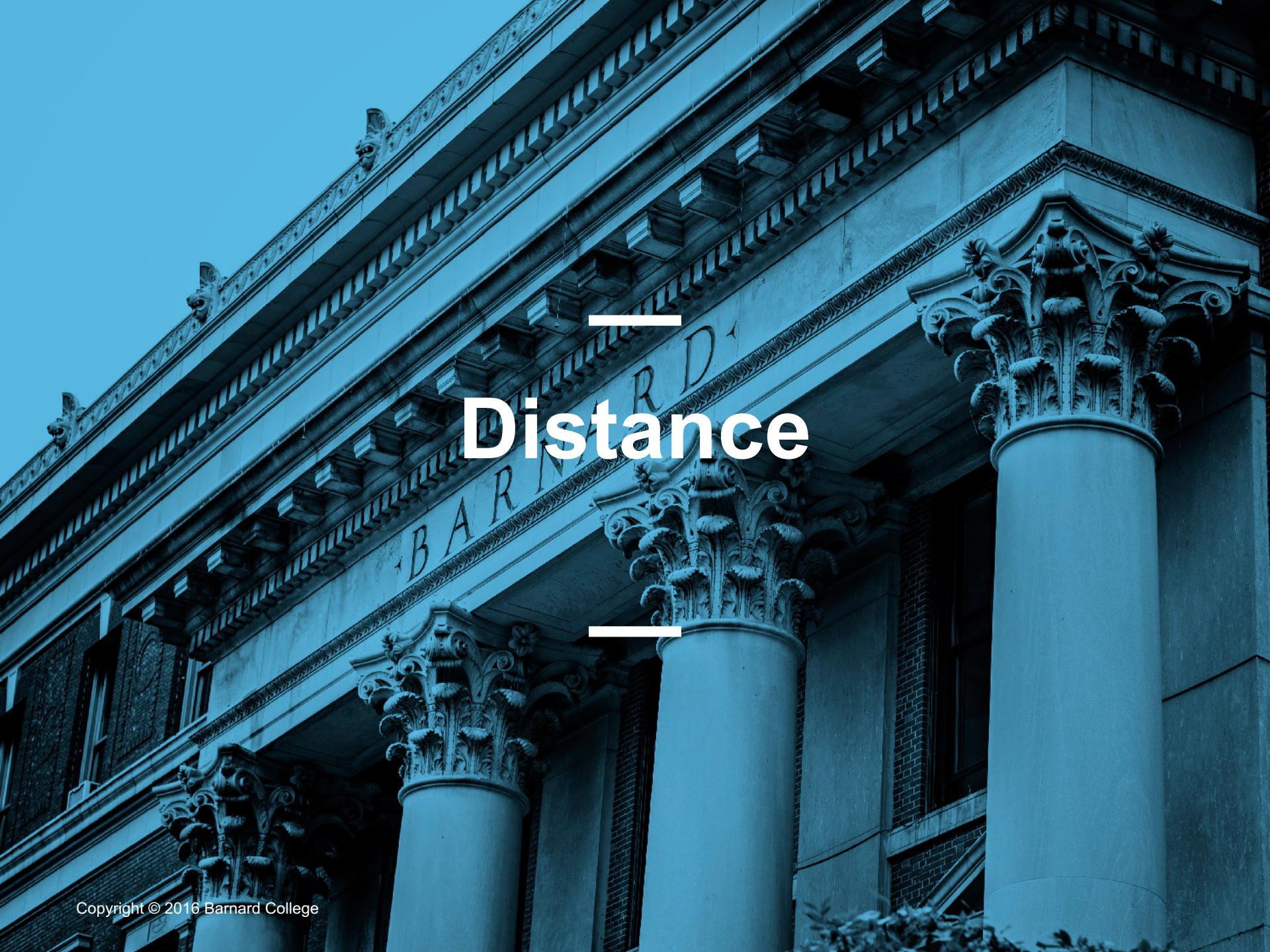


Attributes
(features) of
an example

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

Predicted
label of the
example



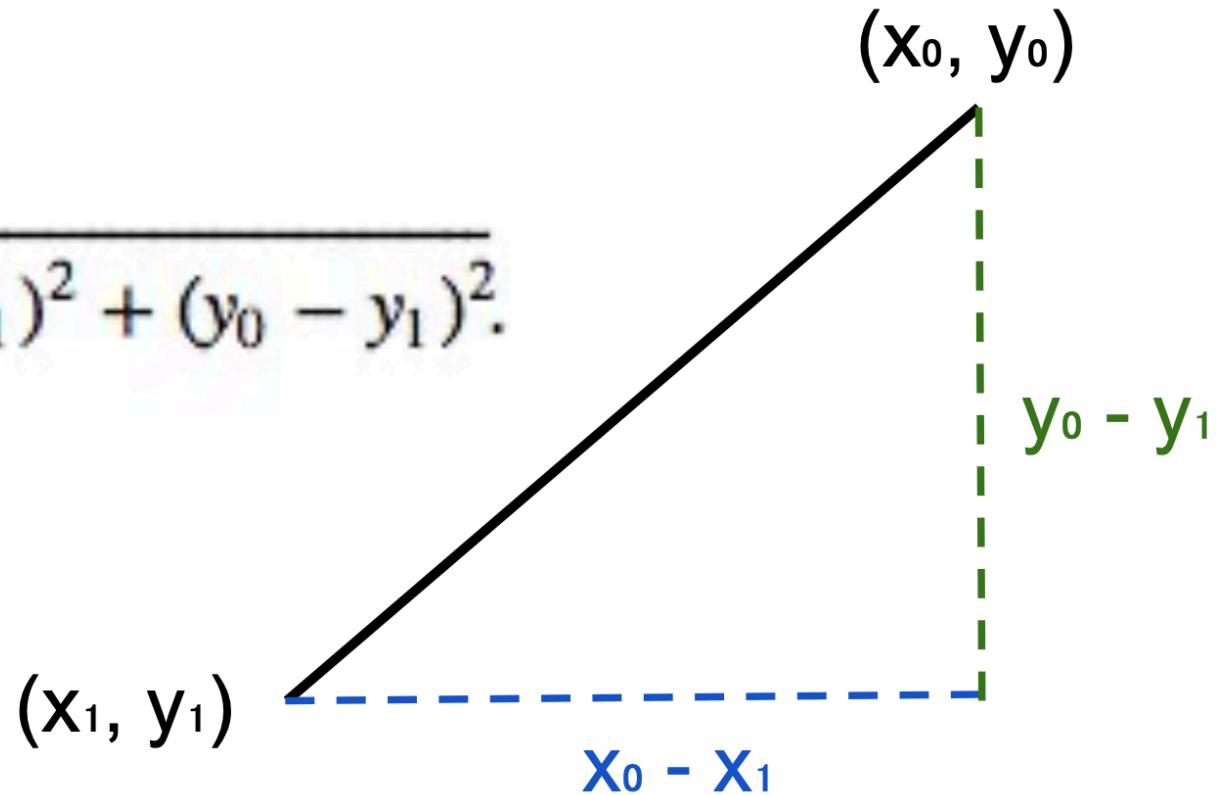


Distance



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:

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