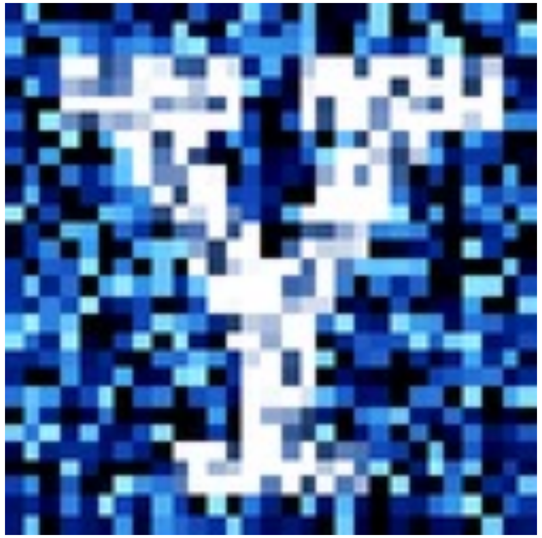
YData: Introduction to Data Science



Lecture 23: Classification

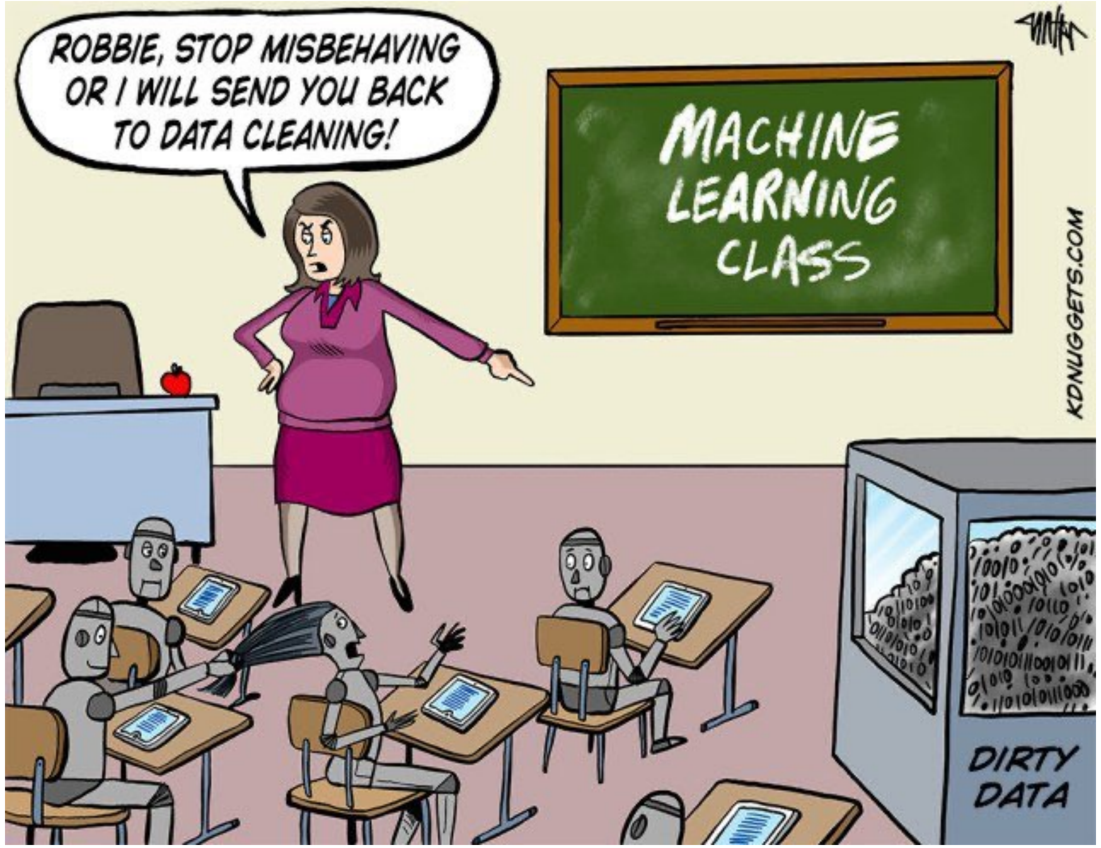
Overview

Quick review of hypothesis tests and confidence intervals

Creating confidence intervals (using hypothesis tests)

Classification

- KNN classifier
- If there is time: overitting


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

Project timeline

~~Tuesday, April 11th~~

- ~~Projects are due on Gradescope at 11pm on~~
- ~~Also, email a pdf of your project to your peer reviewers~~
  - ~~A list of whose paper you will review has been posted to Canvas~~

Wednesday, April 19th

- Jupyter notebook files with your reviews need to be sent to the authors and a pdf needs to be submitted to Gradescope
- A template for doing your review is available on Canvas

Sunday, April 30th

- Project is due on Gradescope
  - Add peer reviews to an Appendix of your project

Homework 9 has been posted

- It is due April 23rd

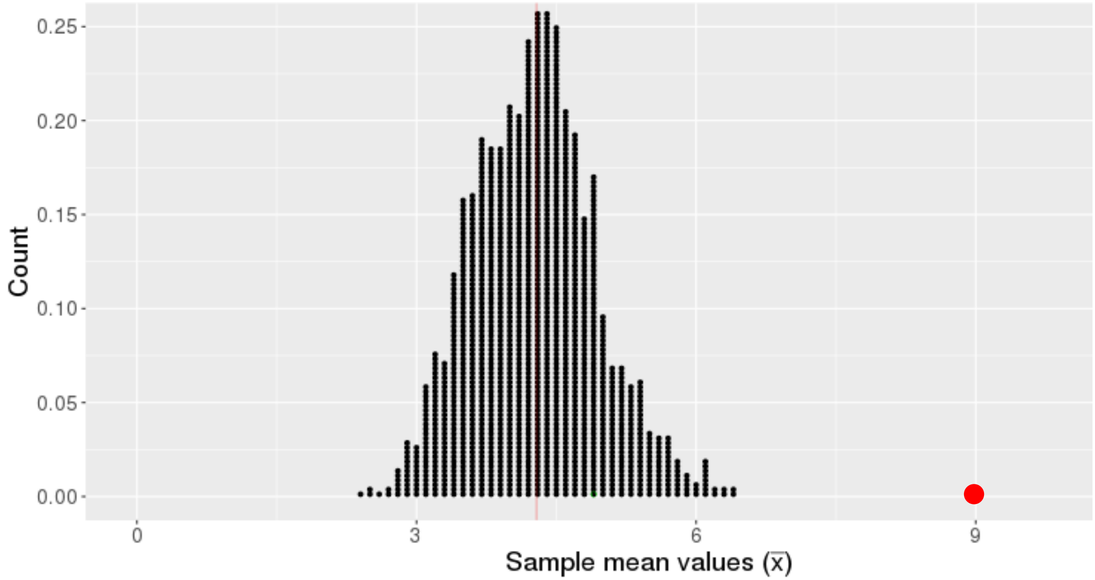
Review of hypothesis tests and confidence intervals

Basic hypothesis test logic

We start with a claim about a population parameter

- E.g., μ = 4[illegible]

This claim implies we should get a certain distribution of statistics

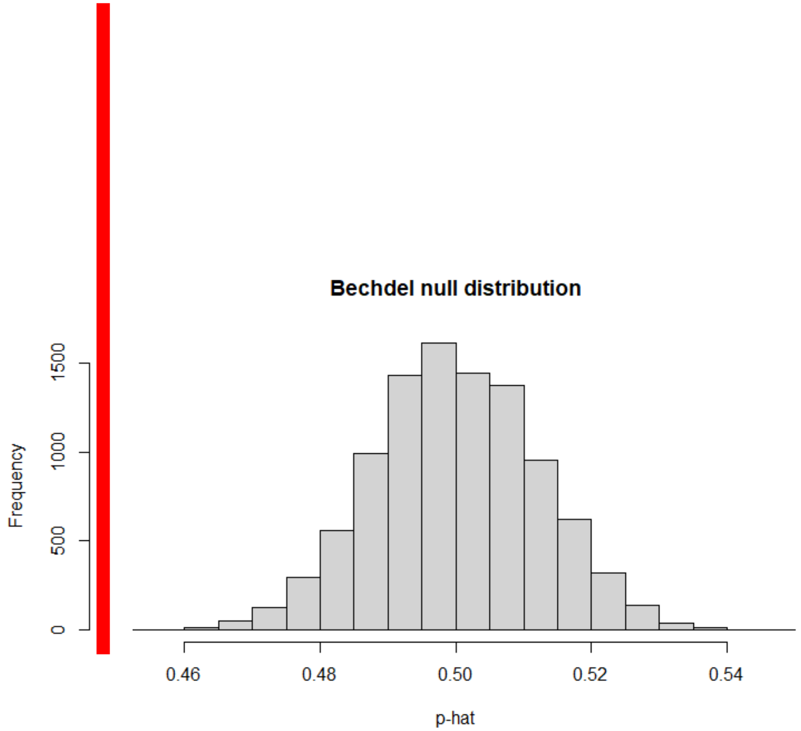


If our observed statistic is highly unlikely, we reject the claim

Null and Alternative hypotheses

Null hypothesis

- A hypothesis where ”nothing interesting” happened
  - E.g., our experiment failed
- We can simulate data under the assumptions of this model to get a "null distribution" of statistics

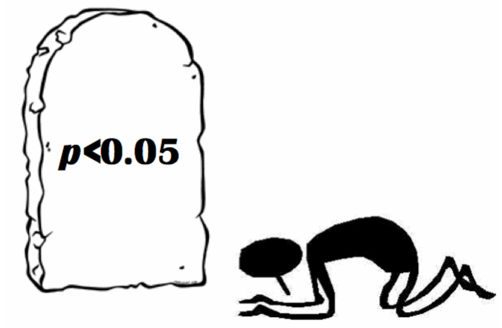


Alternative hypothesis

- The hypothesis we believe in (would like to see true)

p-value: the probability, that we get a statistic as or more extreme than the observed statistic from the null distribution

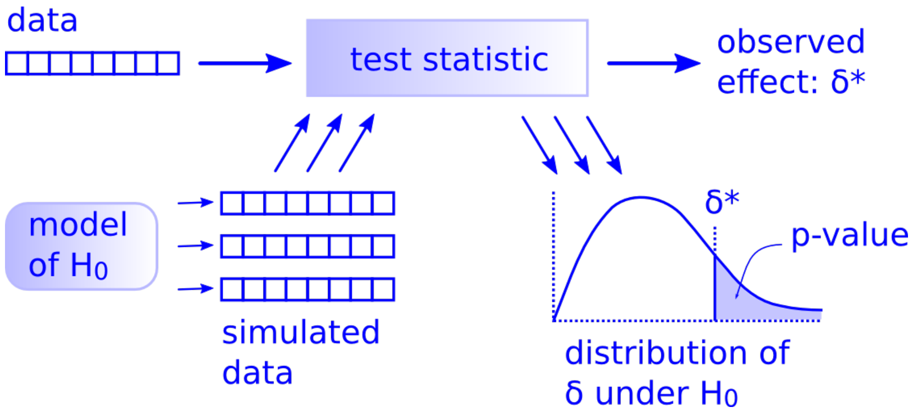
- P(Null_Stat ≥ obs_stat | H_0)

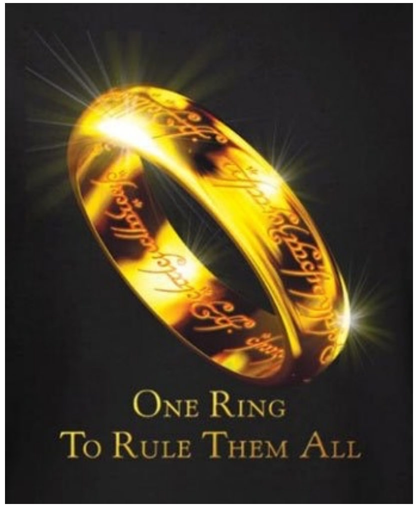


Steps needed to run a hypothesis test

To run a hypothesis test, we can use 5 steps:

1. State the null and alternative hypothesis
2. Calculate the observed statistic of interest
3. Create the null distribution
4. Calculate the p-value
5. Make a decision





Bechdel (hypothesis) test

WELL, ..IDUNNO. I HAVE THIS RULE, SEE...
I ONLY GO TO A MOVIE IF IT SATISFIES THREE BASIC REQUIREMENTS. ONE, IT HAS TO HAVE AT LEAST TWO WOMEN IN IT.
BARBARIAN
WHO, TWO, TALK TO EACH OTHER ABOUT, THREE, SOMETHING BESIDES A MAN.
THE VIGILANTE
PRETTY STRICT, BUT A GOOD IDEA.
NO KIDDING. LAST MOVIE I WAS ABLE TO SEE WAS ALIEN...

1. State the null hypothesis and the alternative hypothesis

- 50% of the movies pass the Bechdel test: H_0: $\pi = 0.5$
- Less than 50% of movies pass the: H_A: $\pi < 0.5$

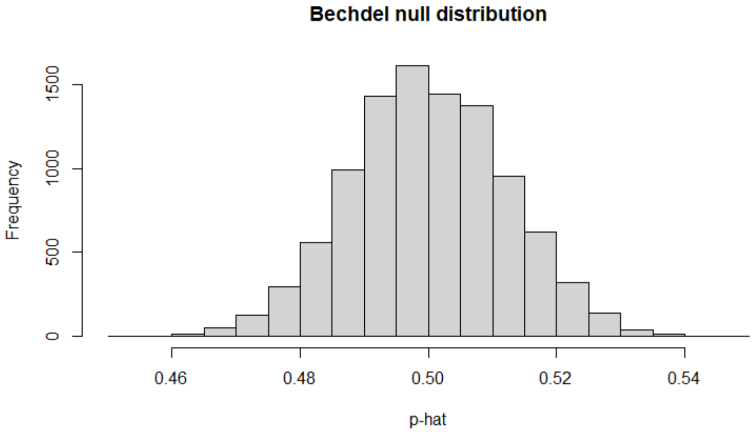
2. Calculate the observed statistic

- 803 out of 1794 movies passed the Bechdel test

$\hat{p} = .448$

3. Create a null distribution that is consistent with the null hypothesis

- i.e., the statistics we expect if 50% of the movies passed the Bechdel test



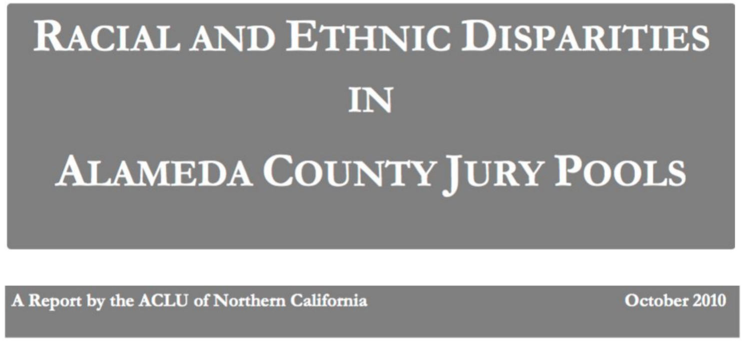
4. Examine how likely the observed statistic is to come from the null distribution

- What is the probability that only 803 of 1794 movies would pass the Bechdel test ($\hat{p} = .448$) if the null hypothesis was true?
- i.e., what is the p-value?

5. Make a judgement

- A small p-value this means that $\pi = .5$ is unlikely, and so it is likely $\pi < .5$
- i.e., we say our results are ‘statistically significant’

Jury selection in Alameda county



Racial and Ethnic Disparities in Alameda County Jury Pools

A Report by the ACLU of Northern California October 2010

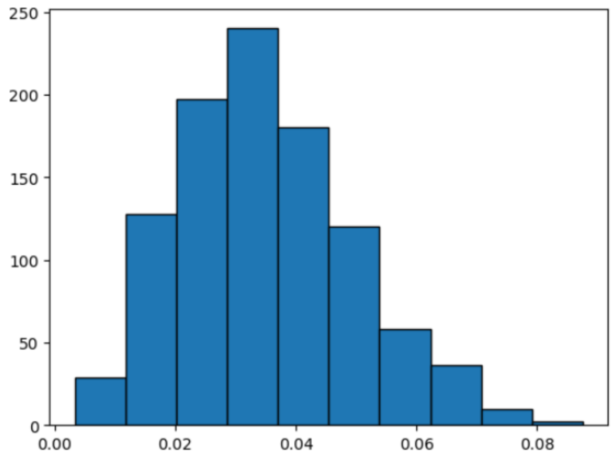
1. State the null hypothesis and the alternative hypothesis
   - Jury panels match population demographics: H_0: $\pi_A = .15$, $\pi_L = 0.12$, etc.
   - At least one ethnicity is not correctly represented: H_A: π_i differs from H_0

2. Calculate the observed statistic

$$TVD = \sum_{i=1}^{k} |\pi_i - \hat{p}_i|$$

TVD = .28

3. Create a null distribution that is consistent with the null hypothesis
   - The TVD statistics we expect if the null hypothesis was true
   - i.e., the TVD statistics we would expect if the sample demographics matched the population demographics



4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that we would get a TVD statistic larger than 0.28 if the null hypothesis was true?
   - i.e., what is the p-value?

5. Make a judgement
   - A small p-value this means that at least one demographic on juries does not match their representations in the population
   - i.e., we say our results are 'statistically significant'

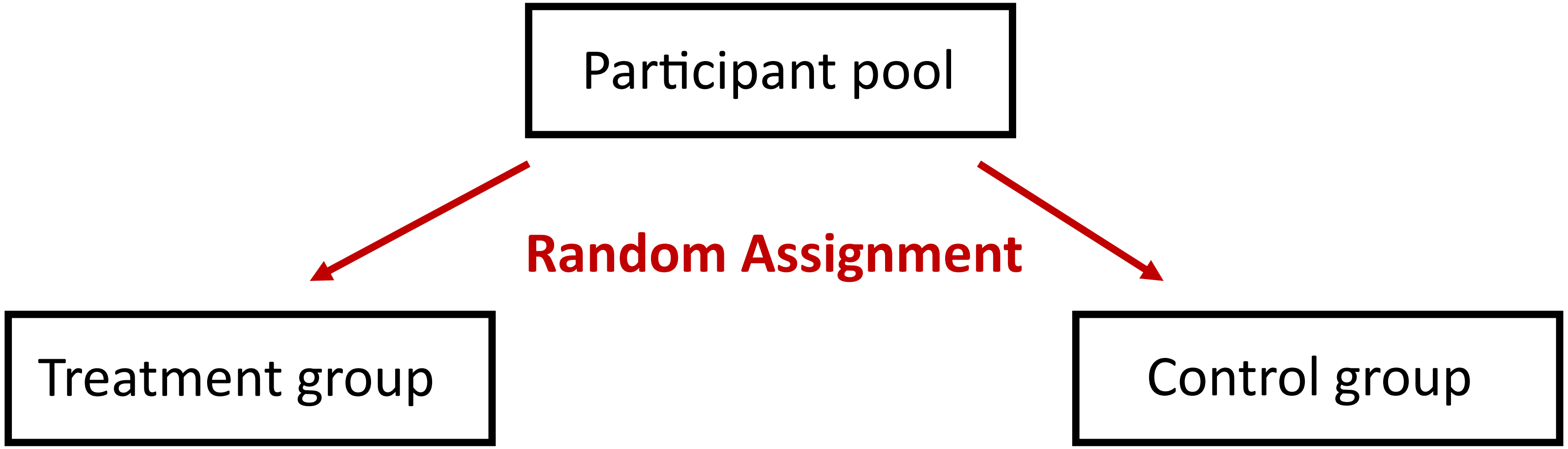
Using random assignment to find causal relationships

An association is the presence of a reliable relationship between the treatments an outcome

A causal relationship is when changing the value of a treatment variable influences the value outcome variable

We can use **random assignment** to find causal relationships!

- i.e., we can run a "Randomized Controlled Trial"



BTA for back pain relief

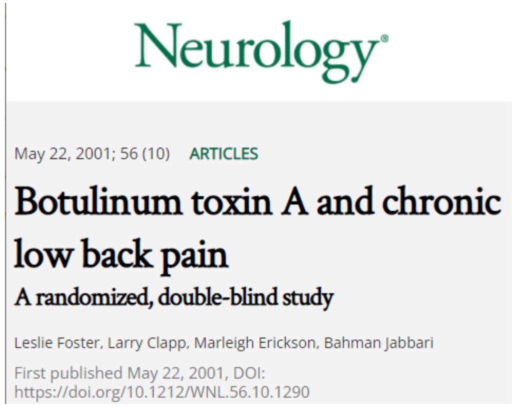
1. State the null hypothesis and the alternative hypothesis
   - BTA does not lead to an increase in pain relief: H_0: $\pi_{treat} = \pi_{control}$
   - BTA leads to an increase in pain relief: H_A: $\pi_{treat} > \pi_{control}$

2. Calculate the observed statistic: $\hat{p}_{treat}$ - $\hat{p}_{control}$

3. Create a null distribution that is consistent with the null hypothesis
   - The $\hat{p}_{treat}$ - $\hat{p}_{control}$ statistics we expect if the null hypothesis was true
   - i.e., statistics we would expect if there was no difference in pain relief between the two groups

4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that we would get a $\hat{p}_{treat}$ - $\hat{p}_{control}$ statistic larger than 0.475 if the null hypothesis was true?
   - i.e., what is the p-value?

5. Make a judgement
   - A small p-value this means that at the proportion of pain relief differed between the two groups
     - i.e., we say our results are 'statistically significant'
   - Because our analysis is based on a randomized controlled trial (using random assignment) we can say that BTA causes an increase in pain relief



Neurology®

May 22, 2001; 56 (10) ARTICLES

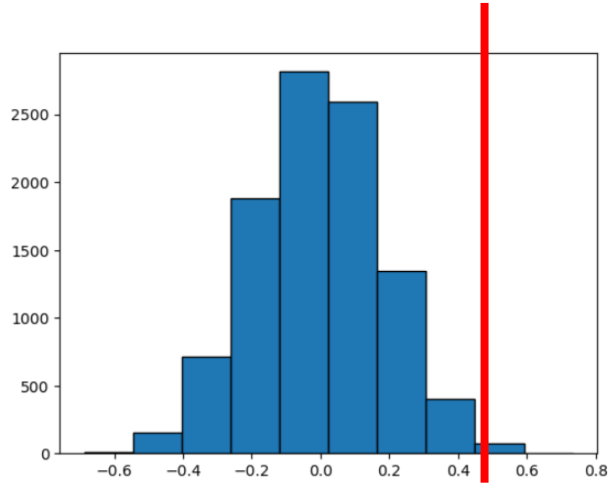
Botulinum toxin A and chronic low back pain

A randomized, double-blind study

Leslie Foster, Larry Clapp, Marleigh Erickson, Bahman Jabbari

First published May 22, 2001, DOI: https://doi.org/10.1212/WNL.56.10.1290

$\hat{p}_{treat}$ - $\hat{p}_{control}$ = .475



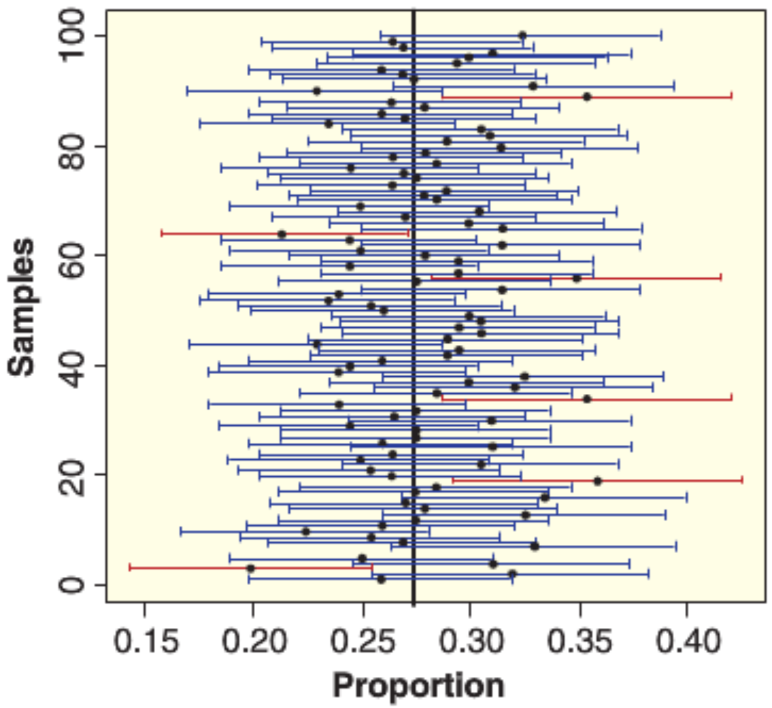
Confidence Intervals

A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

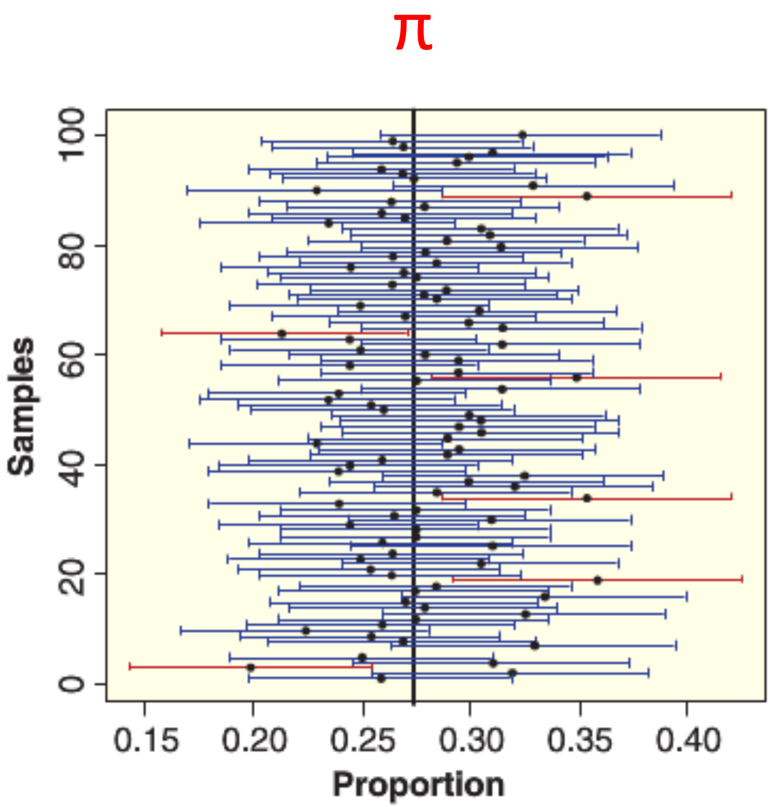
The **confidence level** is the percent of all intervals that contain the parameter

π



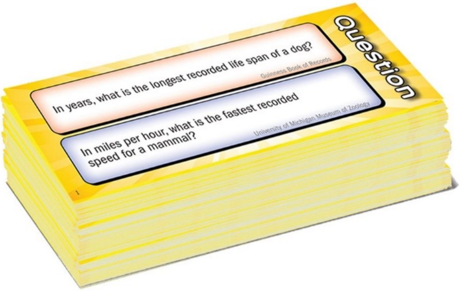
90% Confidence Intervals

For any given confidence interval, we don't know of it has the parameter in it



We just know that it will be in the interval most of the time

- E.g., 9 out of 10 times for a 90% confidence level



In years, what is the longest recorded life span of a dog?

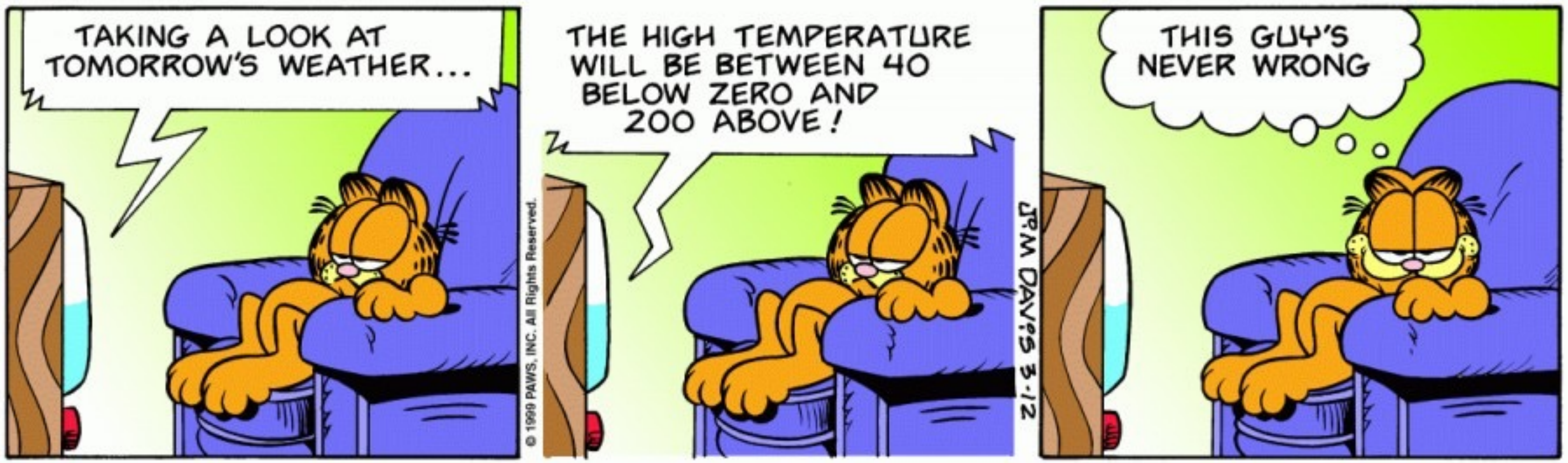
Guinness Book of Records

In miles per hour, what is the fastest recorded speed for a mammal?

University of Michigan Museum of Zoology

The Greek National Anthem

Tradeoff between interval size and confidence level


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

There is a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Using hypothesis tests to construct confidence intervals

Constructing confidence intervals

There are several methods that can be used to construct confidence intervals including

- ”Parametric methods” that use probability functions
    - E.g., confidence intervals based on the normal distribution
- A “bootstrap method” where data is resampled from our original sample to approximate a sampling distribution

To learn more about these methods, take Introductory Statistics!

Constructing confidence intervals

We are going to use a less conventional method to get confidence intervals based on the relationship between confidence intervals and hypothesis tests

- The method we will discuss is valid, but can be more computationally expensive than other methods

What we will do is to run a series of hypothesis test with different null hypothesis parameter values

Our confidence interval will be all parameters values where we **fail to reject** the null hypothesis

H_0: $\pi = \pi_0$

Failure to reject $\pi = \pi_0$

means π_0 is plausible

Motivation: Bechdel Confidence Interval

From running a hypothesis test on the Bechdel data, we saw that H_0: $\pi = .5$ is unlikely

- i.e., it was not plausible that 50% of movies pass the Bechdel test

WELL, ...IDUNNO. I HAVE THIS RULE, SEE...
I ONLY GO TO A MOVIE IF IT SATISFIES THREE BASIC REQUIREMENTS. ONE, IT HAS TO HAVE AT LEAST TWO WOMEN IN IT..
THE BARBARIAN
..WHO, TWO, TALK TO EACH OTHER ABOUT, THREE, SOMETHING BESIDES A MAN.
THE VIGILANTE
PRETTY STRICT, BUT A GOOD IDEA.
NO KIDDING. LAST MOVIE I WAS ABLE TO SEE WAS ALIEN...

But what is a reasonable range of values for the population proportion of movies that pass the Bechdel test?

Let's create a confidence interval for π_{Bechdel} to find out!

Let's explore this in Jupyter!

Classification

Prediction: regression and clasification

We “learn” a function f

- f(**x**) $\longrightarrow$ y

Input: **x** is a data vector of "features"

Output:

- Regression: output is a real number (y ϵ R)
- Classification: output is a categorical variable y_k

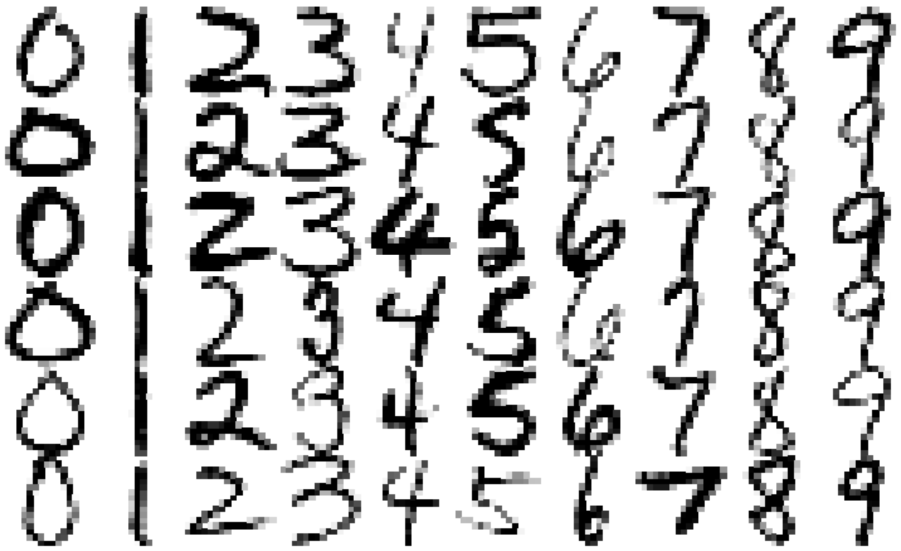
Example: salmon or sea bass?

What are the features and labels in this task?

- Labels (y): Salmon or Sea bass
- Features (X): Length, width, lightness, number and shape of fins, etc.

From Duda, Hart and Stork, 2001

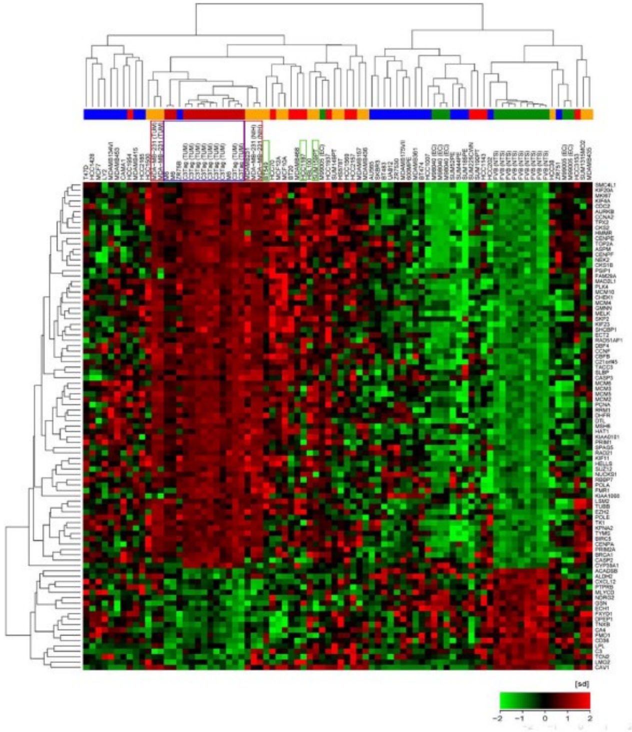
Example: what is in this image?



What are the features and labels in this task?

- Labels (y): cat, dog, etc., or numbers 0 to 9
- Features (X): Pixel values

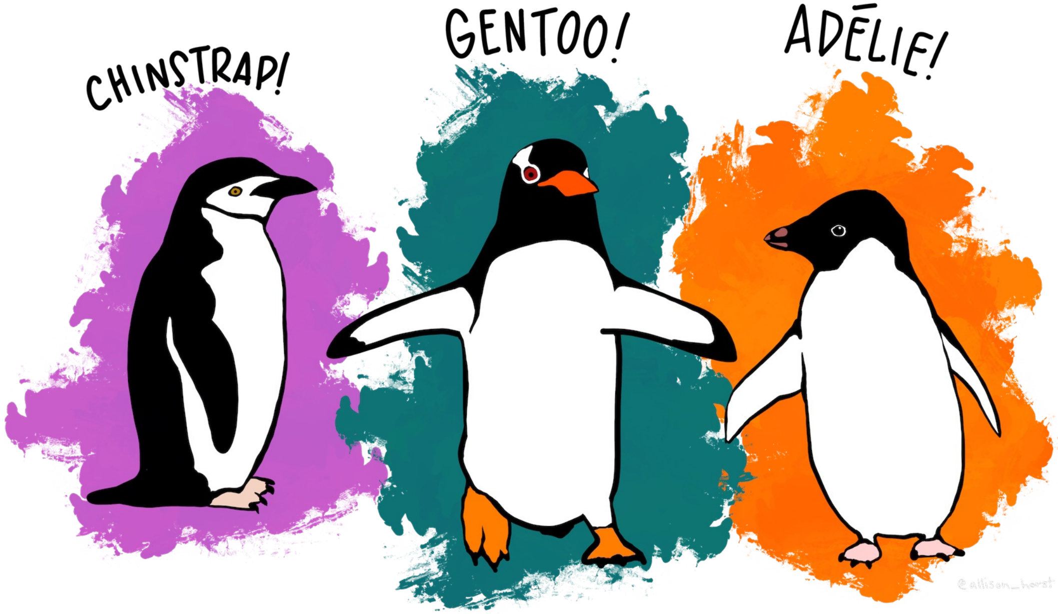
Example: predicting cancer



What are the features and labels in this task?

- Labels (y): Cancer/no caner
- Features (X): The expression level of different genes

Example: Penguin species

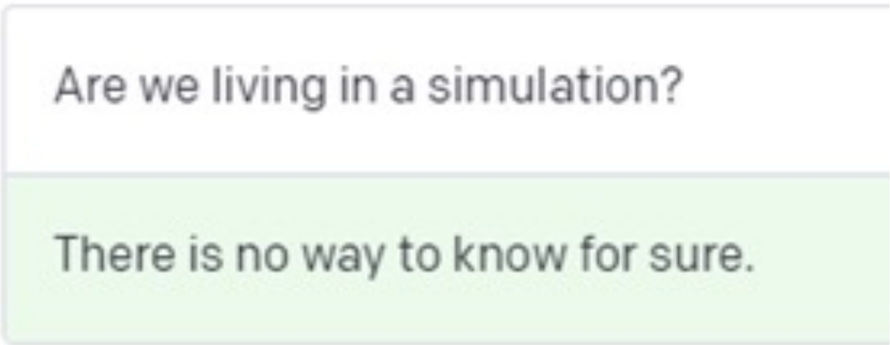


What are the features and labels in this task?

- Labels (y): Chinstrap, Gentoo, Adelie
- Features (X): Flipper length, bill length, body mass, …

Example: GPT-3 predicting/generating text

Question answering:

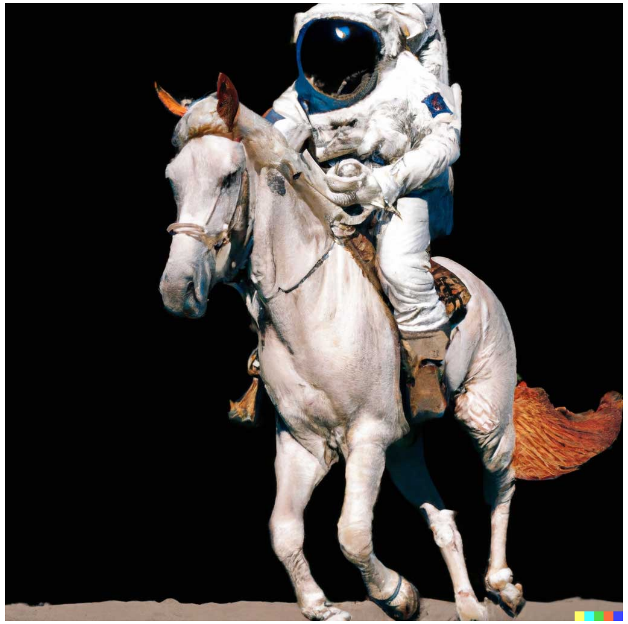


Are we living in a simulation?

There is no way to know for sure.

Image generation

"Draw an astronaut riding a horse"



What are the features and labels in this task?

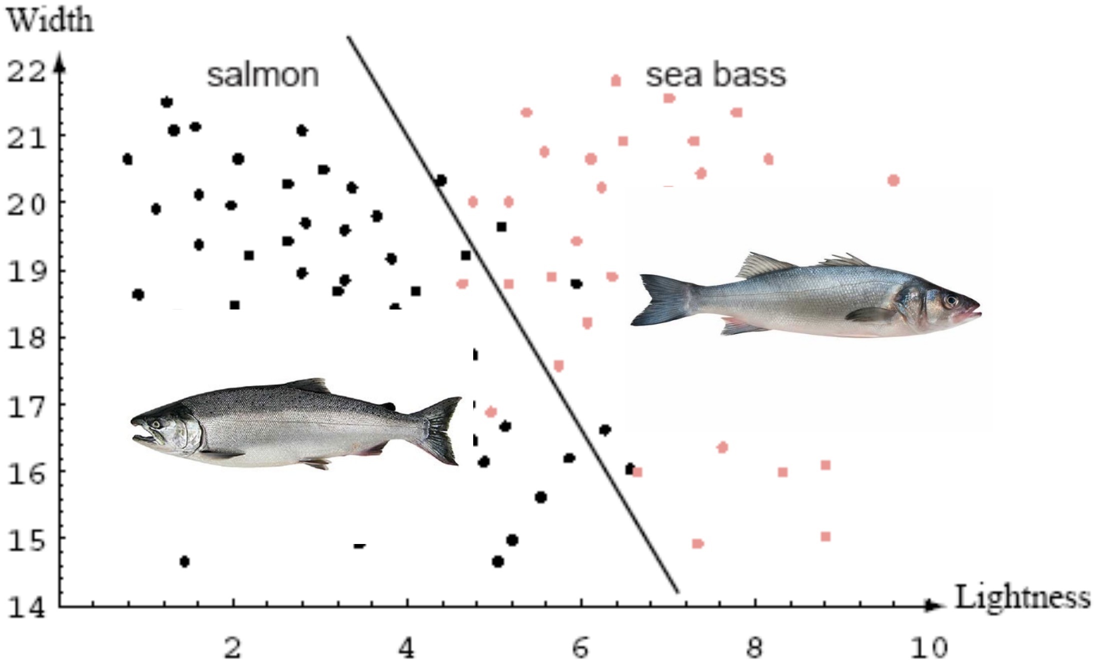
- Labels (y): Next word in a sentence (or an image)
- Features (X): Previous words/prompt words

Example of classifier on a feature set

A binary classifier is a function from the set of input features to {0, 1}

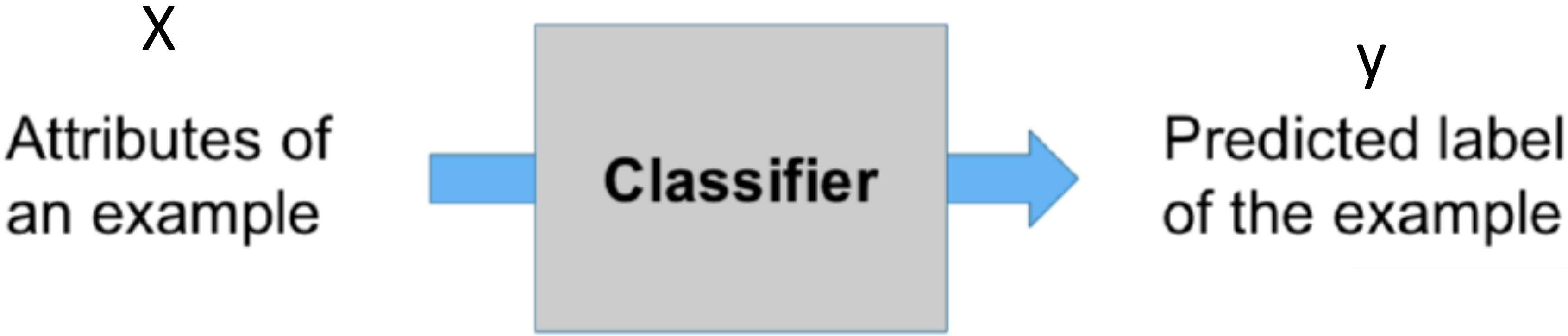
- E.g., f(pixel values) ⟶ salmon or bass

A linear classifier draws a straight line (or a multi-dimensional plane) between the two predicted values



Let's explore features and labels in Jupyter!

Training a classifier



X

Attributes of an example

Classifier

y

Predicted label of the example

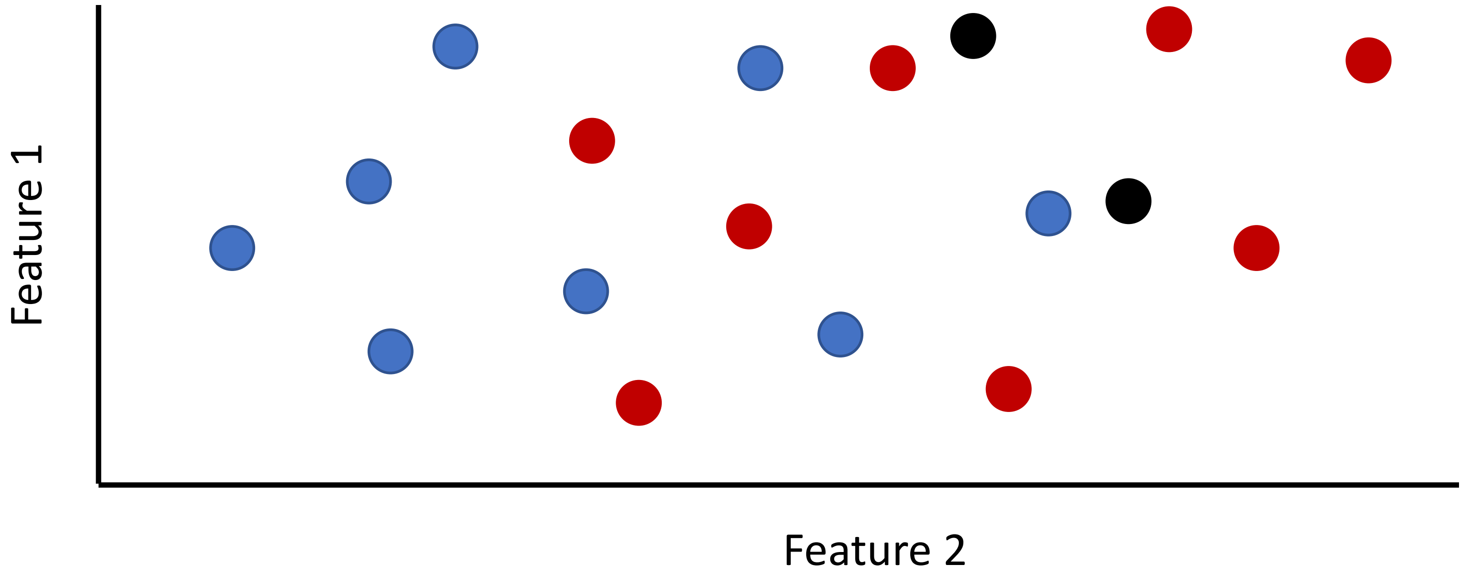
k-Nearest Neighbor classifier

Nearest Neighbor Classifier (k = 1)

Training the classifier: Store all the features with their labels

Making predictions: The label of closest training point is returned

- i.e., we find the distance between a test point and all training points, and the closest point is the prediction



Class 1

Class 2

Prediction:

Class 1

Class 2

Distance between two points

Two features x and y: $D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$

Three features x, y, and z: $D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$

- And so on for more features…

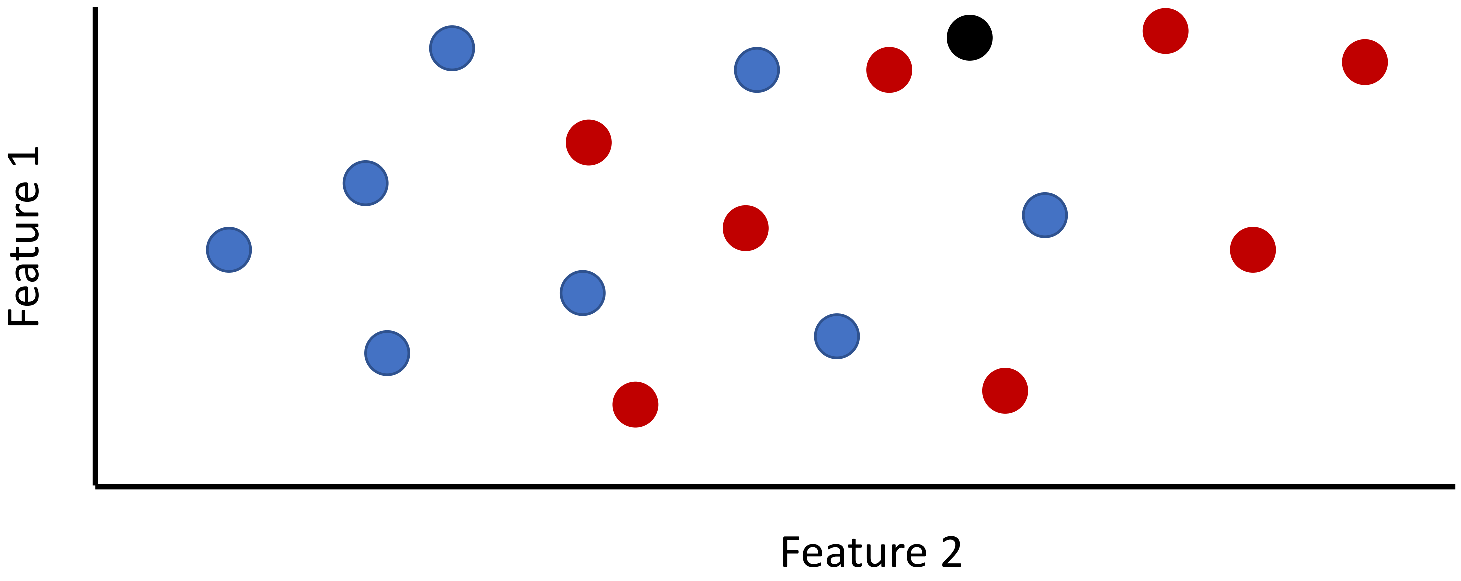
It's important the features are standardized

- If not, features that typically have larger values will dominate the distance measurement

Finding the k Nearest Neighbors (k ≥ 1)

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



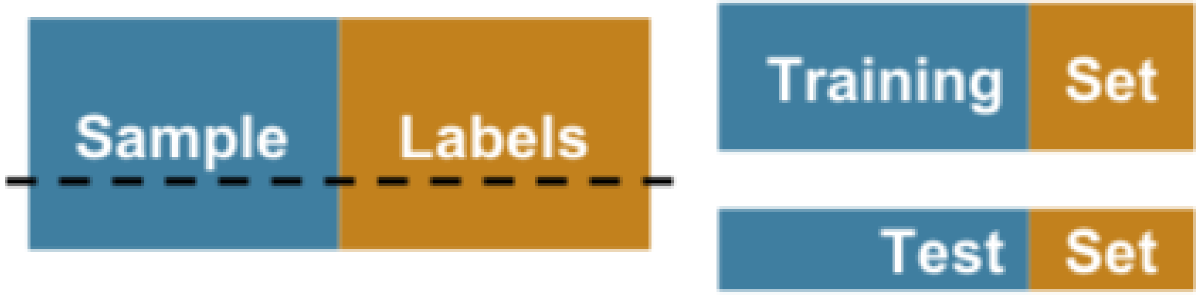
Let's explore this in Jupyter!

Evaluation

Accuracy of a classifier

The accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly on the *test set*

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



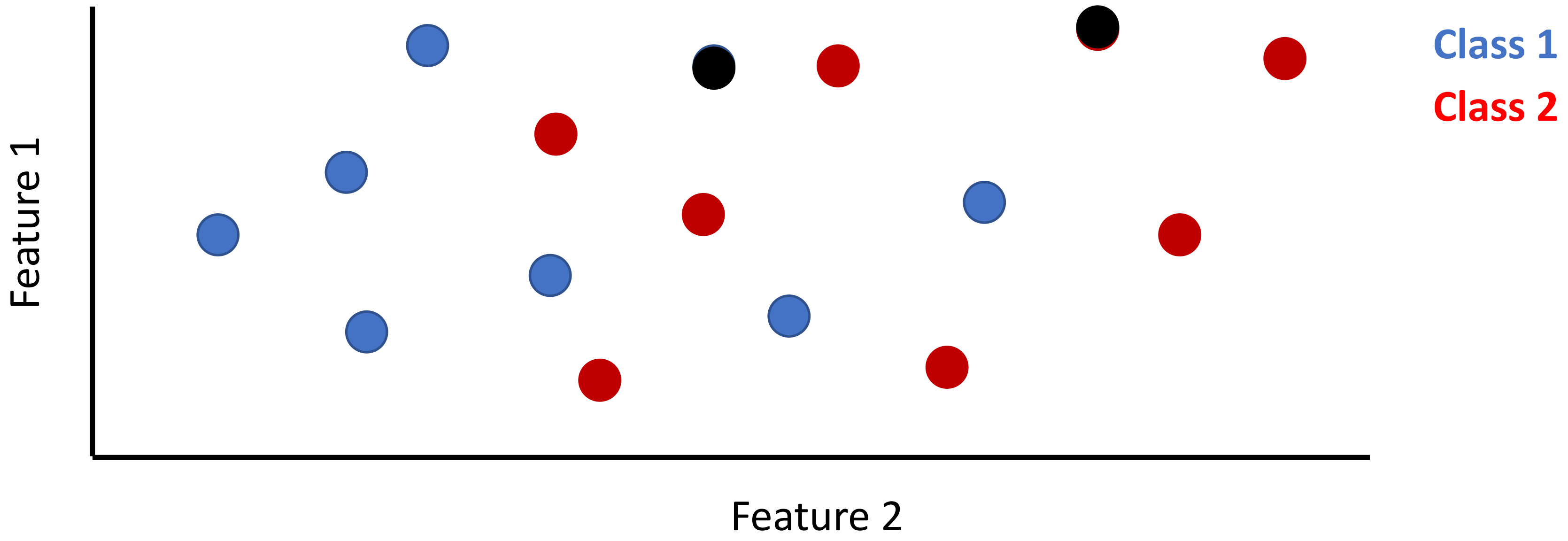
Training and test accuracy

Q: What would happen if we tested the classifier using the training data with k = 1?

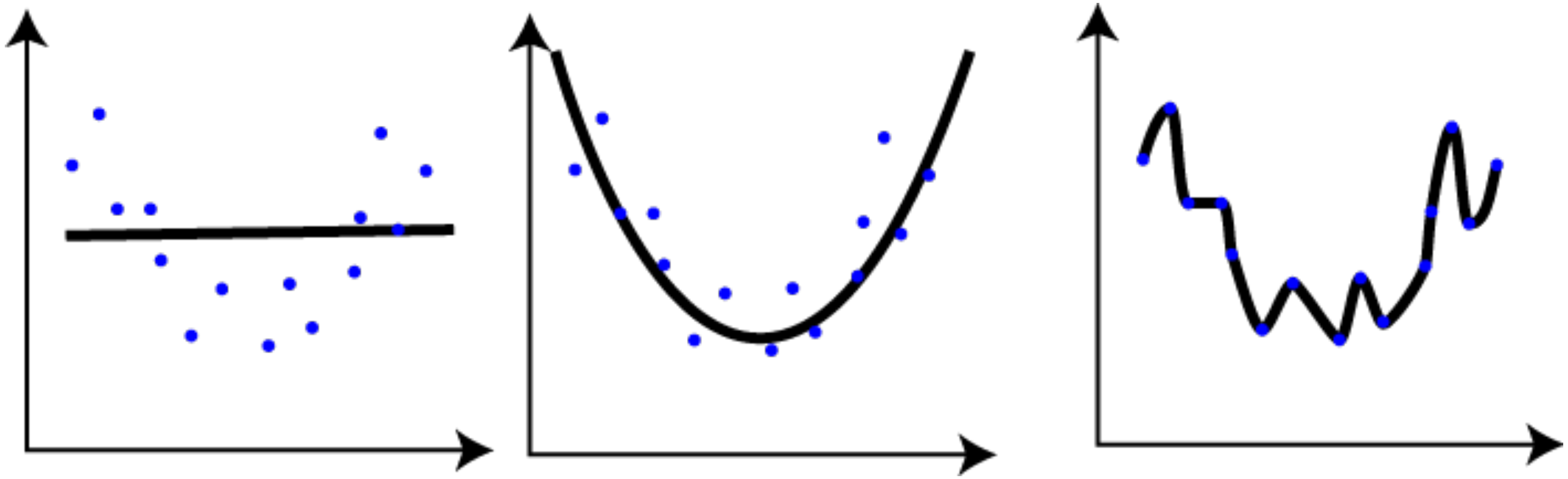
A: We would have 100% accuracy

Q: Would this indicate that the classifier is good?

- A: No!



Review: overfitting



Review: overfitting

If our classifier has over-fit to the training data then:

a) We might not have a realistic estimate of how accurate its predictions will be on new data

b) There might be a better classifier that would not over-fit to the data and thus can make better predictions

What we really want to estimate is how well the classifier will make predictions on new data, which is called the **generalization (or test) error**

Overfitting song…

Review cross-validation

Training error rate (training accuracy): model predictions are made on using the same data that the model was fit with

Test error rate (test accuracy): model predictions are made on a separate set of data

k-fold cross-validation

Are there any downsides to using half the data for training and half for testing?

Since we are only using half the data for training, potentially can build a better model if we used more of our data

- Say 90% of our data for training and 10% of testing

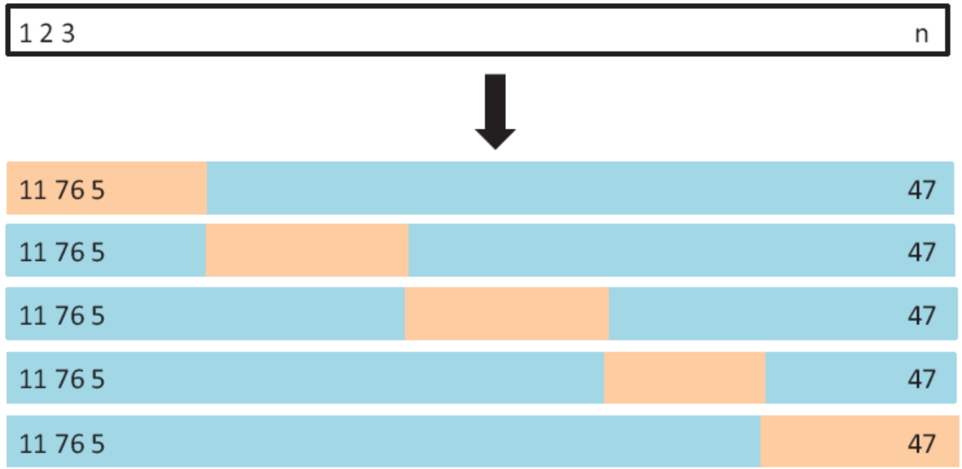
Problem: Then our estimate of the generalization error would be worse

Solutions?

k-fold cross-validation

k-fold cross-validation

- Split the data into k parts
- Train on k-1 of these parts and test on the left out part
- Repeat this process for all k parts
- Average the prediction accuracies to get a final estimate of the generalization error



Let's explore this in Jupyter!

Next class, building the KNN classifier ourselves