

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)

Final 05/07 9:30am Old Library 116

Outline

Review Classification

Evaluation Metrics

Comparing two classifiers



Classification



BRYN MAWR
COLLEGE





Classifiers



BRYN MAWR
COLLEGE



A Classifier

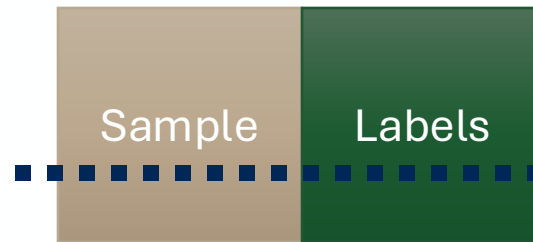


Training and Evaluating a Classifier

Attributes
(features) of
an example



Predicted
label of the
example



Model
association
between
attributes and
labels



Estimate
classifier's
accuracy



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

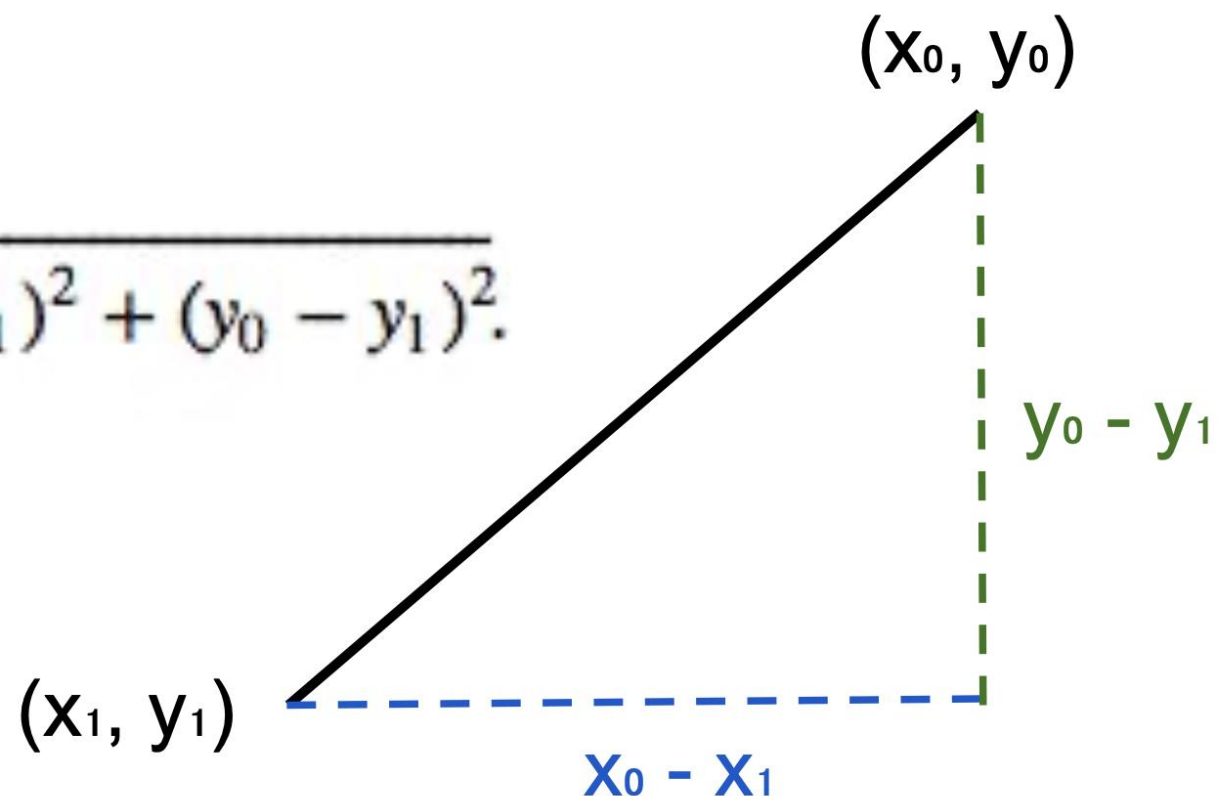


BRYN MAWR
COLLEGE



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

Population

Sample Labels

Training

Set

Test

Set





Evaluation



BRYN MAWR
COLLEGE



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



Classify a tweet as viral or not



Taylor Swift  @taylorswift13 · Jan 27



The Lavender Haze video is out now. There is lots of lavender. There is lots of haze. There is my incredible costar [@laith_ashley](#) who I absolutely adored working with.



7,985



104.6K



435.1K



18.2M



BRYN MAWR
COLLEGE



Accuracy

- Model A performs 60% accuracy, would you say this is good, decent, or awful?
- Model A performs 80% accuracy, would you say this is good, decent, or awful
- Model A performs 98% accuracy, would you say this is good decent or awful?

Evaluation: Accuracy

- Imagine we saw 1 million tweets
 - 100 of them were viral
 - 999,900 were not
- We could build a dumb classifier that just labels every tweet "not viral"
 - It would get 99.99% accuracy!!! Wow!!!!
 - But useless! Cant find the viral tweets!
- When should we not we use **accuracy** as our metric?
 - When data isn't balanced across labels/classes



The 2-by-2 confusion matrix

true positive	false positive
false negative	true negative

The 2-by-2 confusion matrix

		<i>gold standard labels</i>	
		gold positive	gold negative
<i>system output labels</i>	system positive	true positive	false positive
	system negative	false negative	true negative

The 2-by-2 confusion matrix

		<i>gold standard labels</i>		
		gold positive	gold negative	
<i>system output labels</i>	system positive	true positive	false positive	precision = $\frac{tp}{tp+fp}$
	system negative	false negative	true negative	
		recall = $\frac{tp}{tp+fn}$		accuracy = $\frac{tp+tn}{tp+fp+tn+fn}$

Evaluation: Precision

- % of items the system detected (i.e., items the system labeled as positive) that are in fact positive (according to the human gold labels)

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Evaluation: Recall

- % of items actually present in the input that were correctly identified by the system.

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Why Precision and recall

- Our dumb viral-classifier
 - label no tweets as "viral"

Accuracy=99.99%

but

Recall = 0

- (it doesn't get any of the 100 viral tweets)

Precision and recall, unlike accuracy, emphasize true positives:

- finding the things that we are supposed to be looking for.

A combined measure: F

- F measure: a single number that combines P and R:

$$F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- We almost always use balanced F_1 (i.e., $\beta = 1$)

$$F_1 = \frac{2PR}{P + R}$$



Comparing Models

- Model A performs 60% accuracy, Model B performs 60.5% accuracy
- Which is better?