Annotated Version

Machine Learning Course - CS-433

Regression

Sept 15, 2020

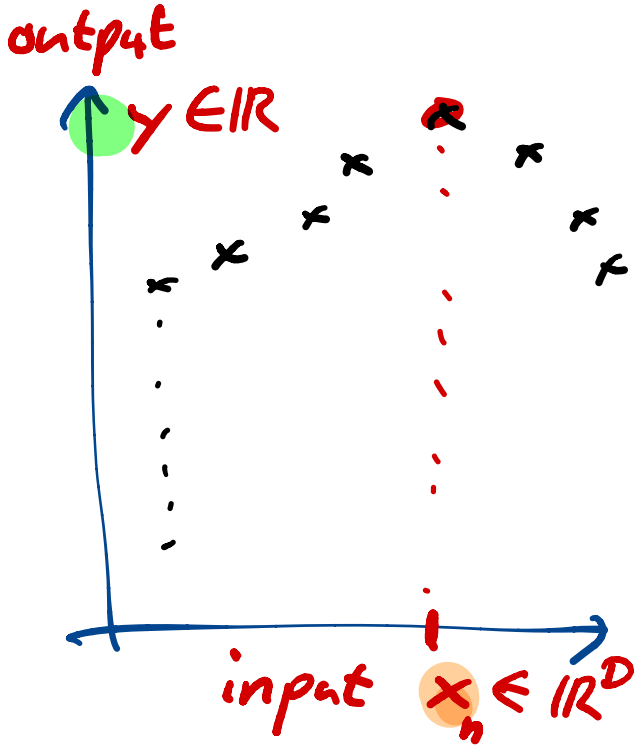
minor changes by Martin Jaggi 2020,2019,2018,2017,2016;

Last updated on: September 14, 2020

EPFL

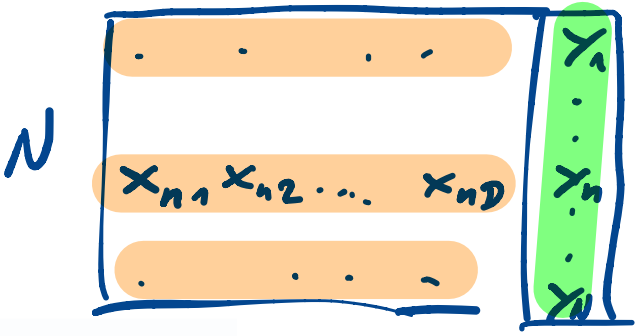
What is regression?

Regression is to relate input variables to the output variable, to either predict outputs for new inputs and/or to understand the effect of the input on the output.



$$\text{Data} = \{(x_n, y_n)\}_{n=1}^{N}$$

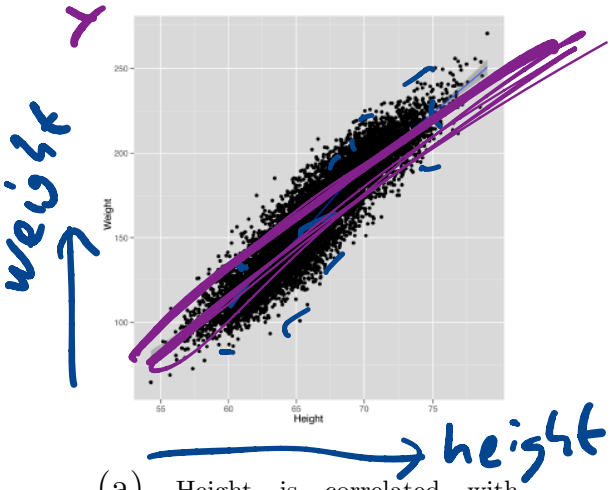
$x_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$



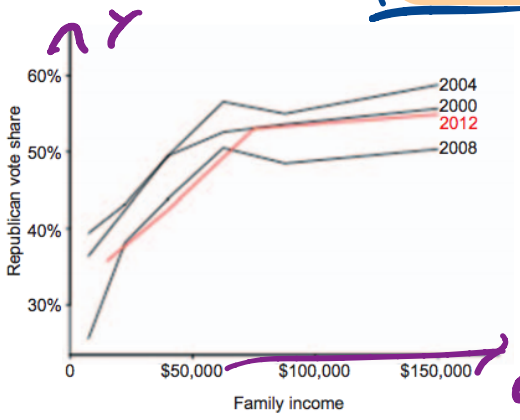
Dataset for regression

In regression, data consists of pairs $(\mathbf{x}_n, y_n)$, where y_n is the n'th output and $\mathbf{x}_n$ is a vector of D inputs. The number of pairs N is the data-size and D is the dimensionality.

Examples of regression

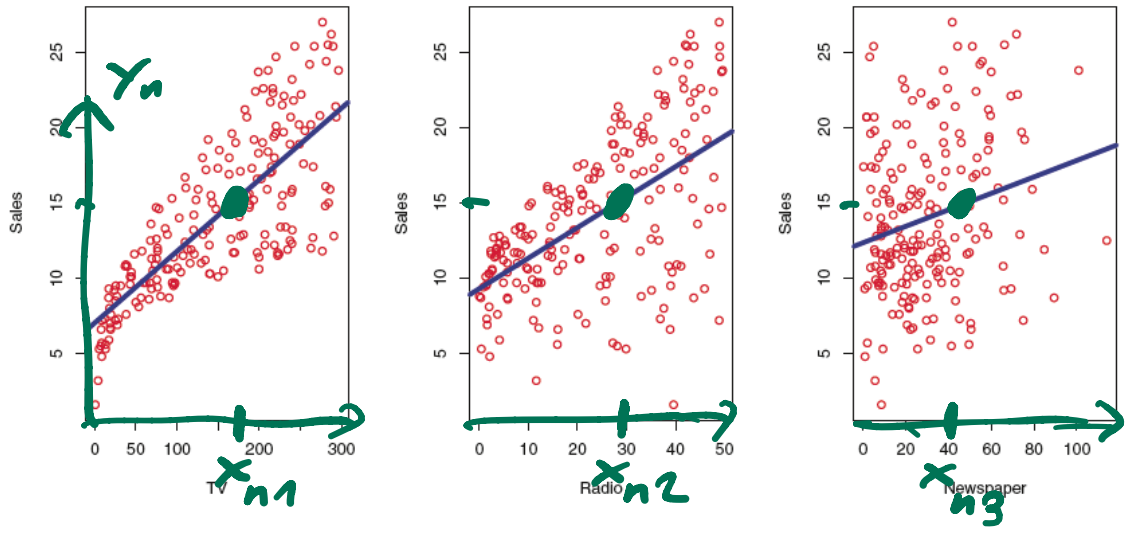


(a) Height is correlated with weight. Taken from "Machine Learning for Hackers"



(b) Do rich people vote for republicans? Taken from Avi Feller et. al. 2013, Red state/blue state in 2012 elections.

one-dimensional
D = 1



(c) How does advertisement in TV, radio, and newspaper affect sales? Taken from the book "An Introduction to statistical learning"

Two goals of regression

In prediction, we wish to predict the output for a new input vector, e.g. what is the weight of a person who is 170 cm tall?

In interpretation, we wish to understand the effect of inputs on output, e.g. are taller people heavier too?

The regression function

For both the goals, we need to find a function that approximates the output "well enough" given inputs.

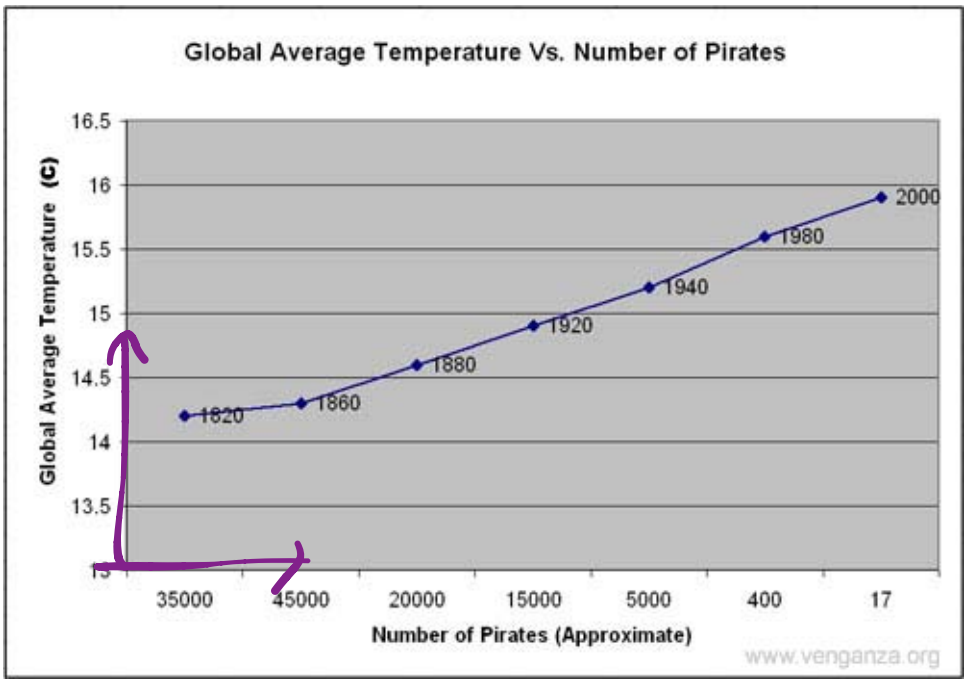
$$y_n \approx f(\mathbf{x}_n), \text{ for all } n$$

model, parameterized by $\mathbf{w}$

Additional Notes

Correlation $\neq$ Causation

Regression finds correlation not a causal relationship, so interpret your results with caution.



This image is taken from www.venganza.org. You can see many more examples at this page: Spurious correlations page.

Machine Learning Jargon for Regression

X_n *Input variables* are also known as features, covariates, independent variables, explanatory variables, exogenous variables, predictors, regressors.

Y_n *Output variables* are also known as target, label, response, outcome, dependent variable, endogenous variables, measured variable, regressands.

Prediction vs Interpretation

Some questions to think about: are these prediction tasks or interpretation task?

1. What is the life-expectancy of a person who has been smoking for 10 years?
2. Does smoking cause cancer?
3. When the number of packs a smoker smokes per day doubles, their life span gets cut in half?
4. A massive scale earthquake will occur in California within next 30 years.
5. More than 300 bird species in north America could reduce their habitat by half or more by 2080.