YData: An Introduction to Data Science

Lecture 35: Classifiers

Jessi Cisewski-Kehe and John Lafferty
Statistics & Data Science, Yale University
Spring 2019

Credit: data8.org

Announcements

Review:
Hypothesis testing
Regression Inference
(continued from Wed.)

Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values
- This displays the **empirical distribution of the statistic under the null hypothesis**
- It is a prediction about the statistic, made by the null hypothesis
  - It shows all the likely values of the statistic
  - Also how likely they are (assuming the null hypothesis is true)

Resolve choice between null and alternative hypotheses

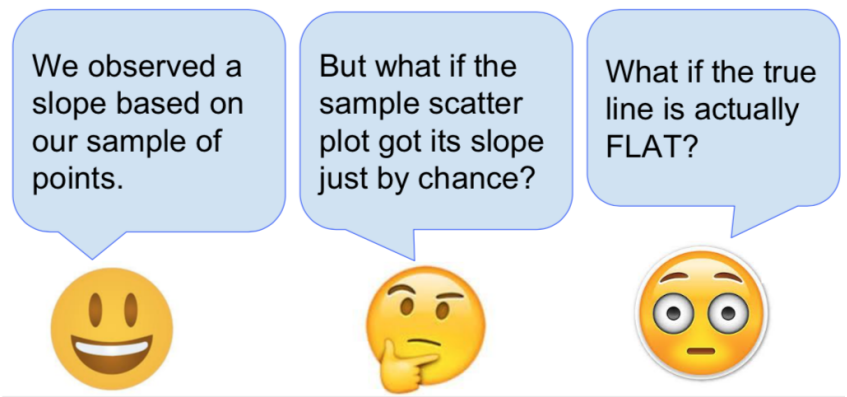
- Compare the **observed test statistic** and its empirical distribution under the null hypothesis
- If the observed value is not consistent with the distribution, then the test favors the alternative – "rejects the null hypothesis"

Using a CI for Testing (Lecture 24)

- Null hypothesis: **Population average = x**
- Alternative hypothesis: **Population average $\neq$ x**
- Cutoff for P-value: p%
- Method:
  - Construct a (100-p)% confidence interval for the population average
  - If x is not in the interval, reject the null
  - If x is in the interval, can't reject the null

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, reject the null hypothesis.
  - If the interval does contain 0, there isn't enough evidence to reject the null hypothesis.



A/B testing: Comparing Two Samples (Lec 19,20)

- Previously, we only considered data from a single group

- Compare values of sampled individuals in Group A with values of sampled individuals in Group B.
  $\rightarrow$ Question: Do the two sets of values come from the same underlying distribution?
  $\rightarrow$ Answering this question by performing a statistical test is called A/B testing.

  Examples:
  (A) Birth weights of babies of mothers who smoked during pregnancy
  (B) Birth weights of babies of mothers who didn't

  (A) Control group
  (B) Treatment group

  Deflategate

A/B testing: Simulating Under the Null

- If the null is true, all rearrangements of the birth weights among the two groups are equally likely

- Plan:
  - Shuffle all the birth weights
  - Assign some to "Group A" and the rest to "Group B", maintaining the two sample sizes
  - Find the difference between the averages of the two shuffled groups
  - Repeat

Discussion question

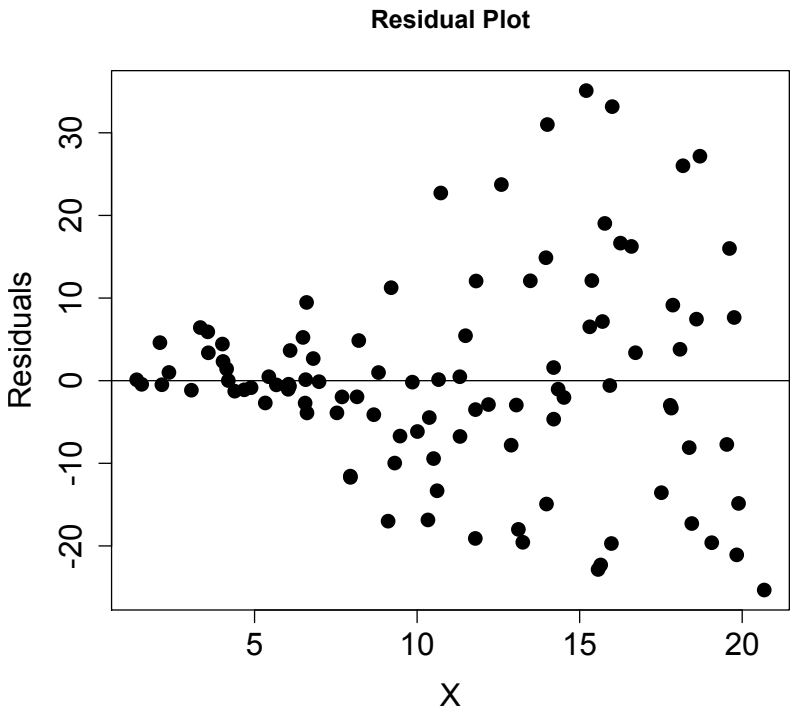
A study on the effect of caffeine involved asking subjects to take a memory test 20 minutes after drinking cola. Some subjects were randomly assigned to drink caffeine-free cola, and some to drink regular cola (with caffeine). For each subjects, a test score (the number of items recalled correctly) was recorded. The subjects were not told which type of cola they had been given.

a. The memory test had a total of 25 items on it. The average number of items recalled was 15 for the caffeine-free group and 16 for the regular cola group. Are the values 15 and 16 statistics or parameters?

b. Can an A/B hypothesis testing framework be used here? How?

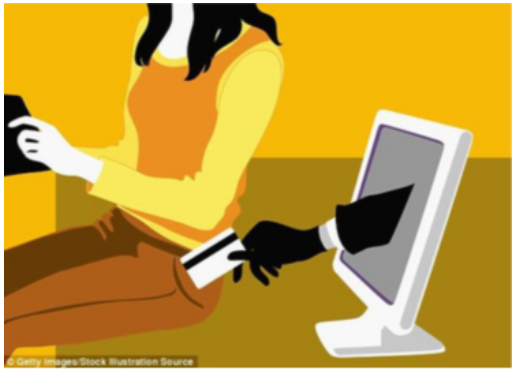
Discussion question

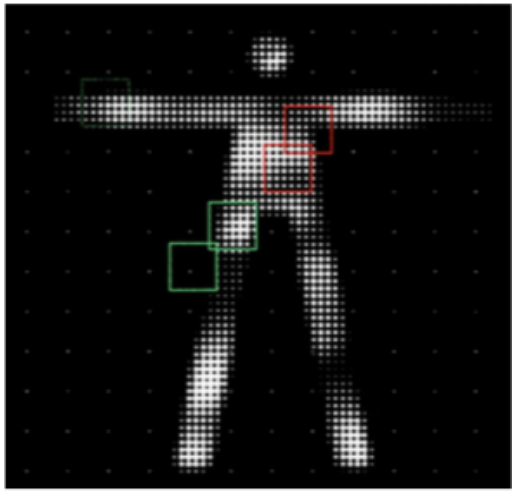
Suppose a Least-squares linear model was fit on explanatory variable X and response variable Y, with the residuals plotted in the figure below against X. What linear model assumption appears to be violated given the residual plot below?



Classification

Classification Example

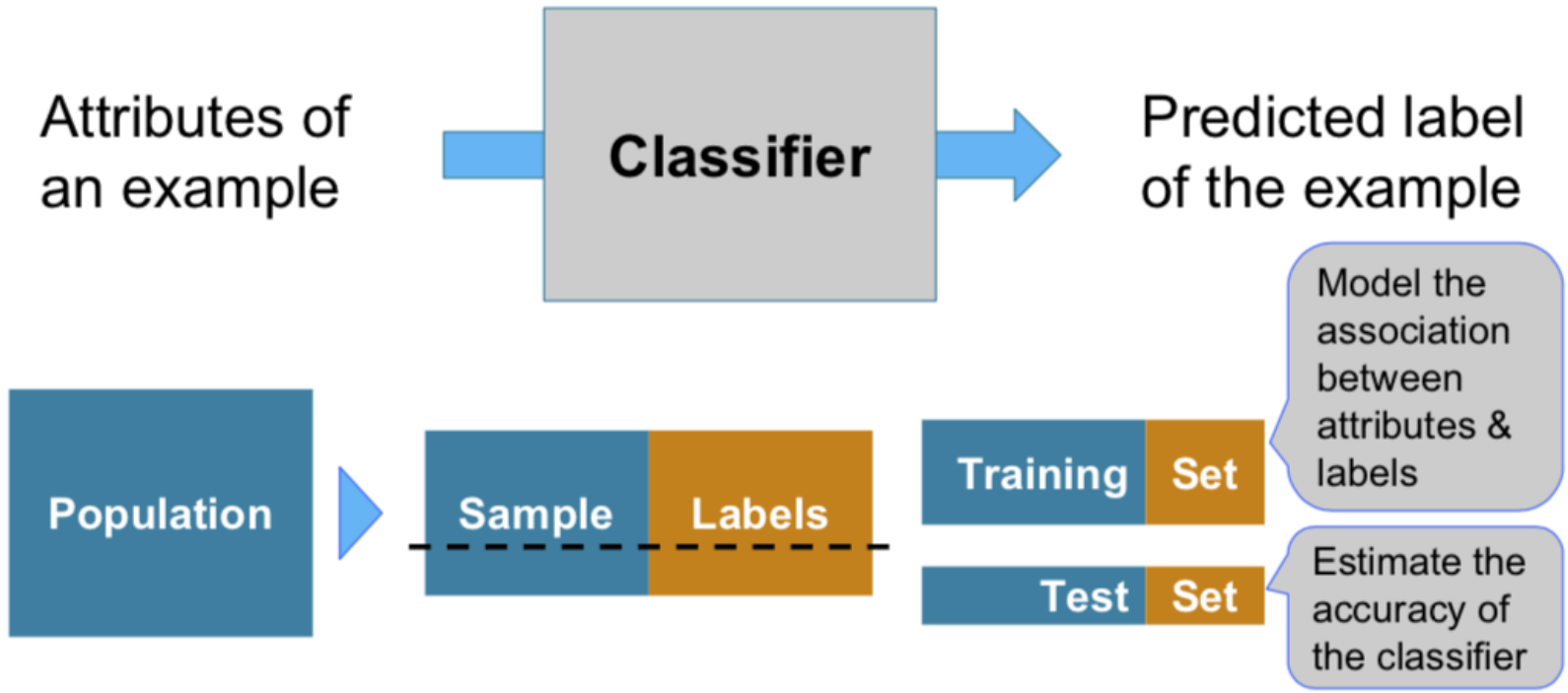




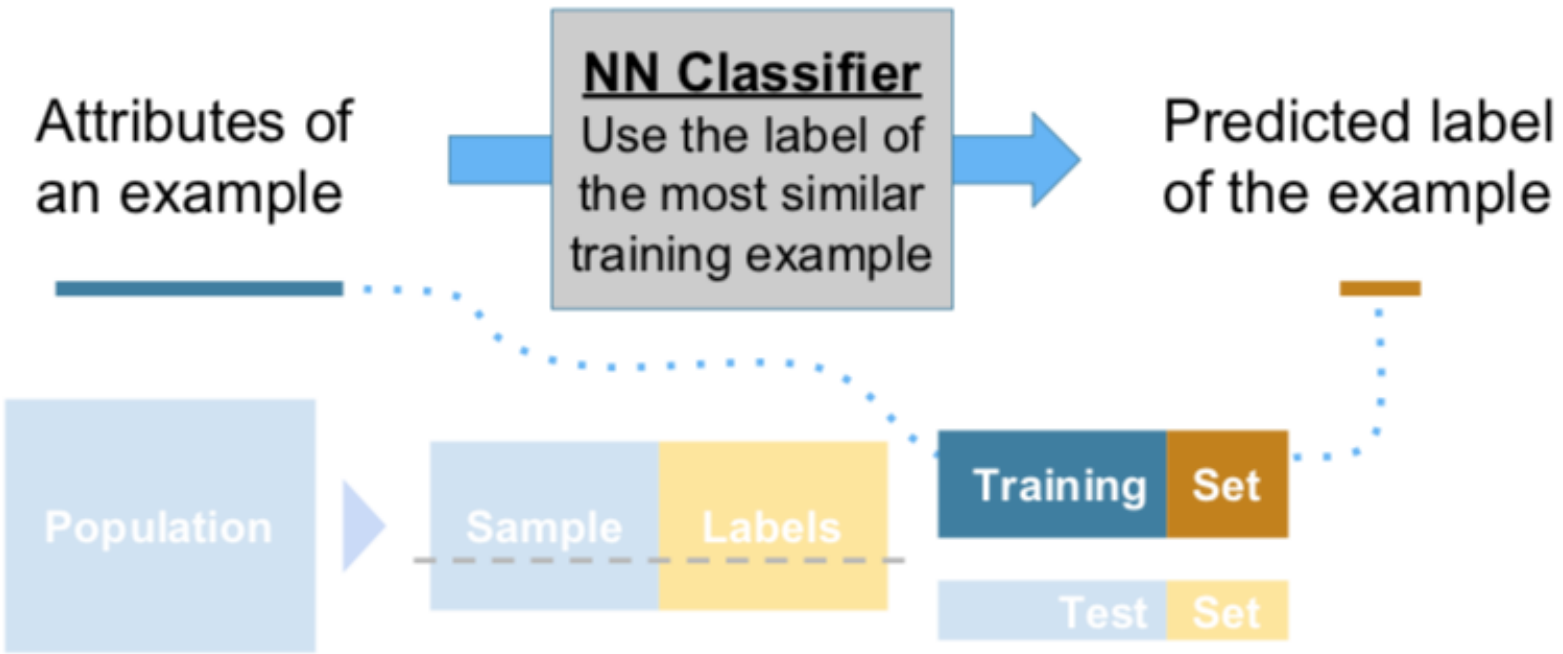
(DEMO from Wed. 4/17)

Classifiers

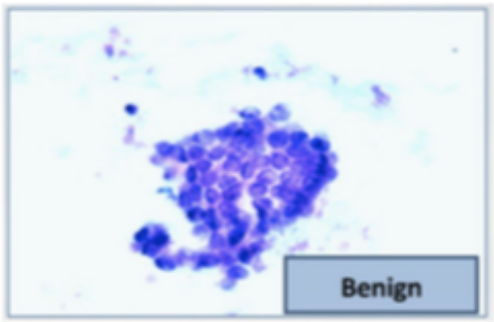
Training a Classifier

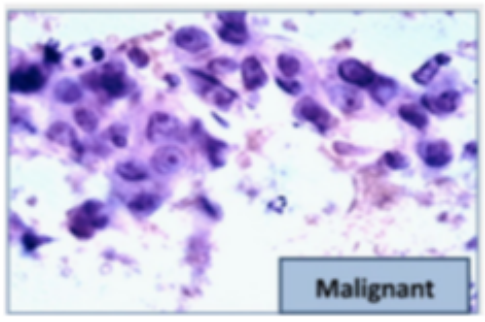


Nearest Neighbor Classifier



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

(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)

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

(DEMO)

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)

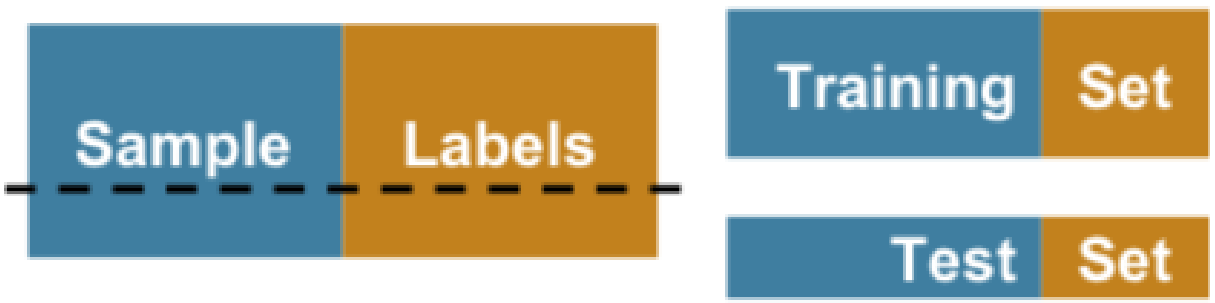
Evaluation

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)