Can statistics help us to understand deep learning?

by

Hannes Smit

2019-03-27

Introduction

Machine Learning

- Successes
  - Self driving cars — safer than human drivers
  - Computer vision — facial recognition
  - Writing text
- Possible problems
  - Self driving cars — what if they crash?
  - Parole decisions in the U.S. — biases
  - GPT2 — the algorithm 'too dangerous to release'

The Black Box

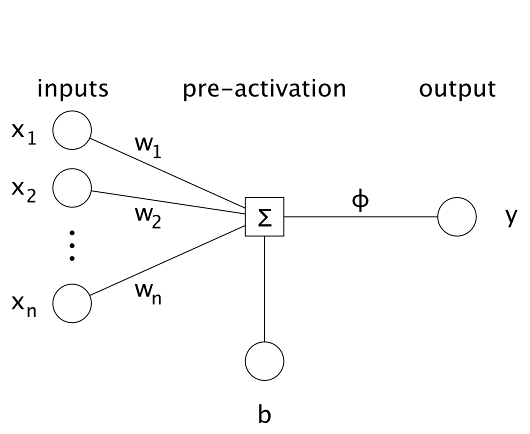
- Machine learning is a black box
- We need to be able to open the black box
- Put the algorithms' decisions into human understandable terms

Artificial Neural Networks

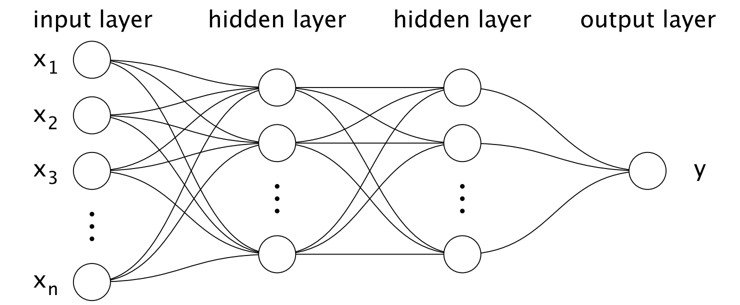
Neuron

1. Weighted sum
2. Bias term b
3. Nonlinear activation function ϕ
   - Identity (linear regression)
   - tanh
   - ReLU

$$y = \phi\left(b + \sum_{i=1}^{n} w_i x_i\right)$$



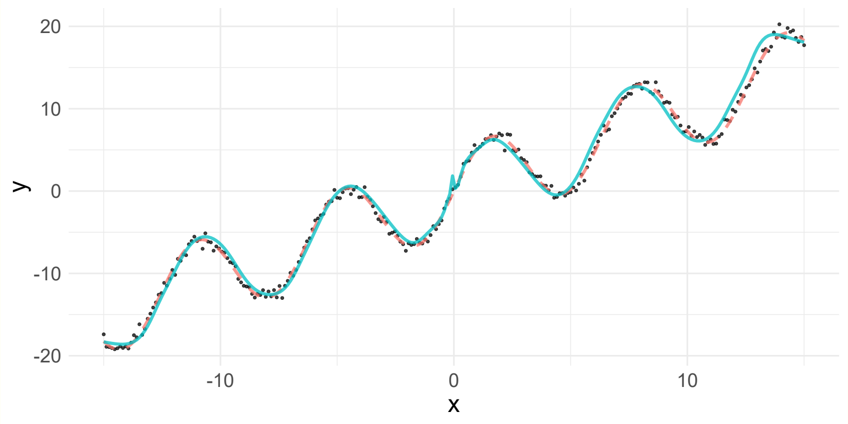
Backpropogation



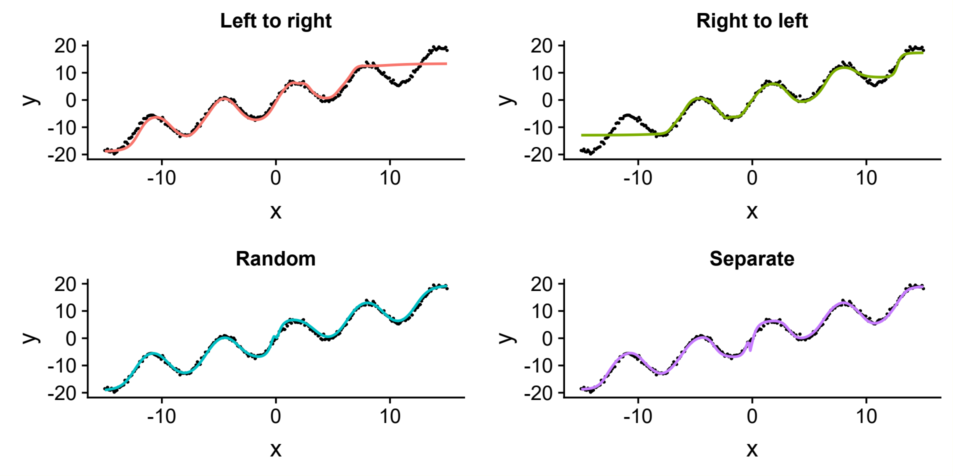
- Neural networks are arranged in layers
- Deep learning involves many layers
- Use gradient descent to optimise the weights
- In practice, stochastic batch gradient descent is used

Simple Example

$$y = x + 5\sin(x) + \epsilon \quad \epsilon \sim N(0, 0.5)$$



Training


Left to right
Right to left
Random
Separate
20
10
0
-10
-20
y
-10
0
10
x

Opening the Black Box

Current Research

Linear Regression

$$y = x + 5\sin(x) + \epsilon \quad \epsilon \sim N(0, 0.5)$$

$$\begin{aligned} y =& \alpha + \beta_0 x + \beta_1 x^2 + \\ & \beta_2 \sin(x) + \beta_3 \sin(2x) + \beta_4 \sin(x/2) + \\ & \beta_5 \cos(x) + \beta_6 \cos(2x) + \beta_7 \cos(x/2) + \epsilon \end{aligned}$$

Linear Regression Results

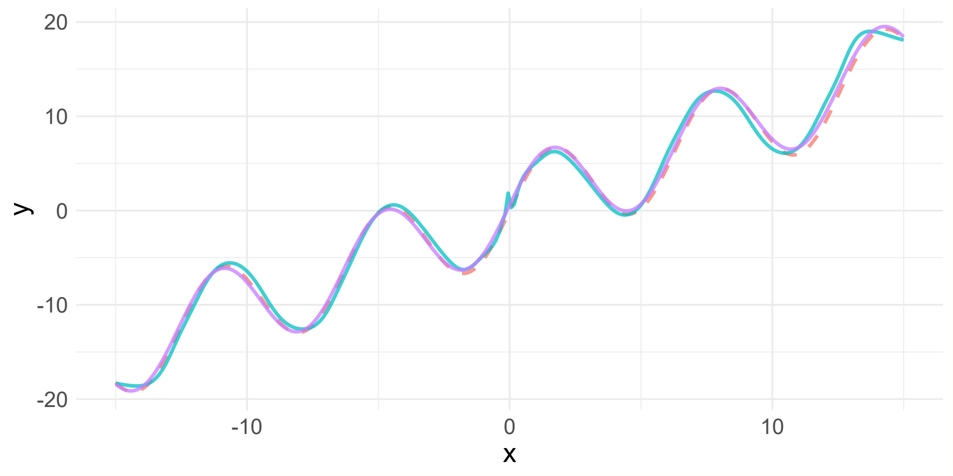
$$y = x + 5\sin(x) + \epsilon \quad \epsilon \sim N(0, 0.5)$$

	Estimate	Std. Error	t value	p value	
(Intercept)	0.1404	0.0695	2.02	0.0444	
x	1.0203	0.0053	193.86	0.0000	←
x^2	0.0006	0.0007	0.87	0.3843	
$\sin(x)$	4.7783	0.0643	74.31	0.0000	←
$\sin(2x)$	0.0720	0.0636	1.13	0.2588	
$\sin(x/2)$	0.0099	0.0662	0.15	0.8808	
$\cos(x)$	0.2785	0.0656	4.25	0.0000	←
$\cos(2x)$	0.0654	0.0647	1.01	0.3130	
$\cos(x/2)$	0.0901	0.0705	1.28	0.2028	

Stepwise Regression

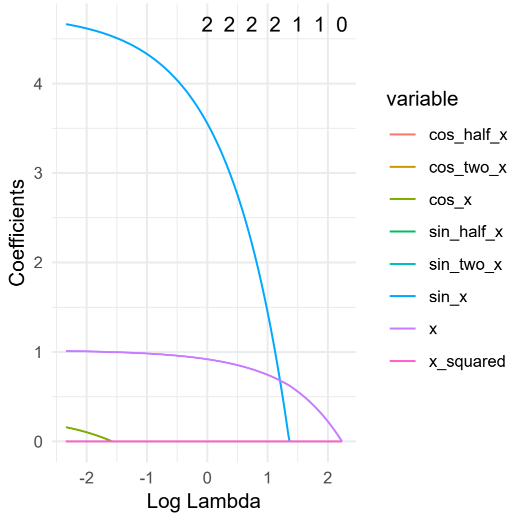
	Estimate	Std. Error	t value	p value
(Intercept)	0.1840	0.0458	4.02	0.0001
x	1.0201	0.0052	195.05	0.0000
$sin(x)$	4.7815	0.0633	75.54	0.0000
$cos(x)$	0.2864	0.0652	4.40	0.0000
$cos(x/2)$	0.1146	0.0637	1.80	0.0732

Stepwise Regression



Lasso

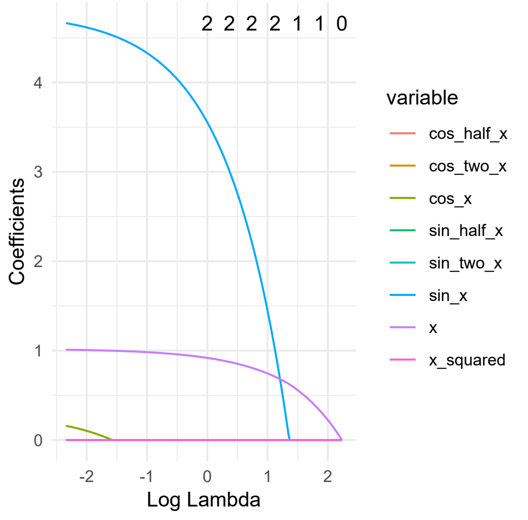
- Least Absolute Shrinkage and Selection Operator
- Constrains the sum of the absolute values of the model parameters
- Regularises the least influential parameters to zero
- Use k-fold cross validation to find the optimal hyperparameter λ



Lasso Results

Result of 10 fold CV: $\lambda = -2.232497$

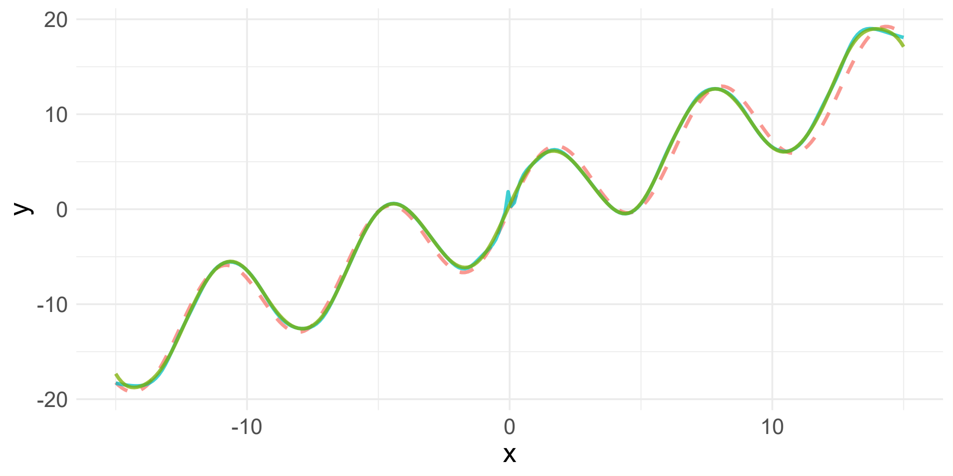
variable	estimate
(Intercept)	0.20
x	1.01
$\sin(x)$	4.64
$\cos(x)$	0.13



Gaussian Processes

- A Gaussian Process is the infinite dimensional analogue to the multivariate normal distribution
- Instead of a mean vector it uses a mean function and instead of a covariance matrix it uses a covariance function
- GPs have statistical properties that are understood

Using a Gaussian Process



Future Research

- Using these techniques on more complex and realistic problems
- Use the Lasso method as a mean function for a Gaussian process
- Sensitivity analysis can also find significant variables

Thank you for listening