YData: An Introduction to Data Science

Lecture 35: Classifiers

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

Credit: data8.org

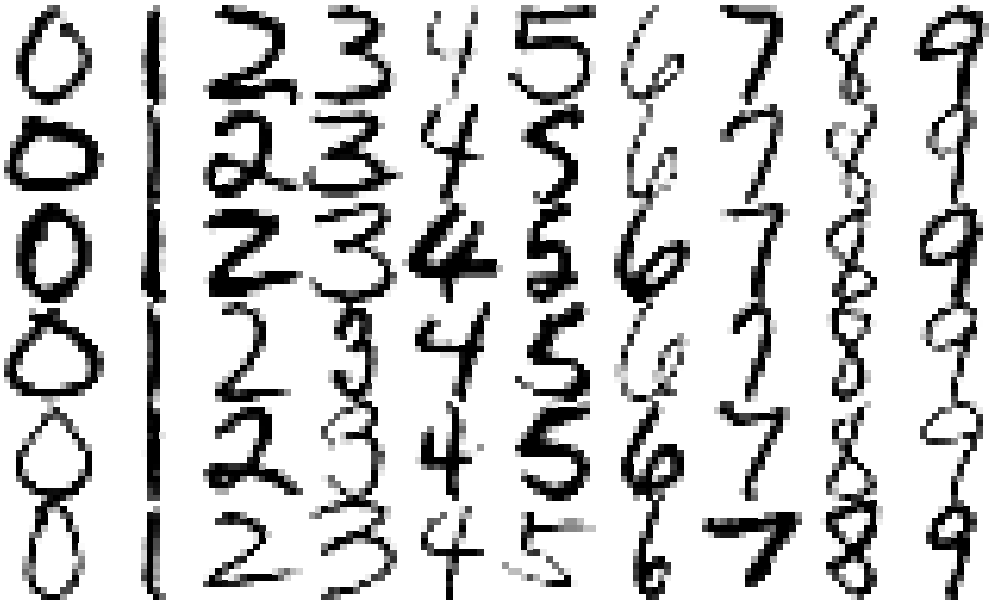
Announcements

- Project 3 due Friday 4/30
- Assignment 11 out; due Thursday 5/6
- Late assignments may be posted to Canvas; disregard notifications
- We'll have info on prep for the final exam next week

Classification

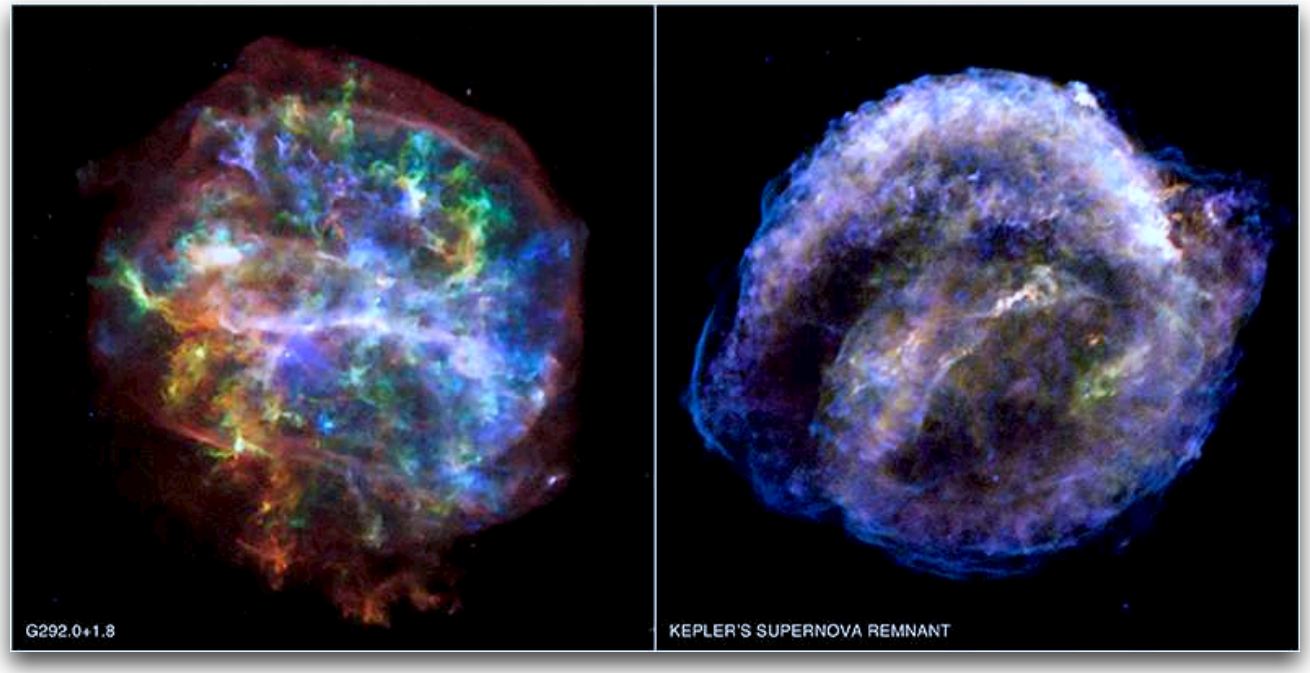
Classification tasks

- 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.



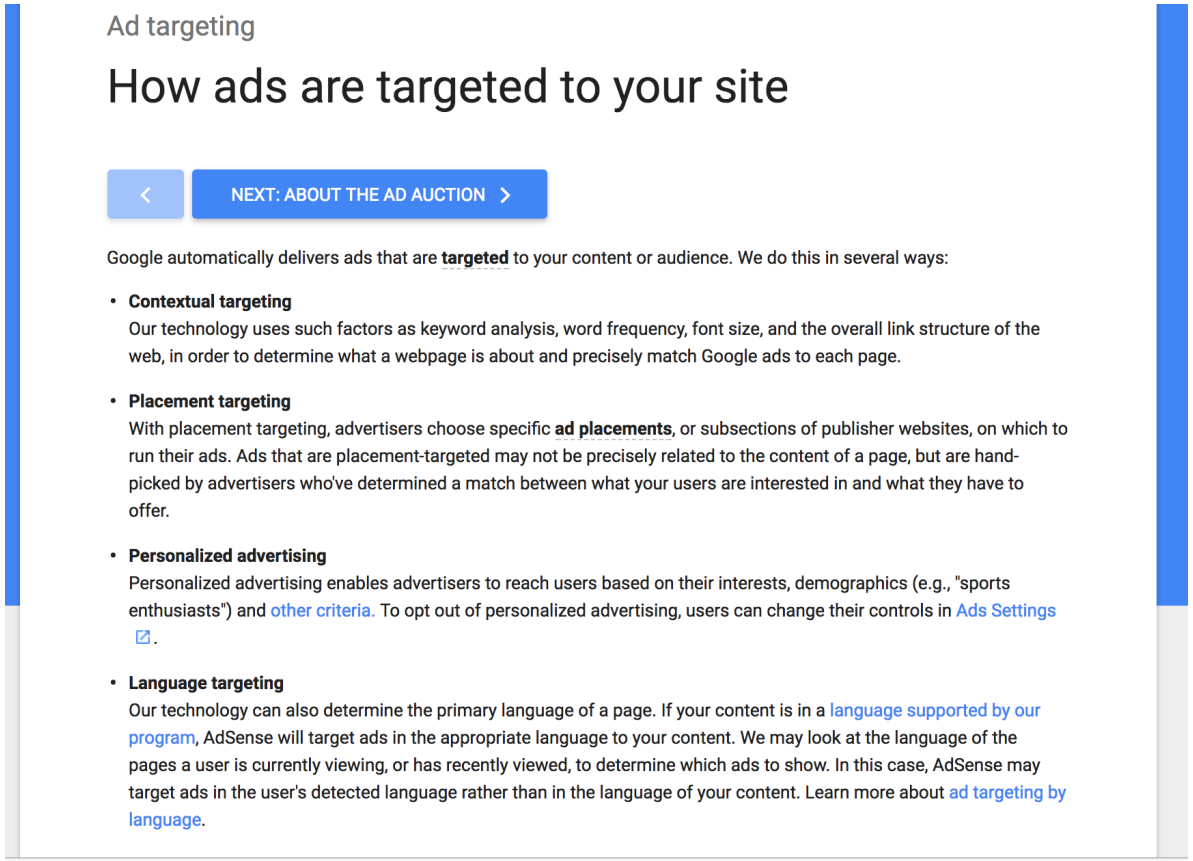
Classification tasks

- 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).

Radiology

Stanford ML Group

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar*, Jeremy Irvin*, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng

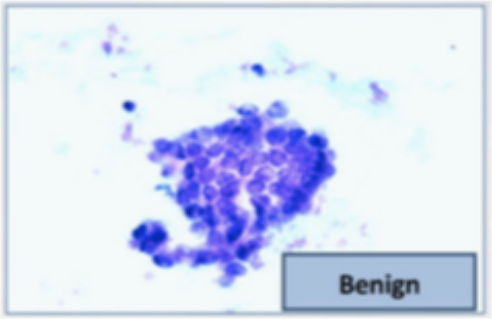
We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists.

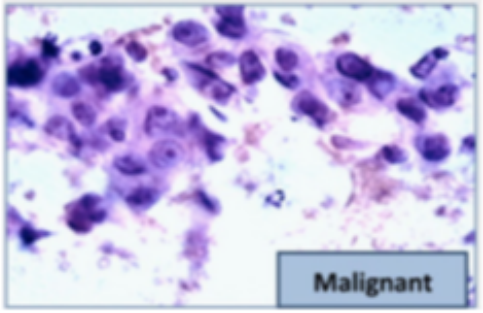
Chest X-rays are currently the best available method for diagnosing pneumonia, playing a crucial role in clinical care and epidemiological studies. Pneumonia is responsible for more than 1 million hospitalizations and 50,000 deaths per year in the US alone.

READ OUR PAPER

https://stanfordmlgroup.github.io/projects/chexnet/

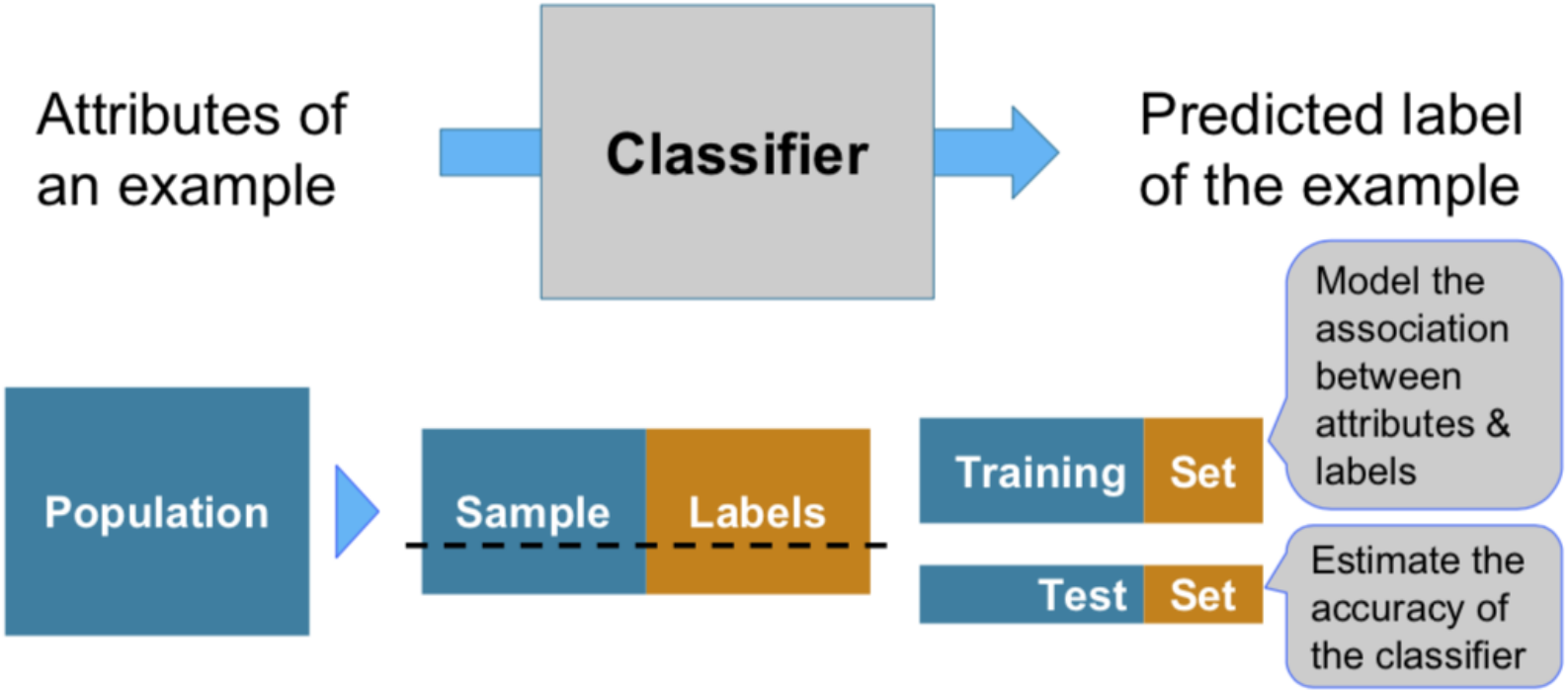
The Google Science Fair



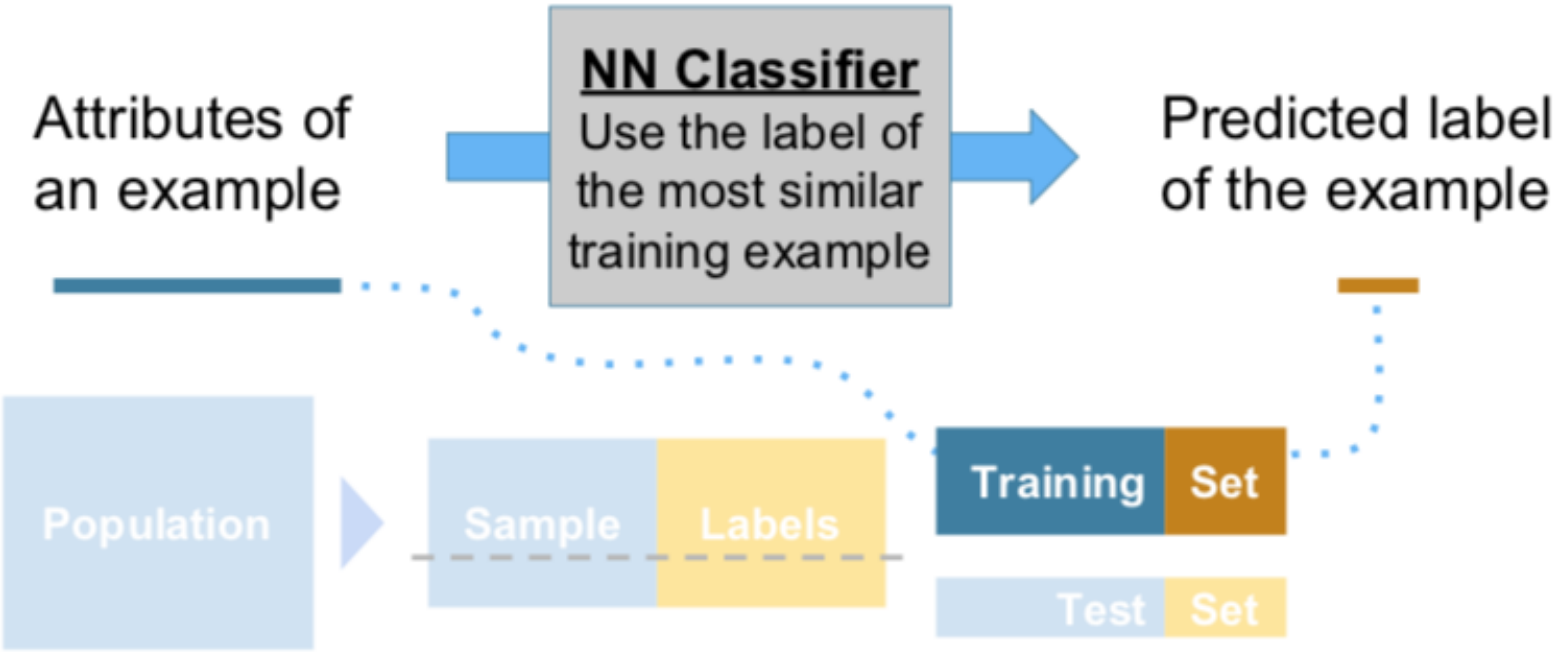


- Brittany Wenger, a 17-year-old high school student in 2012
- Won by building a breast cancer classifier with 99% accuracy

Training a Classifier



Nearest Neighbor Classifier



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)

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 ...

(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)