Life as an ML Engineer

David Rasch - Infinia ML

October 30, 2018

Intro

Things you already know

1. Interchangable Parts

Things you already know

1. Interchangable Parts
2. Testing

Things you already know

1. Interchangable Parts
2. Testing
3. Integration

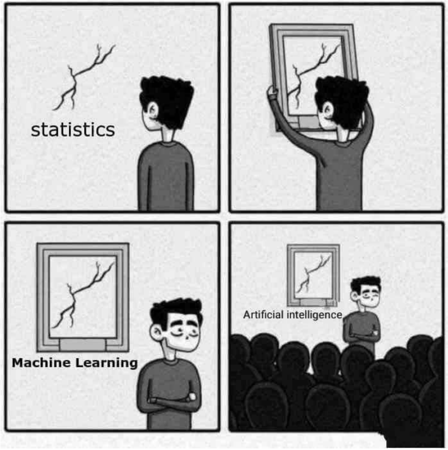
Objectives

1. Think about ML from an engineering perspective

Objectives

1. Think about ML from an engineering perspective
2. Learn some of the terminology used to help converse between Data Scientists and Engineers like:

ai vs statistics



you're going to need some data

you need to know what you're trying to do

- user stories

you need to know what you're trying to do

- user stories
- business problem?

you need to know what you're trying to do

- user stories
- business problem?
- black box function

picking your algorithm

- look at your data

picking your algorithm

- look at your data
- look at your inputs

picking your algorithm

- look at your data
- look at your inputs
- look at your outputs

picking your algorithm

- look at your data
- look at your inputs
- look at your outputs
- phone a friend

picking your algorithm

- look at your data
- look at your inputs
- look at your outputs
- phone a friend
  - scikit learn flow chart

picking your algorithm

- look at your data
- look at your inputs
- look at your outputs
- phone a friend
  - scikit learn flow chart
  - or just use deep learning, it's cool

picking your algorithm

- look at your data
- look at your inputs
- look at your outputs
- phone a friend
  - scikit learn flow chart
  - or just use deep learning, it's cool
- interpretability

don't forget to look for prior art

- Look at UNet, YOLO, ResNet51, RetinaNet, and many other hyped algorithms.

don't forget to look for prior art

- Look at UNet, YOLO, ResNet51, RetinaNet, and many other hyped algorithms.
- Tensorflow has many sets of "pre-trained" weights

this was a whole section on data prep

- new API

this was a whole section on data prep

- new API
- new CSV from a customer

things that matter for ML

- normalizing or "whitening"

things that matter for ML

- normalizing or "whitening"
- binning

things that matter for ML

- normalizing or "whitening"
- binning
- missing values

things that matter for ML

- normalizing or "whitening"
- binning
- missing values
- dimensionality reduction

things that matter for ML

- normalizing or "whitening"
- binning
- missing values
- dimensionality reduction
- class imbalance

algorithms

pre-jargon

- letters

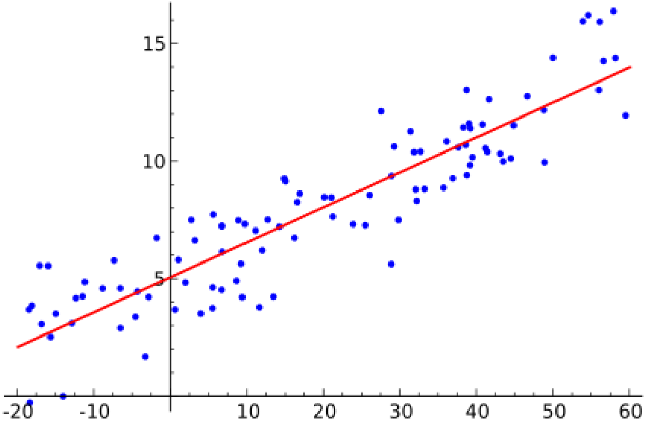
pre-jargon

- letters
- $Y = mx + b$

pre-jargon (cont'd)

$$Y = Wx + b$$

regression



what if there are multiple variables?

- $y = W_1 x_1 + b$

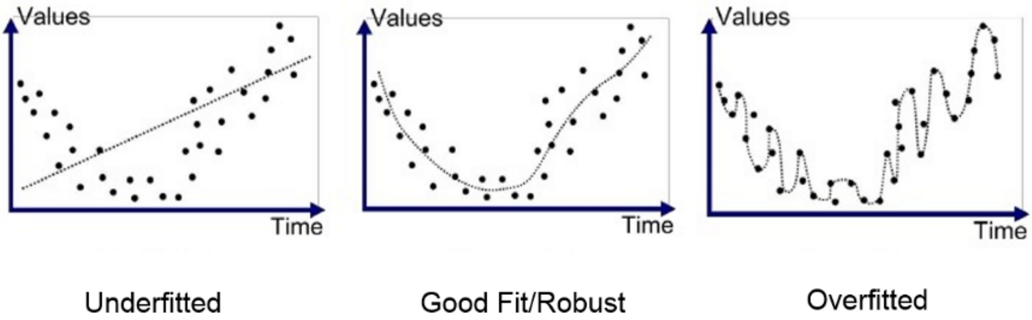
what if there are multiple variables?

- $y = W_1 x_1 + b$
- $y = W_1 x_1 + W_2 x_2 + \ldots + b$

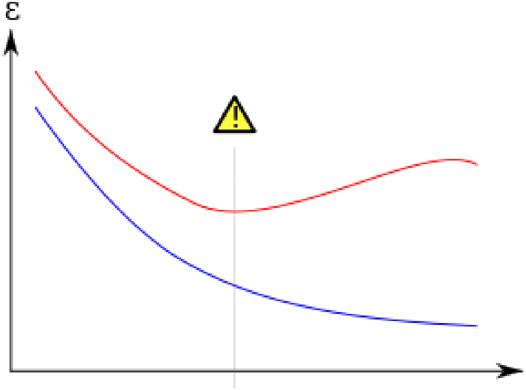
what if there are multiple variables?

- $y = W_1x_1 + b$
- $y = W_1x_1 + W_2x_2 + \ldots + b$
- $y = Wx + b$

overfitting



overfitting



data requirements

- large data

data requirements

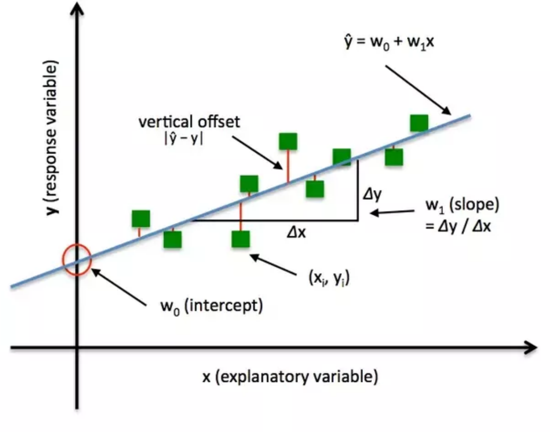
- large data
- data hacks

data requirements

- large data
- data hacks
  - data augmentation - zoom, rotate, flip images

gradient descent

losing it (loss function)



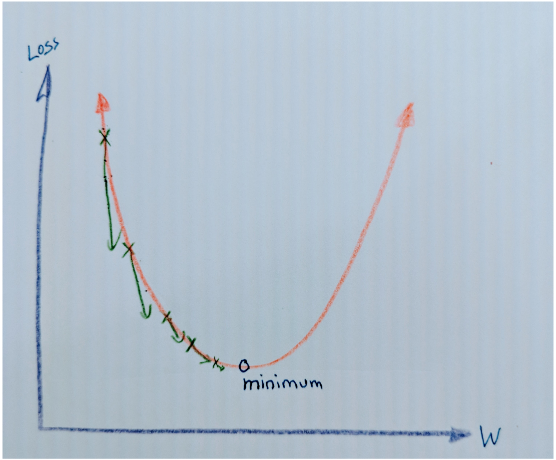
little bit of math

$$\sum_{x}(Wx + b - y_x)^2$$

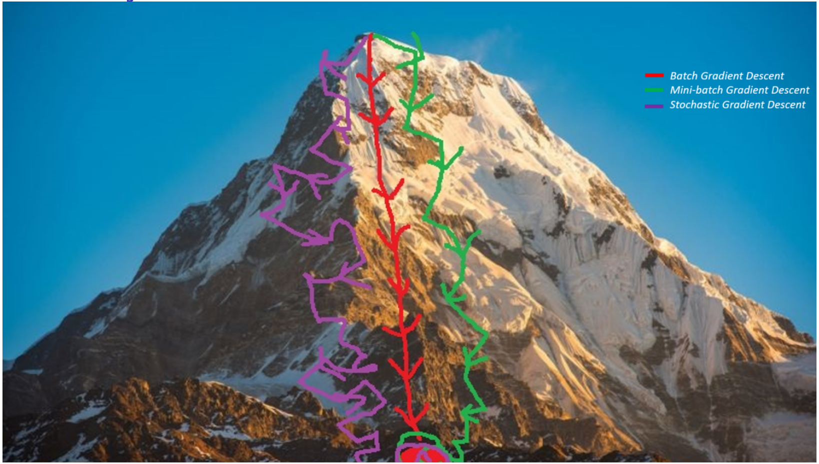
little bit more math

$$\arg\min_{W,b} \sum_{x} (Wx + b - y_x)^2$$

gradient descent



stochasticity, batches, and mini-batches



inference aka "pushing to production"

trained model

- what is truth?

trained model

- what is truth?
- testing?

trained model

- what is truth?
- testing?
- what can go wrong?

trained model

- what is truth?
- testing?
- what can go wrong?
- "master" branch?

scaling (performance, speed)

- easy

scaling (performance, speed)

- easy
- well defined interfaces

scaling (performance, speed)

- easy
- well defined interfaces
- shared nothing

scaling (performance, speed)

- easy
- well defined interfaces
- shared nothing
- load balancing

model health

- what if incoming data is different than training data?

model health

- what if incoming data is different than training data?
  - e.g., hot dog vs not hot dog, and someone gives it a brautwurst

model health

- what if incoming data is different than training data?
  - e.g., hot dog vs not hot dog, and someone gives it a brautwurst
  - or a real example, kangaroos on self driving cars

Operations

- get new data! prompt users for wrong responses

Operations

- get new data! prompt users for wrong responses
- online learning: re-train nightly/hourly/steaming w/ new data

Operations

- get new data! prompt users for wrong responses
- online learning: re-train nightly/hourly/steaming w/ new data
- active learning: figure out what labels you need to improve model performance

tensors and flow graph

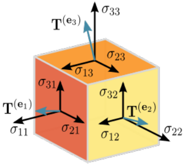
tensors

- linear relation between vectors, scalars, or other tensors

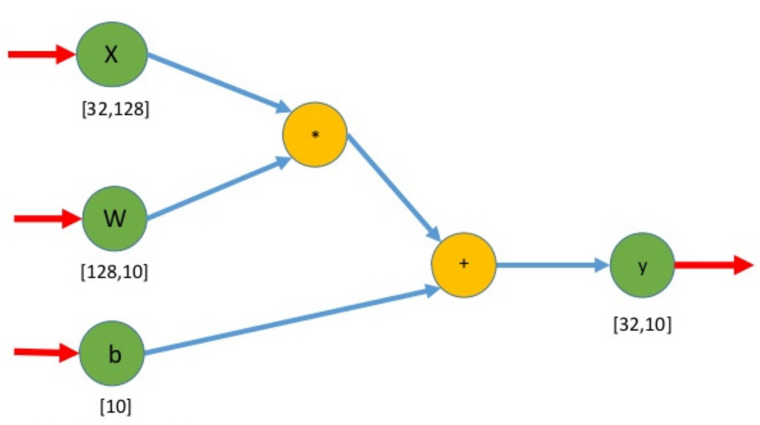
tensors

- linear relation between vectors, scalars, or other tensors
- practically: multi-dimensional array

tensors

- linear relation between vectors, scalars, or other tensors
- practically: multi-dimensional array
- 

computational flow graph (Directed-acyclic graph)



questions?

other resources

other learning resources

- http://fast.ai
- https://hackernoon.com/choosing-the-right-machine-learning-algorithm-68126944ce1f
- http://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html

image credits

- ai vs stats
- regression
- overfitting
- more overfitting
- loss functions
- gradient descent
- tensors
- tensorflow graph