YData: Introduction to Data Science

Lecture 34: classification

Overview

Quick review

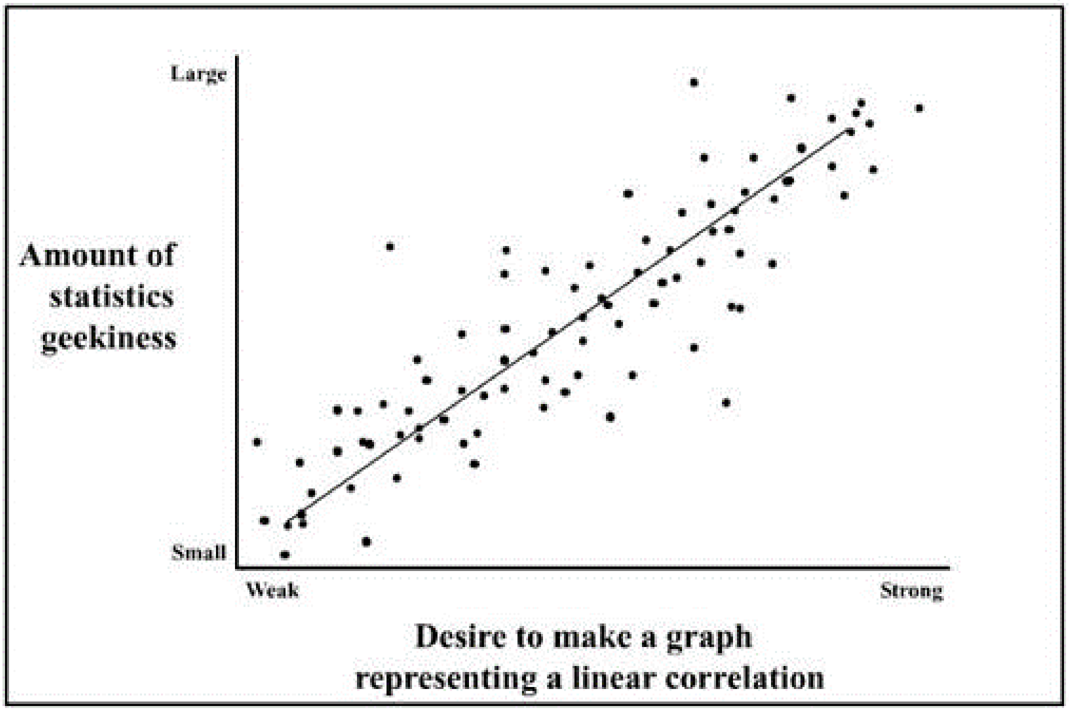
- Polynomial regression
- The correlation coefficient as the proportion of variability explained

Inference for regression

- Confidence intervals
- Hypothesis tests

If there is time:

- Classification



Announcements

Homework 10 has been posted

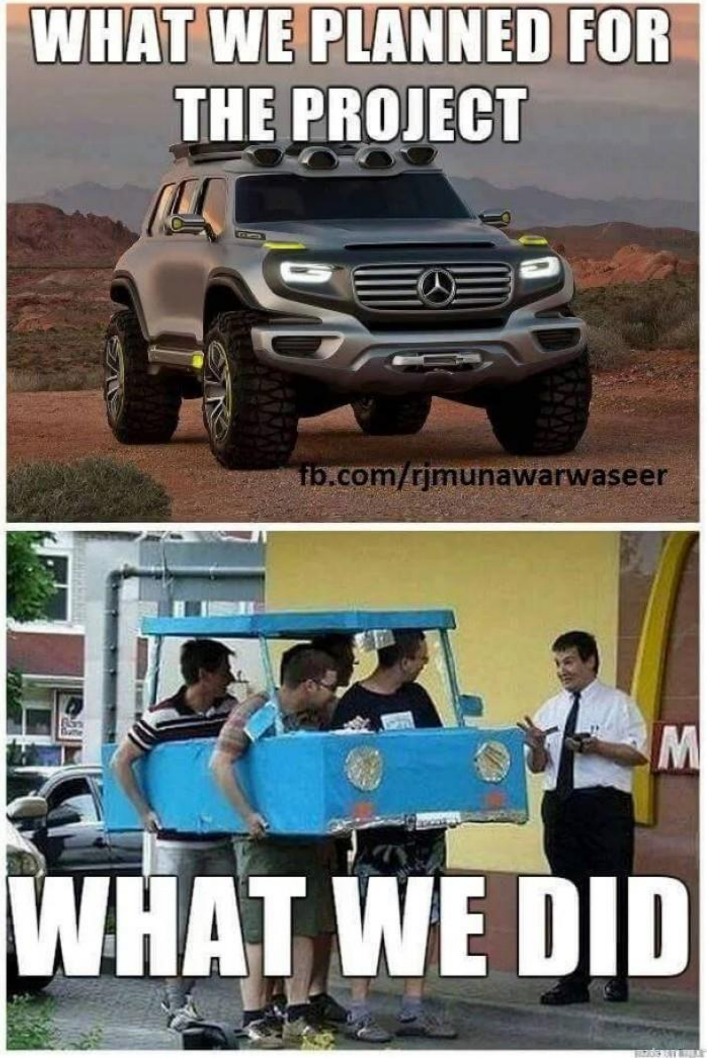
- It is due on Sunday the 24th

Practice 10 has been posted

- It is not due

Project 3 has been posted

- It is due Wednesday the 27th



Review

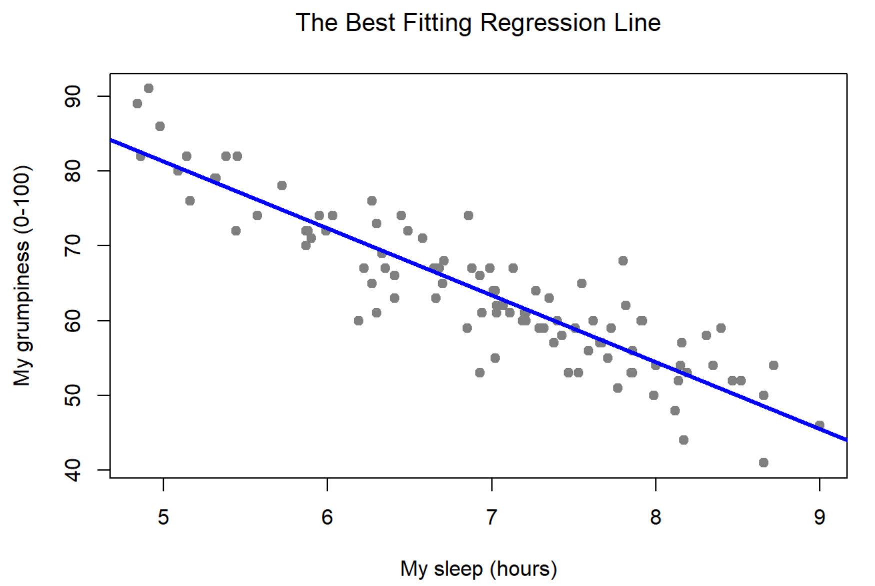
Regression

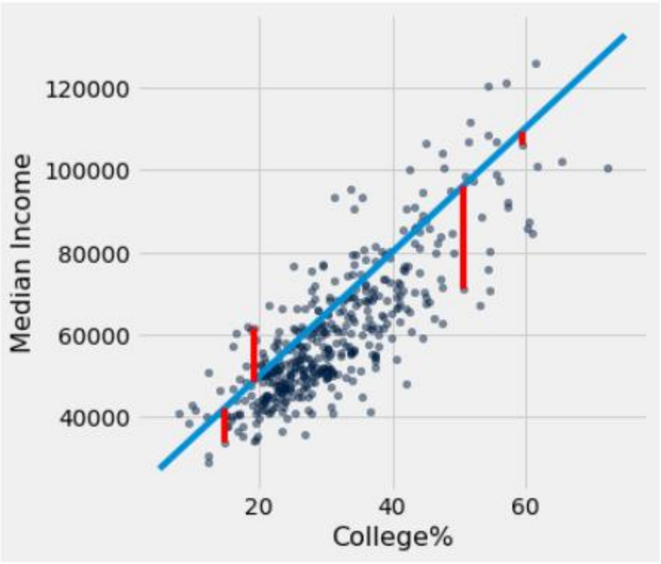
Regression is method of using one variable ***x*** ***to predict*** the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$
- Linear regression: $\hat{y} = \text{slope} \cdot x + \text{intercept}$
- Polynomial regression: $\hat{y} = a + b \cdot x + c \cdot (x)^2 + \ldots$

The coefficients for these regression models are found by minimizing the sum of the squared residuals

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

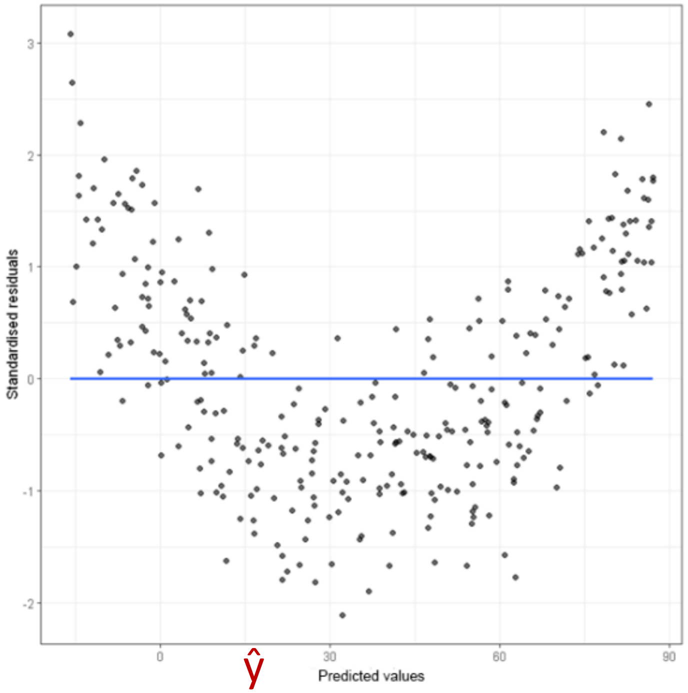




Regression diagnostics

A scatter diagram of residuals

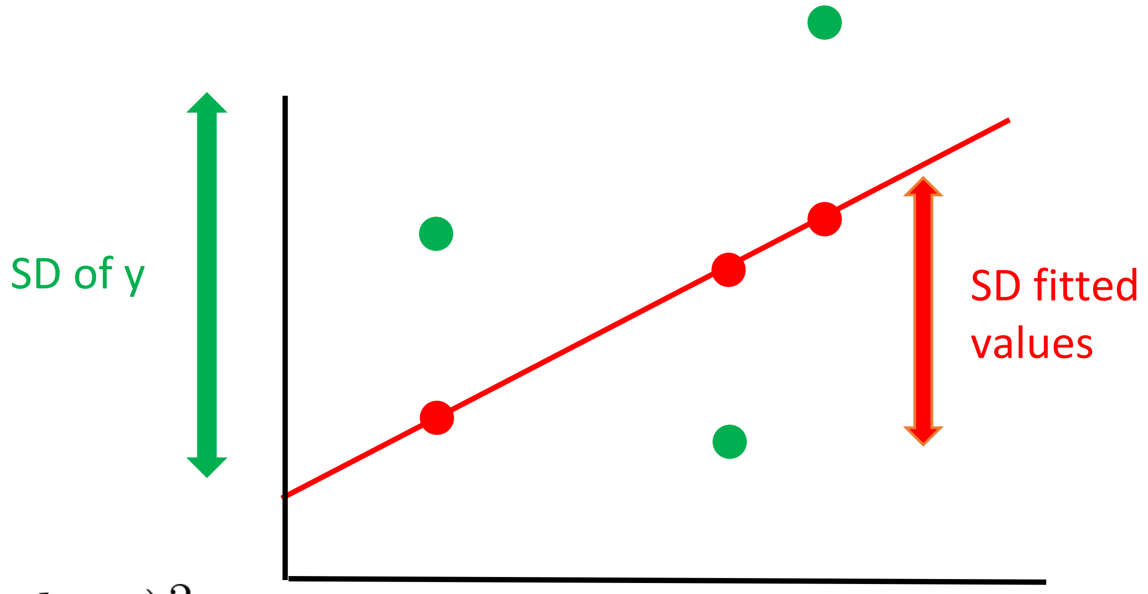
- Should look like an unassociated blob for linear relations
- But will show patterns for non-linear relations
- Used to check whether linear regression is appropriate
  - If not appropriate, we can fit a polynomial



Relationship between correlation and residuals

$$\frac{\text{SD of fitted values}}{\text{SD of y}} = |r|$$

The proportion of the total variability (SD y) accounted for by the regression line is |r|



$$(\text{SD y})^2 \;=\; (\text{SD residuals})^2 \;+\; (\text{SD fitted values})^2$$

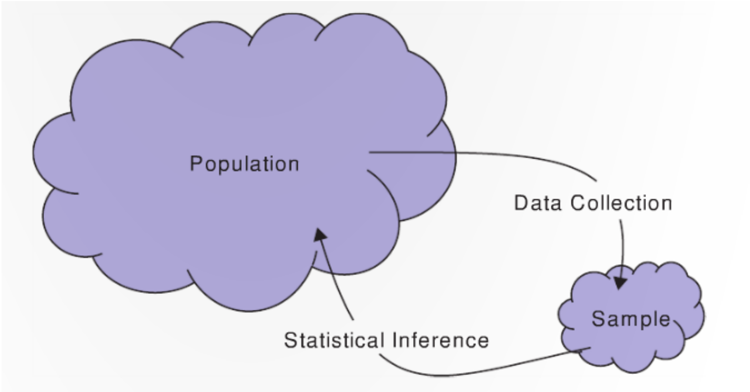
The more variability accounted for by the regression line (larger slope) the less variability left in the residuals

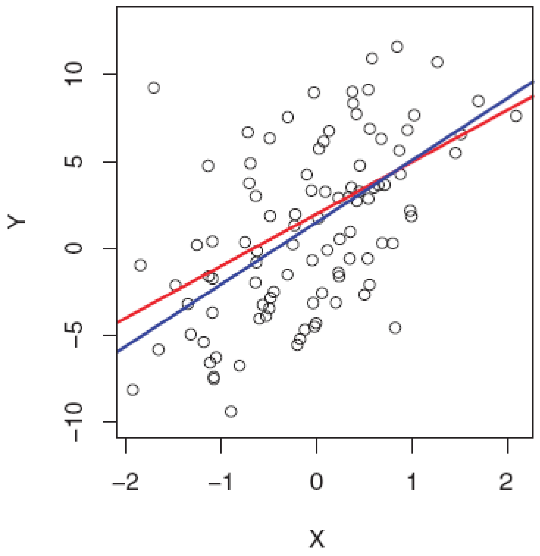
Inference for regression

Inference for regression

A regression line from a sample of data is only an estimate of the true population regression line

- i.e., if we had a different sample of data, we would get a different regression line





Population: regression lines

Sample estimates:
"lines of best fit" based on a sample of data

Q: How accurate is our "line of best fit" from a sample at capturing the true relationship?

Let's explore this in Jupyter!

Confidence interval for linear regression

Confidence interval for regression lines

We can use the bootstrap to create confidence intervals for:

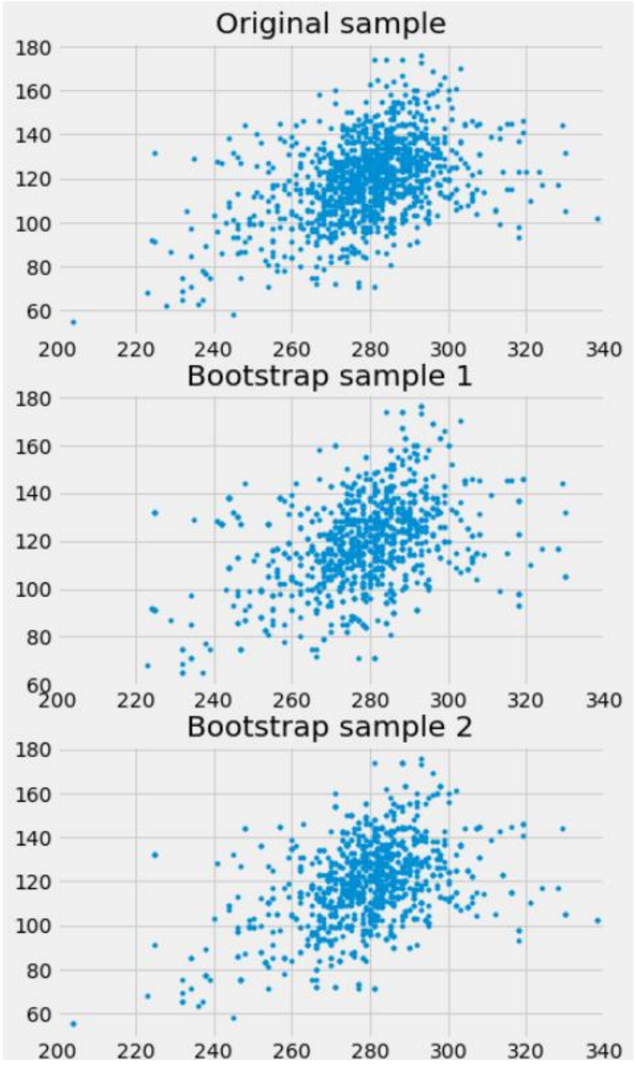
- The regression slope
- The regression intercept
- The whole regression line

To run the bootstrap we need to:

- Resample our data with replacement
- Fit a regression line to the sample of data
- Save the regression slope and intercept
- Repeat many times

To create a 95% confidence interval:

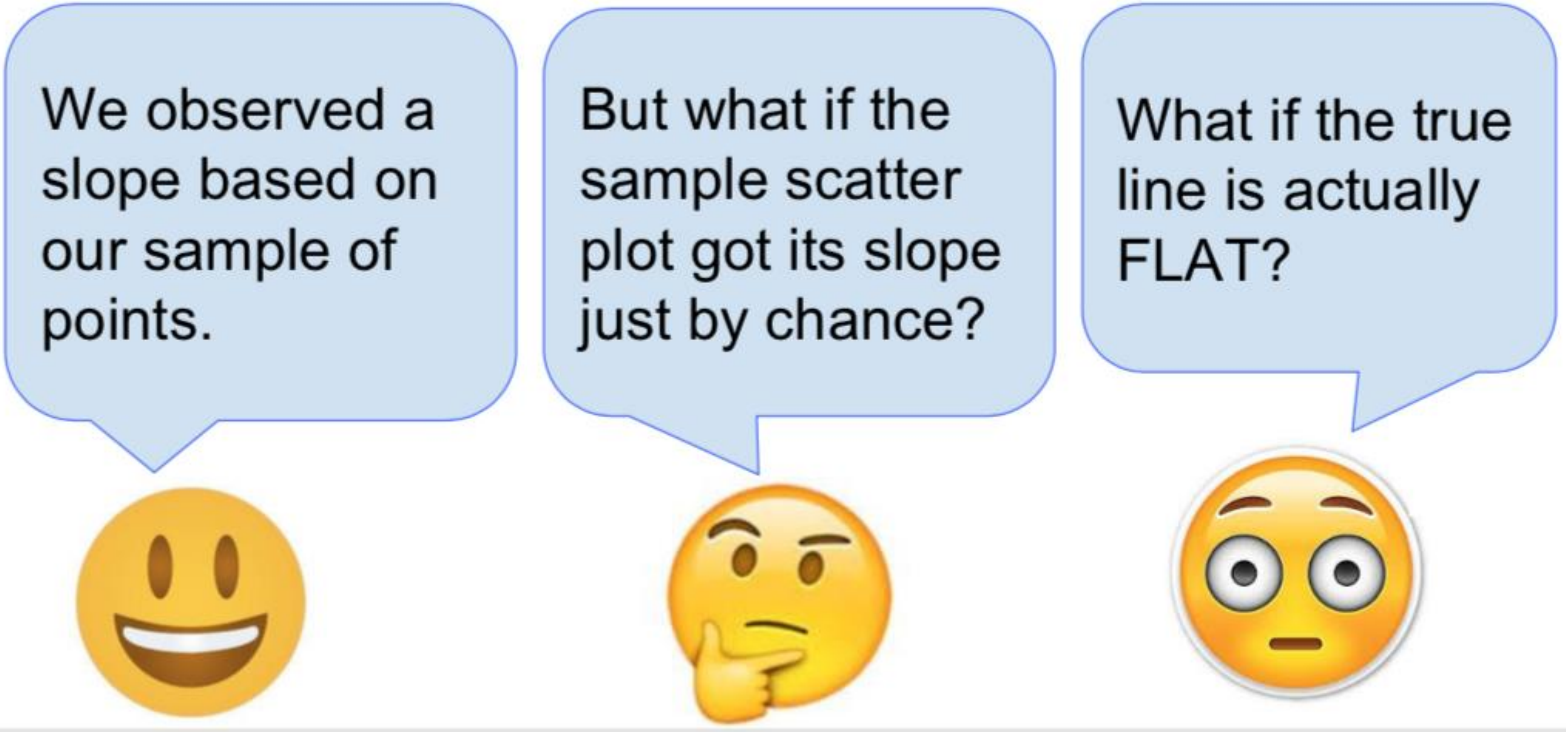
- Get the "middle 95%" of our regression slopes (or intercepts)



Let's explore this in Jupyter!

Hypothesis tests for linear regression

Rain on the regression parade



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

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:

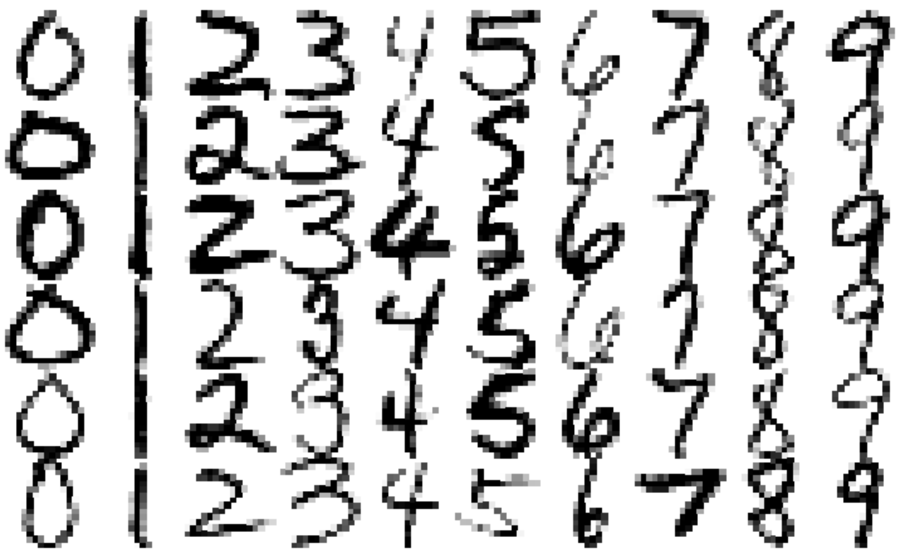
- Regression: output is a real number ($y \in R$)
- Classification: output is a categorical variable y_k

Example: salmon or sea bass?

What could be features in this task?

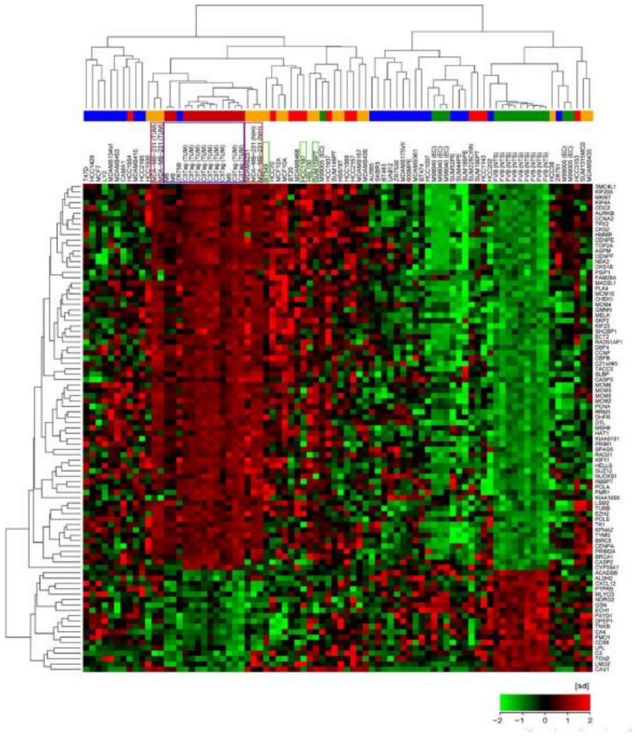
From Duda, Hart and Stork, 2001

Example: what is in this image?



What could be features in this task?

Example: predicting cancer



What could be features in this task?

Example: Fisher’s Iris data set

Setosa

Virginica

Vericolor

What could be features in this task?

Fisher, 1936

Example: GPT-3 predicting/generating text

Question answering:

Are we living in a simulation?

Image generation

"Draw an astronaut riding a horse"

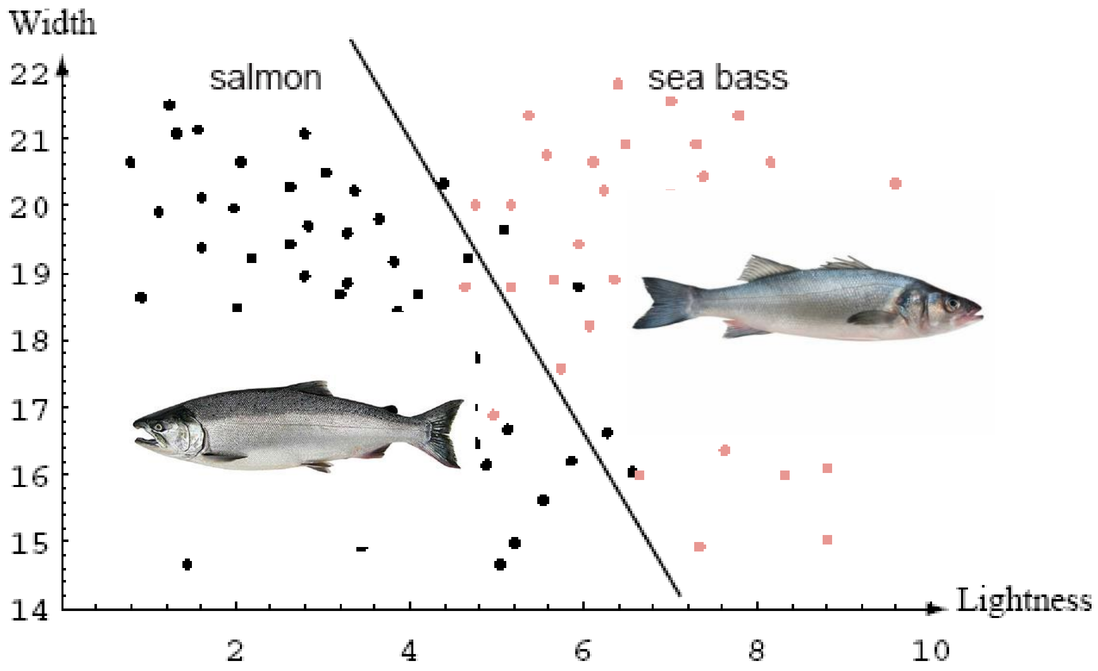
What could be features in this task?

A few key concepts

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

- E.g., f(pixel values) $\longrightarrow$ salmon or bass

It is linear if we can draw a straight line (or a multi-dimensional plane) between the two predicted values



Let's explore this in Jupyter!