Machine Learning

what, why and how

me ...

Harshad

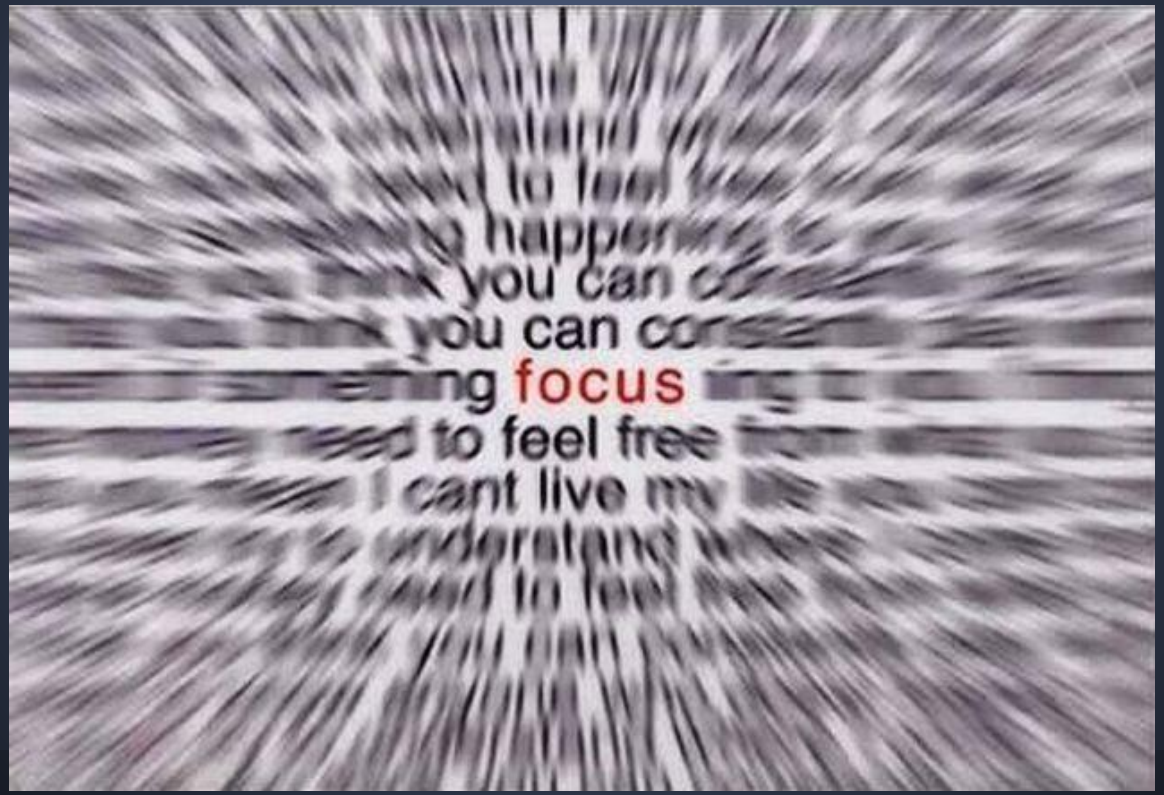
- → Senior Data Scientist @ Sokrati
- → Spent last 4 years trying to understand and apply machine learning

Sokrati is a digital advertising startup based out of Pune

what are we going to do ?

get a 10000 feet view ...

then go to specifics …



10000 feet view

what is ML ?

- Too many definitions!
- Too much debate
- Analytics vs ML vs data mining vs Big Data vs Statistics vs next buzzword in the market

what is ML ?

Teaching machines to take decisions with the help of data

practical man’s definition of machine learning!

bit of history …

That’s too ancient!

That’s not ML...

bit of (relevant) history ...

- Insurance, Banking industry
  - Credit scoring
  - mathematical models in finance
- Artificial Intelligence and other fancy ideas
  - deep blue, Samuel's checkers machine
  - IBM watson computer

$$g(\, y \,) = f(\, \mathbf{X} \,)$$

Find ‘f’ => endeavour to understand the world!

the two cultures in ML

Stats Culture

Vs

AI / ML Culture

the two cultures in ML

Stats Culture

- ➔ Focus on ‘why this model’
- ➔ Goodness of fit, hypothesis tests, residuals
- ➔ or MCMC methods, bayesian modeling
- ➔ regression models, survival analysis

AI / ML Culture

- ➔ Focus on ‘good predictions’
- ➔ Cross validations, ensemble of models
- ➔ Focus on underfit vs overfit analysis
- ➔ neural nets, tree based models (random forests et. al.)

but let’s build bridges

Stats

- ➔ Focus on basics, sound theory
- ➔ Exploration, summaries
- ➔ Models

ML / AI

- ➔ Focus on predictions
- ➔ Model evaluation
- ➔ Feature selection

Computations

- ➔ Focus on application
- ➔ Achieving scale and usability
- ➔ Hadoop , Storm and friends..

Business Knowledge

- ➔ Focus on interpretation
- ➔ Visualizations
- ➔ Creating stories out of data

typical ML process

- → Objective
- → Source data
- → Explore
- → Model
- → Evaluate
- → Apply
- → Validate

objective

in brief!

probability of customer churn

group set of emails by topic

predict rainfall

recommend item to a consumer to increase likelihood of click

bottom line : not in terms of algorithm but outcome

sourcing data

to be covered at end!

explore

R you ready ?

introduction to R

- Starting R
- Data Structures
  - Atomic Vectors, matrix
  - Lists
  - Data frames
- Data types
  - usual suspects in numerics (int, double, character)
  - Factors
  - logical

data frame , the workhorse

- Load sample data frame
- Explore data frame (head, tail)
- Access elements by index
  - access rows
  - access columns
    - single
    - multiple (by name, by index, by -ve index)
- Find metadata
  - names
  - dimensions
- Explore using plot (pairs)

broadcasting / vectorization

- Very important concept
- Subsetting
  - vectors
  - data frames
- Applying operations
  - operations on entire column

data exploration

- summarizing using mean
- quantiles, when mean is not enough
  - outlier detection
- functional roots of R : sapply summary
- summary function
  - on numerical
  - on factors
- plots (basic)
- histograms
- correlations

models

simple & real world

basic types of models

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

linear regression

- load sample dataset (cars)
- build linear regression model
- understand the output
  - summary
  - plots
- understanding train vs predict cycle
  - most important idea!

demo on real world dataset

data exploration, classification

hierarchical clustering

- basic idea of clustering
  - distance as a proxy for similarity, group by distance
  - group anything as long as distance can be calculated
- load and explore eurodist data
- fit hierarchical cluster
- plot dendrogram

demo on real world data

and the most important idea!

concept of vector space model

- Words as axis
- Bag of words defines vector space
- Document as a point in space
- We can
  - define distance
  - measure similarity (cosine similarity)
  - group documents
- what can be a document ?

evaluation

evaluation metrics

- depends on type of model
  - regression : MAPE, MSSE
  - classification : accuracy, precision, recall, F score
  - clustering : within vs between variance
- ML world (ref : two cultures) has much better story
- Not enough to perform well on training set

brief intro regularization

- Bias vs Variance problem
- We want to be ‘just right’
- Concept of regularization
- Intro Cross validation

demo of evaluation

and fantastic Scikits API

sourcing and applying

and the great ML divide

the great ML divide

Lab Culture

- → Theory
- → Small Datasets
- → In memory
- → Not live
- → R, Octave, Python..

Source and Apply

- → The practice
- → Huge datasets
- → Live in production
- → Hadoop and friends, Python ? , R ?

processing data at scale

- Data is not available in final form
- Non standard data
  - click streams
  - event logs
  - free form text
- Process at scale
- Transform , clean, group in final form

5 min intro to Hadoop Ecosystem

- Write in assembly
  - java
- DSLs
  - Pig
  - hive
  - impala
- Functional Languages to rescue!

Introduction to Cascalog

processing data at scale

Recap

- Pragmatic ML
- Key phases
- Supervised learning
- Unsupervised learning
- Evaluation
- Application issues
- Processing data @ scale

Let’s discuss…