YData: An Introduction to Data Science

Lecture 34: Classification

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Announcements

- Project 3 checkpoint today; full project due Friday 4/30
- Lowest project score is dropped
- Assignment 11 posted today; due Thursday 5/6

You are here

Intro, Cause and Effect	Lectures 1–2
Python, Tables, Visualization	Lectures 3–13
Probability and Distributions	Lectures 14–17
Hypothesis Testing and Causality	Lectures 18–20
Midterm exam	—
Confidence and the Normal Distribution	Lectures 23–28
Regression and Classification	Lectures 29–37
Final exam	—

Remaining topics

- Classification
- Multiple regression
- Decisions and Bayes rule
- Pandas primer
- Review

Recap: Regression Inference

Confidence Interval for True Slope

- **Bootstrap the scatter plot**
- **Find the slope of the regression line through the bootstrapped plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the slopes
- The middle 95% interval is an approximate 95% confidence interval for the slope of the true line

(DEMO)

Classification

Classification tasks

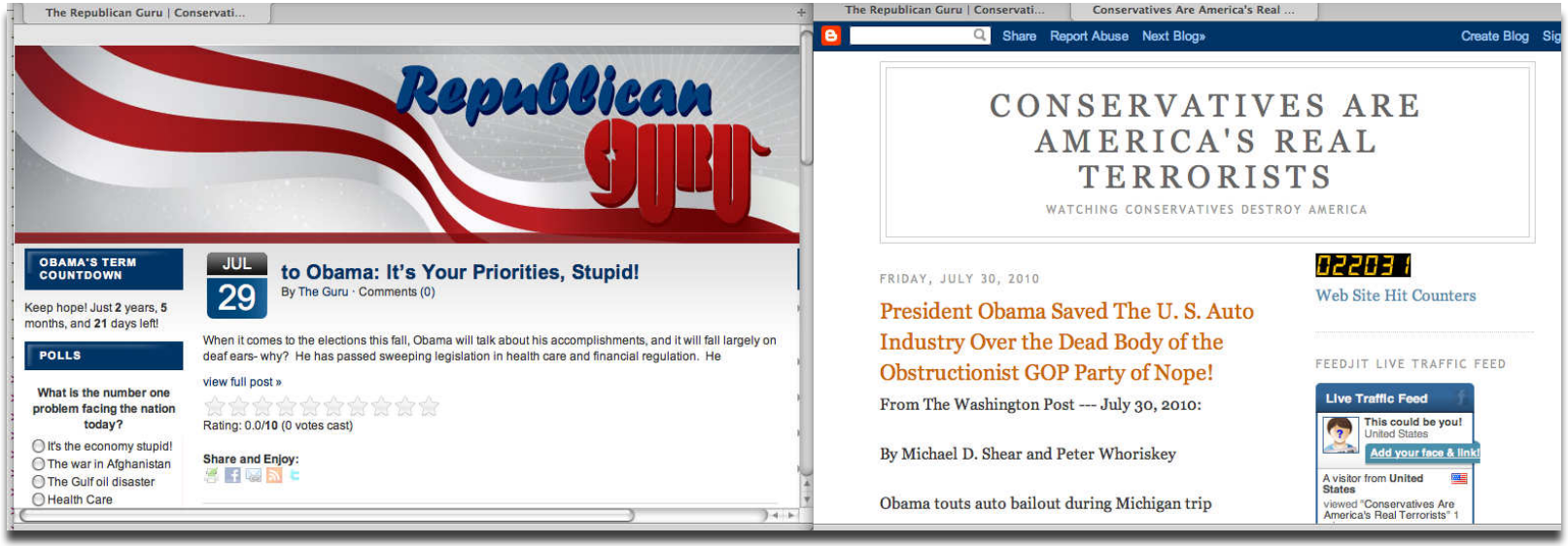
- The Coronary Risk-Factor Study (CORIS). Data: 462 males between ages of 15 and 64 from three rural areas in South Africa.

  Outcome Y is presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease

  9 predictor variables: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age.

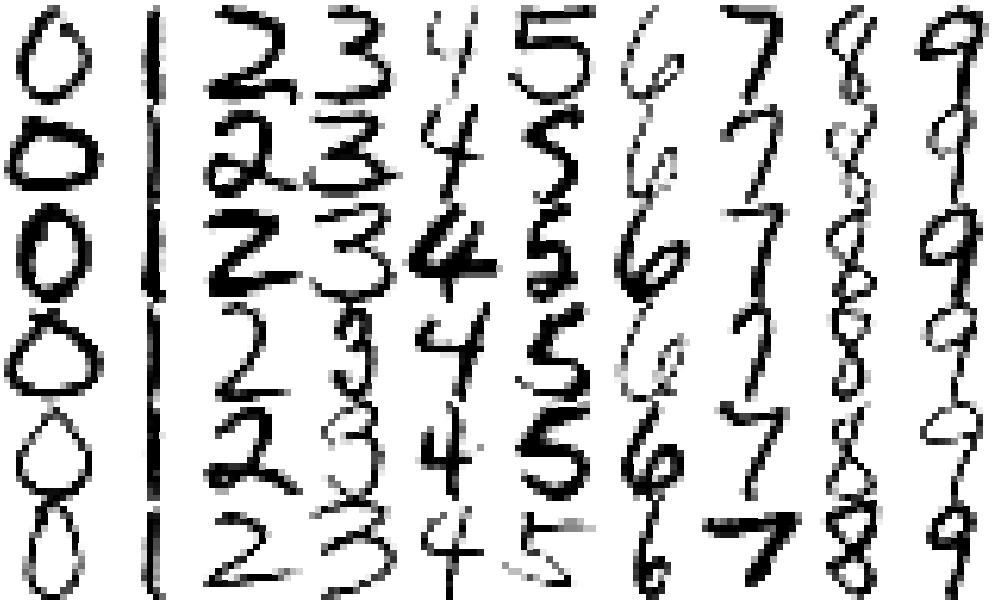
Classification tasks

- Political Blog Classification. A collection of 403 political blogs were collected during two months before the 2004 presidential election. The goal is to predict whether a blog is *liberal* ($Y = 0$) or *conservative* ($Y = 1$) given the content of the blog.



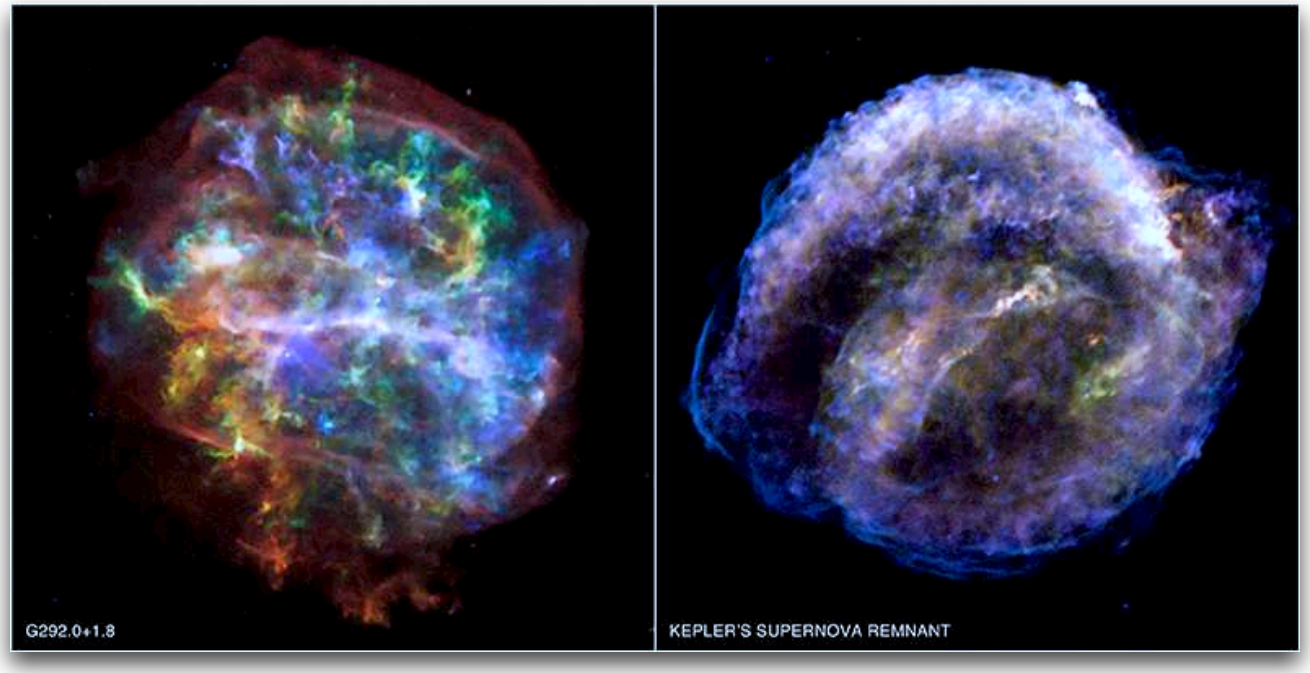
Classification tasks

- Handwriting Digit Recognition. Here each Y is one of the ten digits from 0 to 9. There are 256 input variables $X_1, \ldots, X_{256}$ corresponding to the intensity values of the pixels in a 16×16 image.



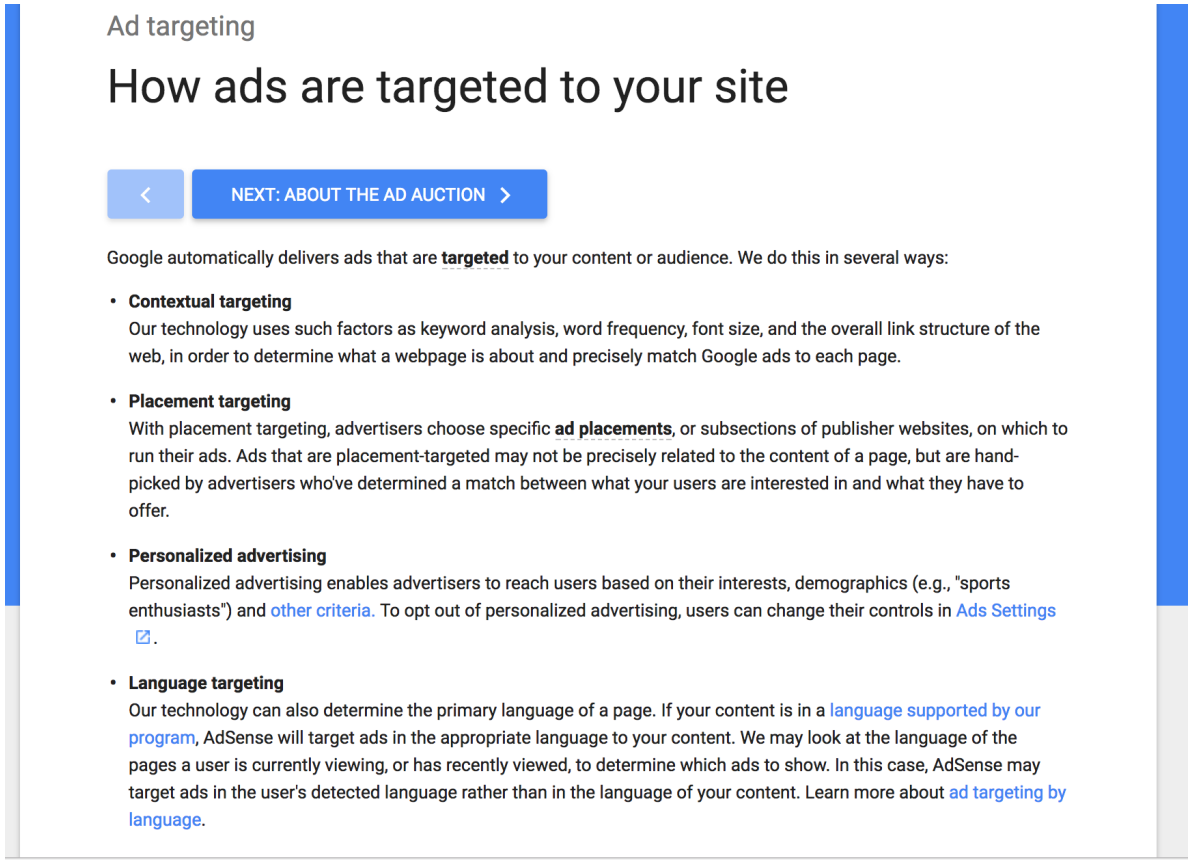
Classification tasks

- A supernova is an exploding star. Type Ia supernovae are a special class of supernovae that are very useful in astrophysics research. These supernovae have a characteristic *light curve*, which is a plot of the luminosity of the supernova versus time.



Classification tasks

- Ad click-through prediction. Predict whether or not a user will click on an ad presented. Used for ranking ads and setting prices.



Ad targeting

How ads are targeted to your site

NEXT: ABOUT THE AD AUCTION

Google automatically delivers ads that are **targeted** to your content or audience. We do this in several ways:

- **Contextual targeting**
  Our technology uses such factors as keyword analysis, word frequency, font size, and the overall link structure of the web, in order to determine what a webpage is about and precisely match Google ads to each page.
- **Placement targeting**
  With placement targeting, advertisers choose specific **ad placements**, or subsections of publisher websites, on which to run their ads. Ads that are placement-targeted may not be precisely related to the content of a page, but are hand-picked by advertisers who've determined a match between what your users are interested in and what they have to offer.
- **Personalized advertising**
  Personalized advertising enables advertisers to reach users based on their interests, demographics (e.g., "sports enthusiasts") and other criteria. To opt out of personalized advertising, users can change their controls in Ads Settings.
- **Language targeting**
  Our technology can also determine the primary language of a page. If your content is in a language supported by our program, AdSense will target ads in the appropriate language to your content. We may look at the language of the pages a user is currently viewing, or has recently viewed, to determine which ads to show. In this case, AdSense may target ads in the user's detected language rather than in the language of your content. Learn more about ad targeting by language.

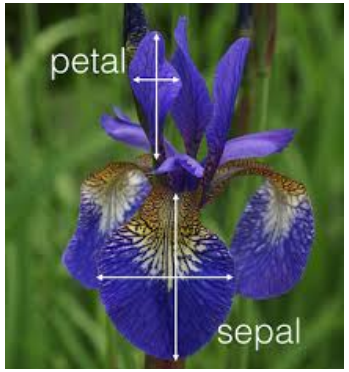
Classification tasks

- The Iris Flower study. The data are 50 samples from each of three species of Iris flowers, *Iris setosa*, *Iris virginica* and *Iris versicolor* The length and width of the sepal and petal are measured for each specimen, and the task is to predict the species of a new Iris flower based on these features.
- App for wildflowers

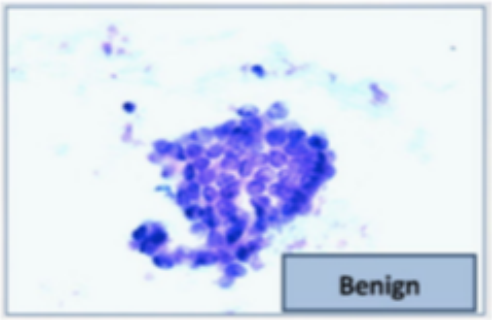
Iris setosa (Left), *Iris versicolor* (Middle), and *Iris virginica* (Right).

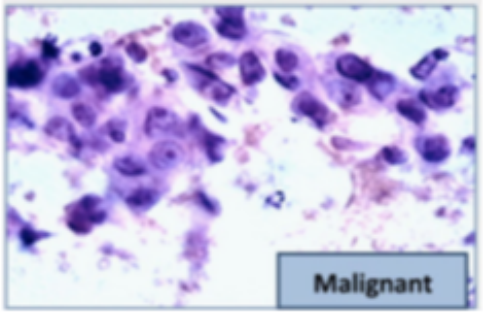
Fisher's iris classification

Iris setosa (Left), *Iris versicolor* (Middle), and *Iris virginica* (Right).



The Google Science Fair





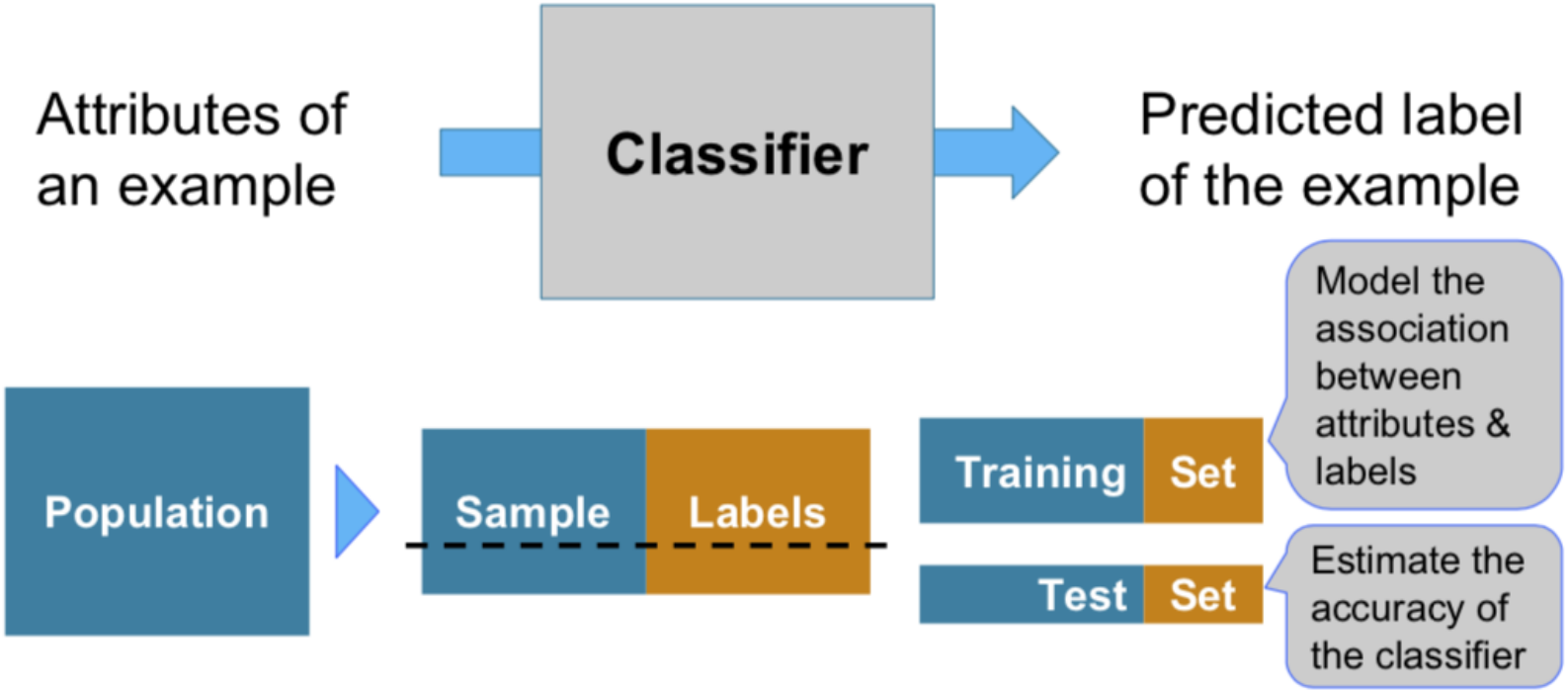
- Brittany Wenger, a 17-year-old high school student in 2012
- Won by building a breast cancer classifier with 99% accuracy

Important concepts

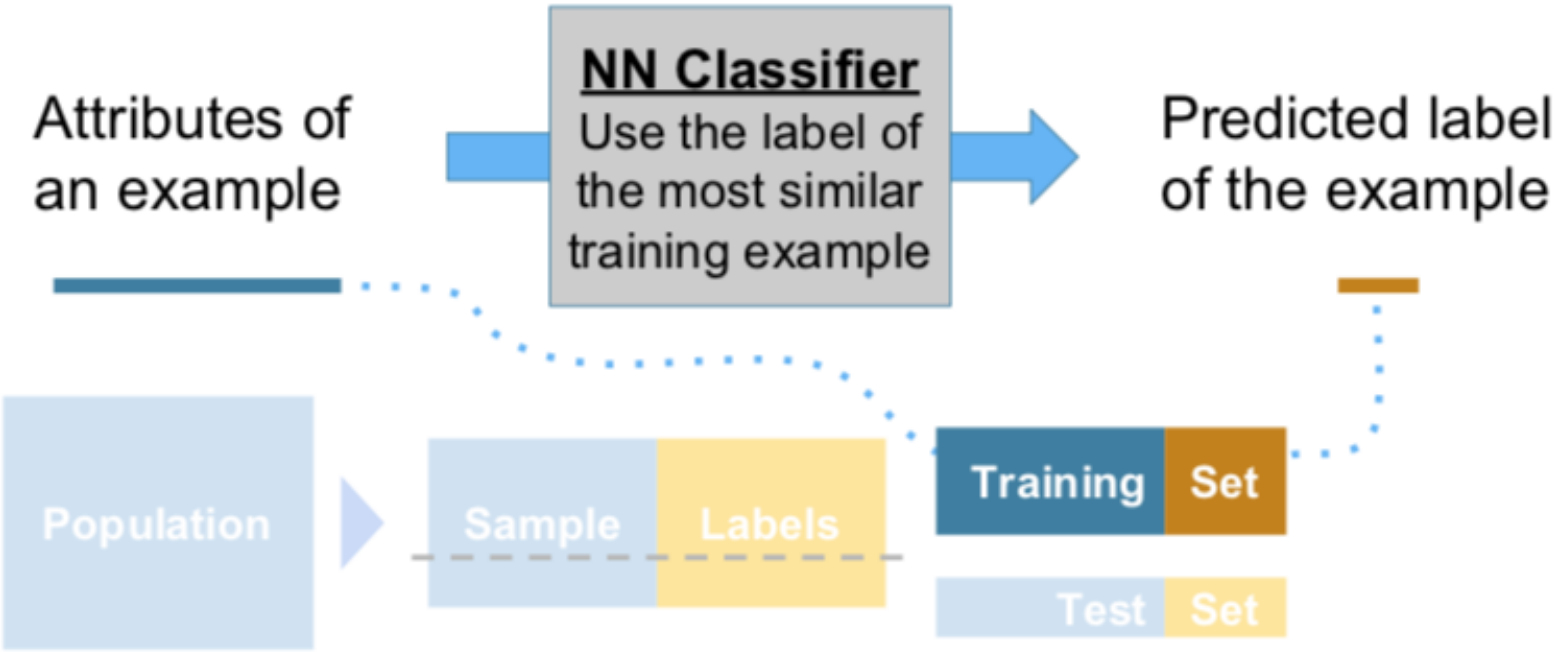
- A binary classifier is a function from the set of inputs to $\{0, 1\}$.
- It is *linear* if we can draw a straight line (or a multi-dimensional plane) between the two predicted values
- The *training error* is the fraction of errors on the training data

(DEMO)

Training a Classifier



Nearest Neighbor Classifier



Distance

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:
...  row.item(j) ...
```

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 ...
- It's important the variables are standardized

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 the class that wins the majority vote

(Demo next class)

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