



DATA 1202

Spring 2024

Lecture 26

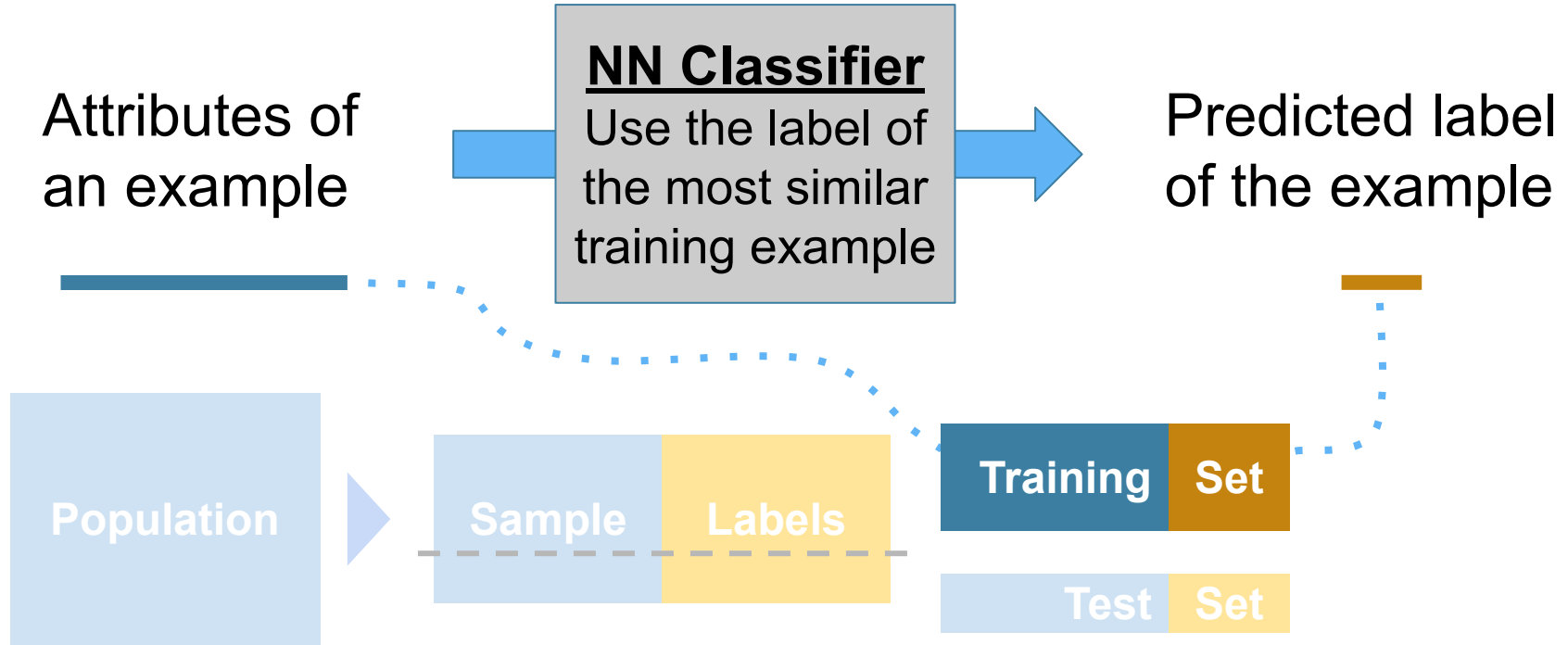
Classifiers

Announcements

- **Homework 13** due tonight
- **Project 2** checkpoint is due Friday
- **Final Exam**
 - Wed. @ 3pm

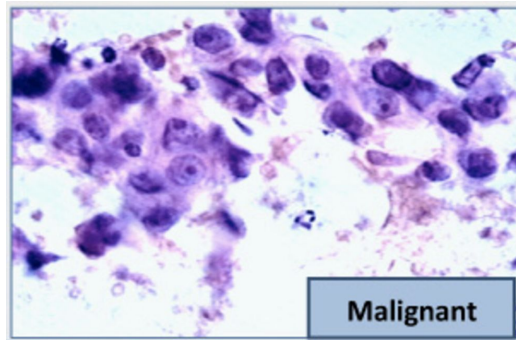
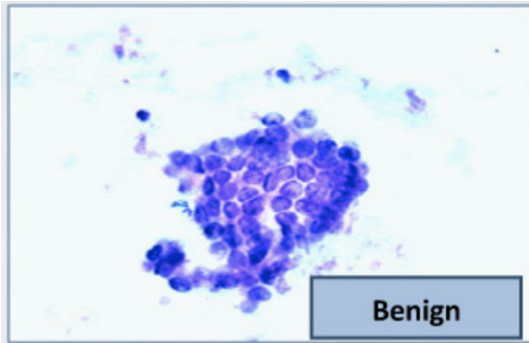
Classifiers

Nearest Neighbor Classifier



The Google Science Fair

- Brittany Wenger, a 17-year-old high school student
- Won by building a breast cancer classifier with 99% accuracy

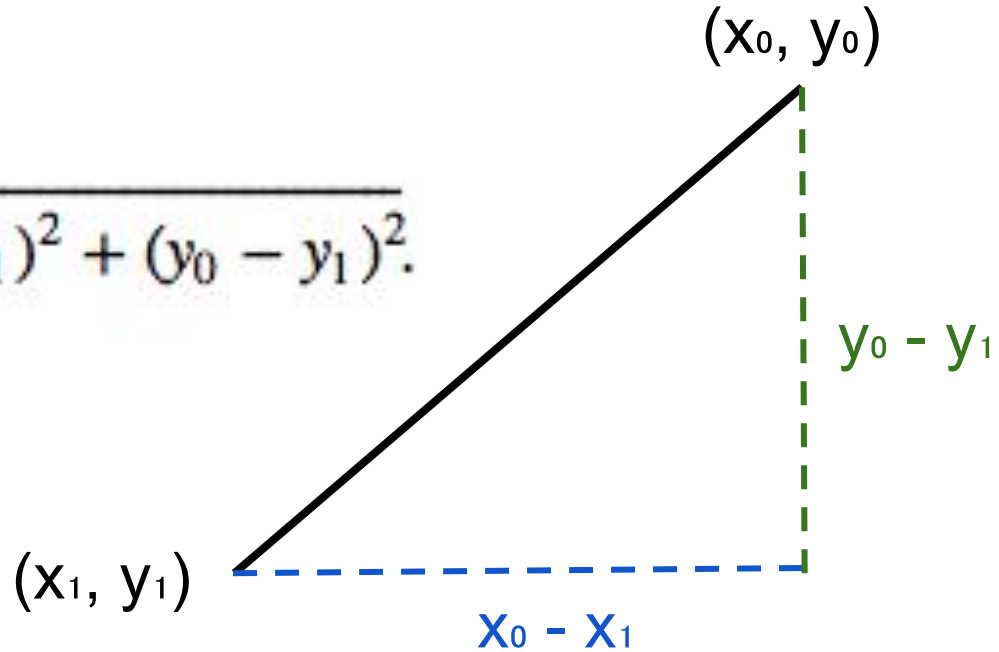


(Demo)

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 :

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

- and so on ...
-

Rows

Rows of Tables

Each row contains all the data for one individual

- `t.row(i)` evaluates to *i*th 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:                                     (Demo)
    ... row.item(j) ...
```

- `t.exclude(i)` evaluates to the table `t` without its *i*th row
-

Nearest Neighbors

Finding the k Nearest Neighbors

To find the k nearest neighbors of an example:

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

The Classifier

To classify a point:

- Find its k nearest neighbors
- Take a majority vote of the k nearest neighbors to see which of the two classes appears more often
- Assign the point to the class that wins the majority vote

(Demo)

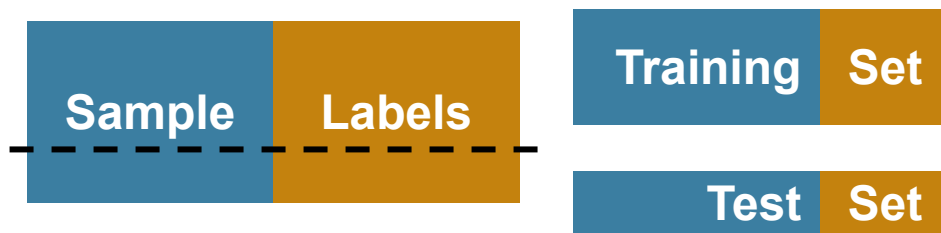
Evaluation

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

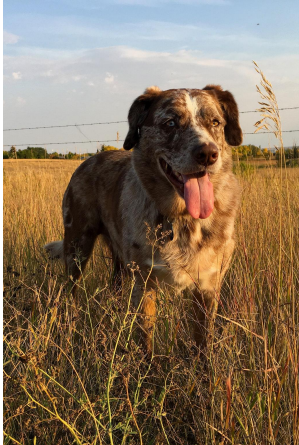
If the labeled data set is sampled at random from a population, then we can infer accuracy on that population



(Demo)

Before Classifying

Dog or Wolf?



Start with a Representative Sample

- Both the training and test sets must accurately represent the population on which you use your classifier
 - **Overfitting** happens when a classifier does very well on the training set, but can't do as well on the test set
-

Standardize if Necessary

Chronic Kidney
Disease data set

Glucose	Hemoglobin	White Blood Cell Count	Class
117	11.2	6700	1
70	9.5	12100	1
380	10.8	4500	1
157	5.6	11000	1

- If the attributes are on very different numerical scales, distance can be affected
- In such a situation, it is a good idea to convert all the variables to standard units

(Demo)
