Women Techmakers

Powered by GDG Bangalore

December Meetup

Learning AI Development with UX

Who are we?

Arjun Rao
NLP Engineer at Stride.ai Inc

Anisha Mascarenhas
Software Engineer at LinkedIn

Agenda

1. The ML Mindset: Identifying good problems for AI
2. Hands on: Gather your ingredients for ML
3. A Text Classification Cookbook
4. Wrapping Up: Quirks of Real World AI Development

The ML Mindset

What is AI?

At the core of every computer program there is a mathematical function at work. It could be as simple as computing the interest on an outstanding loan or as complex as flying an aircraft on autopilot. *Artificial Intelligence*, or *AI*, is a generic name for a computer program whose core mathematical function has been created (almost) automatically; and *Machine Learning*, or *ML*, refers to a collection of techniques which offer ways of creating AI.

Namit Chaturvedi
(PhD in theoretical computer science,
Applied Research Engineer at LinkedIn)

AI can only be as good as the examples and techniques used to train it

Thinking about a problem from a ML Perspective: From programs to experiments

Step	Example
1. Set the research goal.	I want to predict how heavy traffic will be on a given day.
2. Make a hypothesis.	I think the weather forecast is an informative signal.
3. Collect the data.	Collect historical traffic data and weather on each day.
4. Test your hypothesis.	Train a model using this data.
5. Analyze your results.	Is this model better than existing systems?
6. Reach a conclusion.	I should (not) use this model to make predictions, because of X, Y, and Z.
7. Refine hypothesis and repeat.	Time of year could be a helpful signal.

Identifying good problems for ML

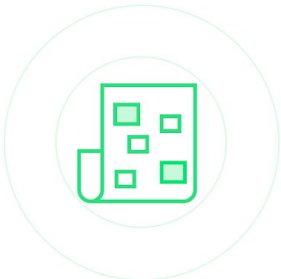
Start with the problem, and not the solution

Identifying good problems for ML: Aim to make decisions, not just predictions.

"I trained a model that predicts the probability that someone will want to watch a video and still click "thumbs down" on youtube!"

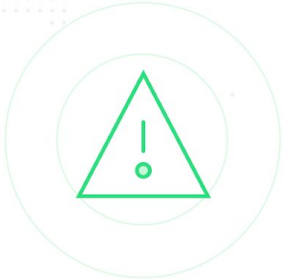
When is traditional computing better than machine learning?

Not enough data



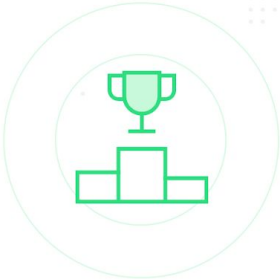
Noisy Data

No time & money



Simple problem to solve

Problems to solve with machine learning

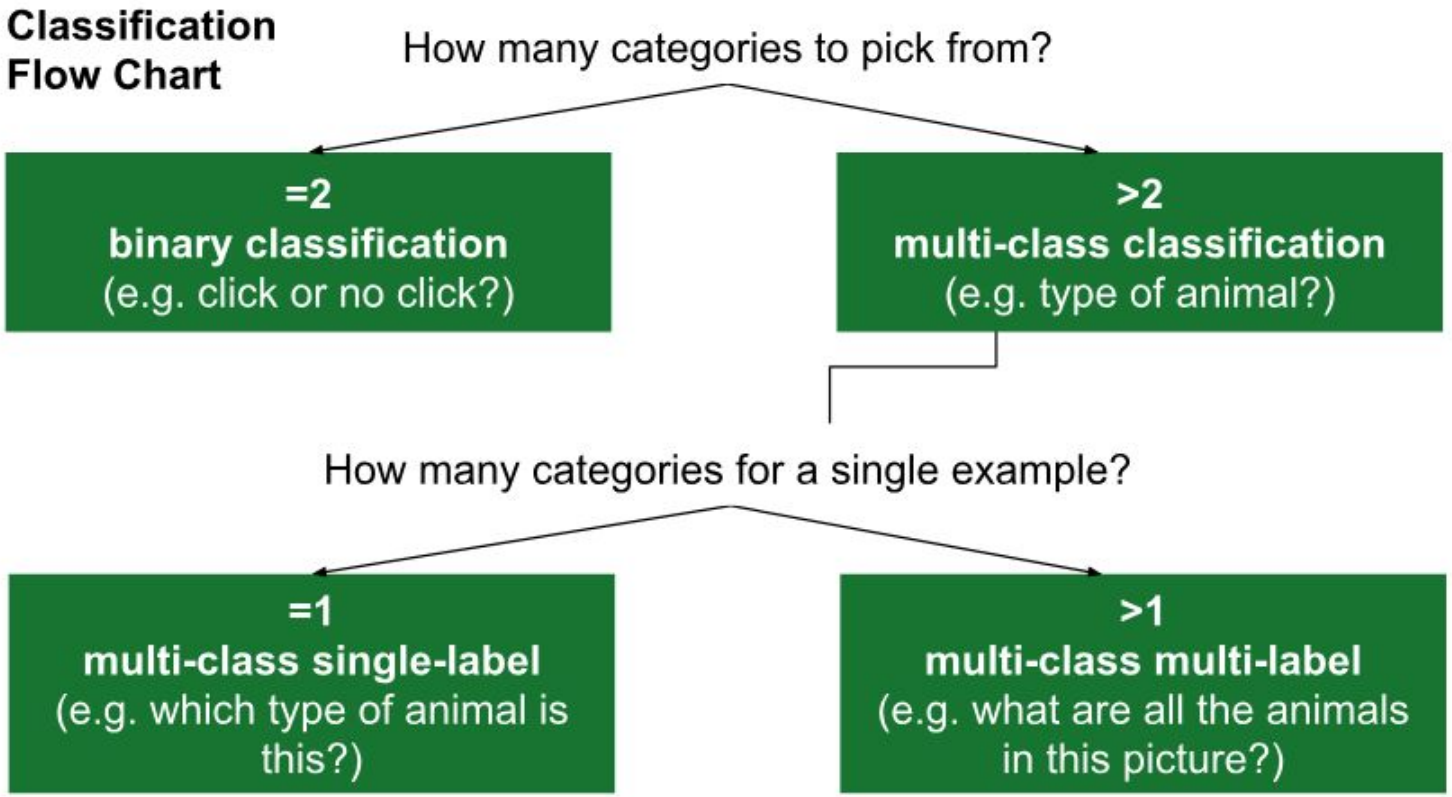


Classification

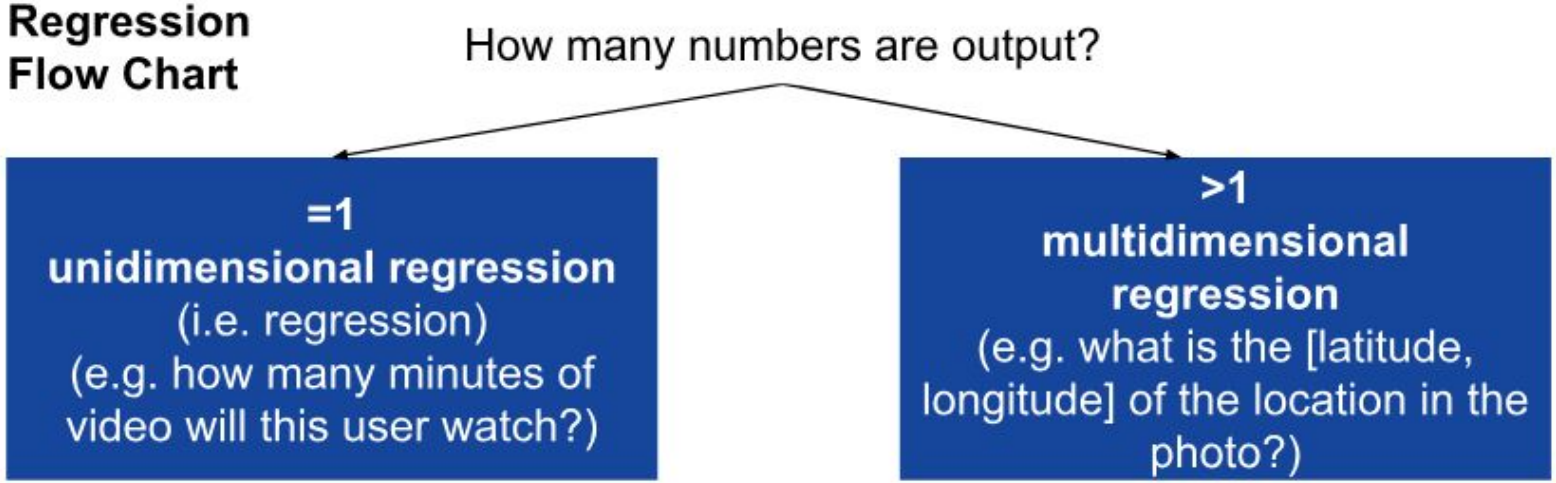
Clustering

Regression

Dimensinality Reduction


Classification
Flow Chart
How many categories to pick from?
=2
binary classification
(e.g. click or no click?)
>2
multi-class classification
(e.g. type of animal?)
How many categories for a single example?
=1
multi-class single-label
(e.g. which type of animal is this?)
>1
multi-class multi-label
(e.g. what are all the animals in this picture?)

Regression Flow Chart



Hands On

Building a good dataset

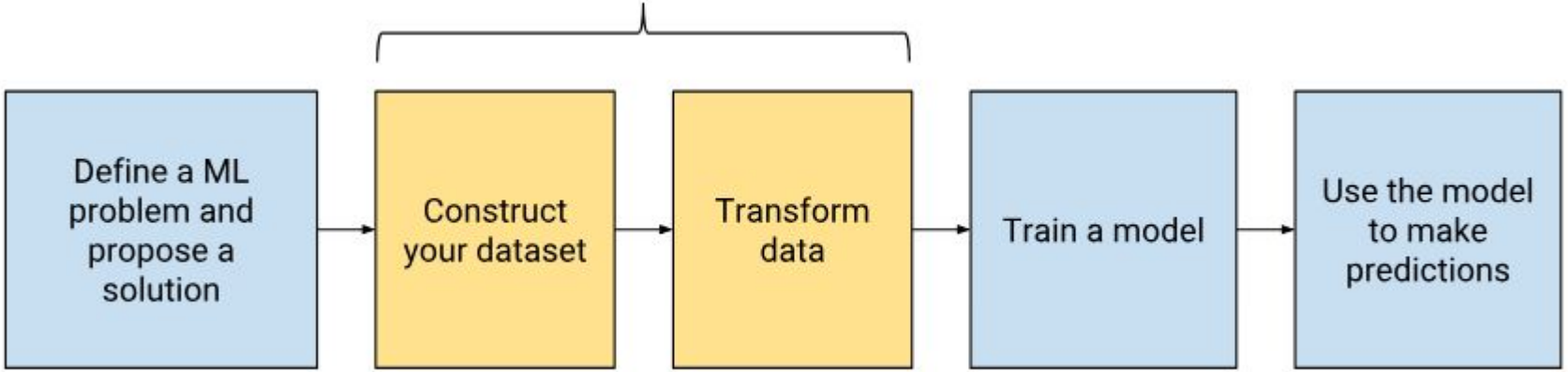
02

Link to Dcult -
https://tinyurl.com/altimetrik-dcult

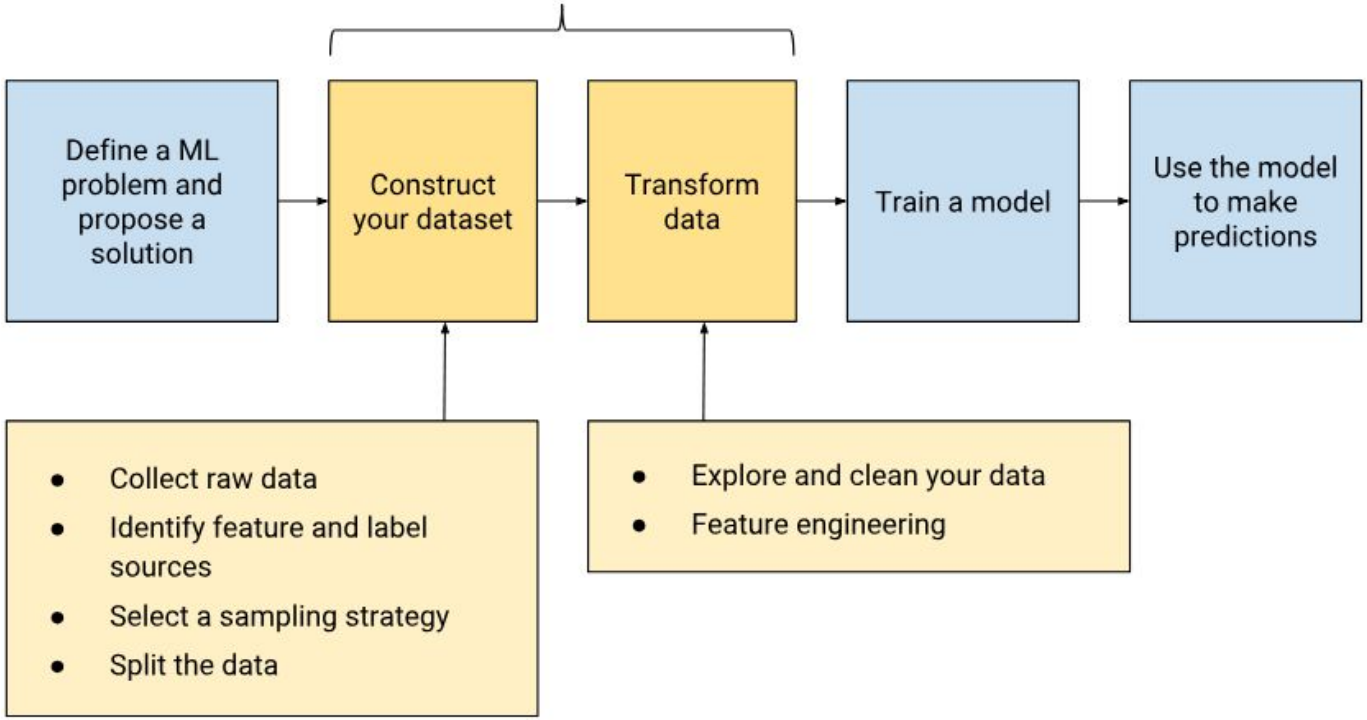
https://dcultaltimetrik.com

Link to Colab - https://goo.gl/tVgzsQ

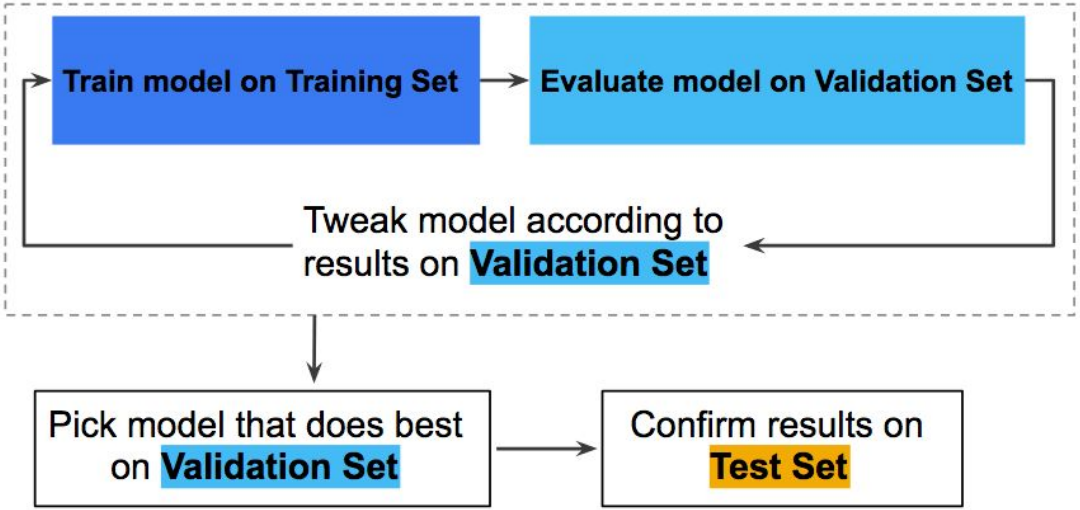
Preparing your dataset



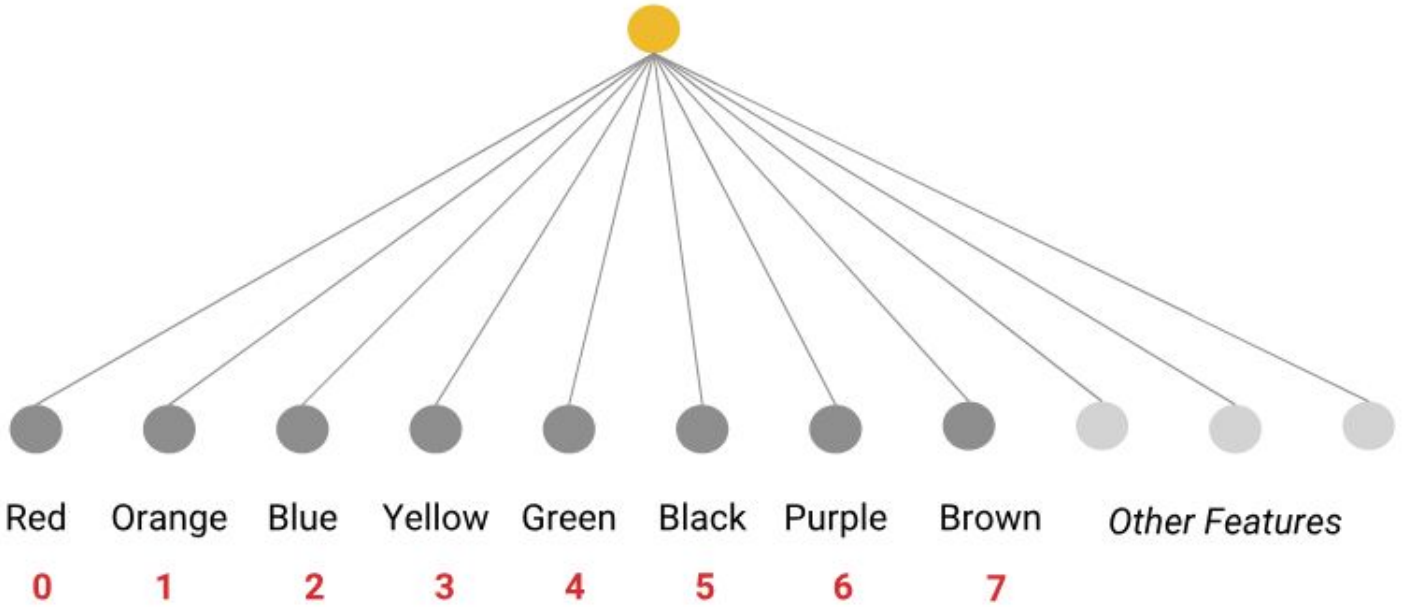
Preparing your dataset



Splitting your dataset:



Transforming data to features:
Example for dealing with Categorical Data

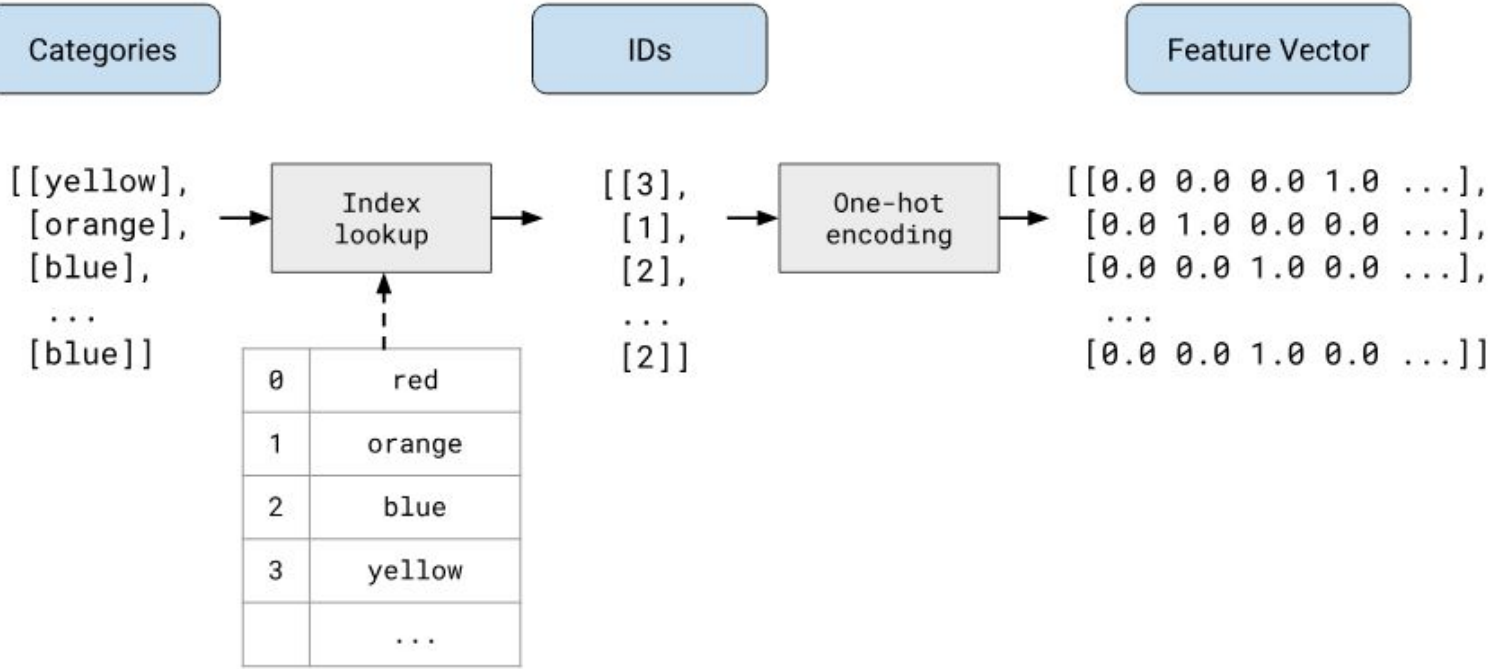


Transforming data to features: Building a vocabulary

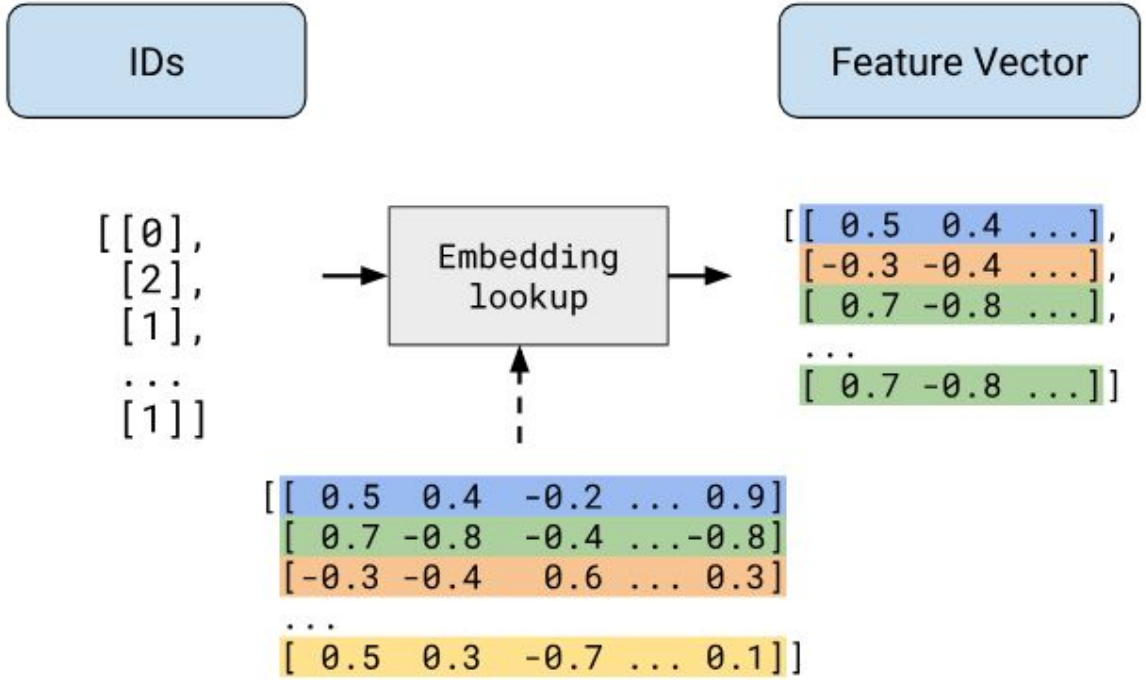
In a vocabulary, each value represents a unique feature.

Index Number	Category
0	Red
1	Orange
2	Blue
...	...

Transforming data to features: Mapping categories to feature vectors



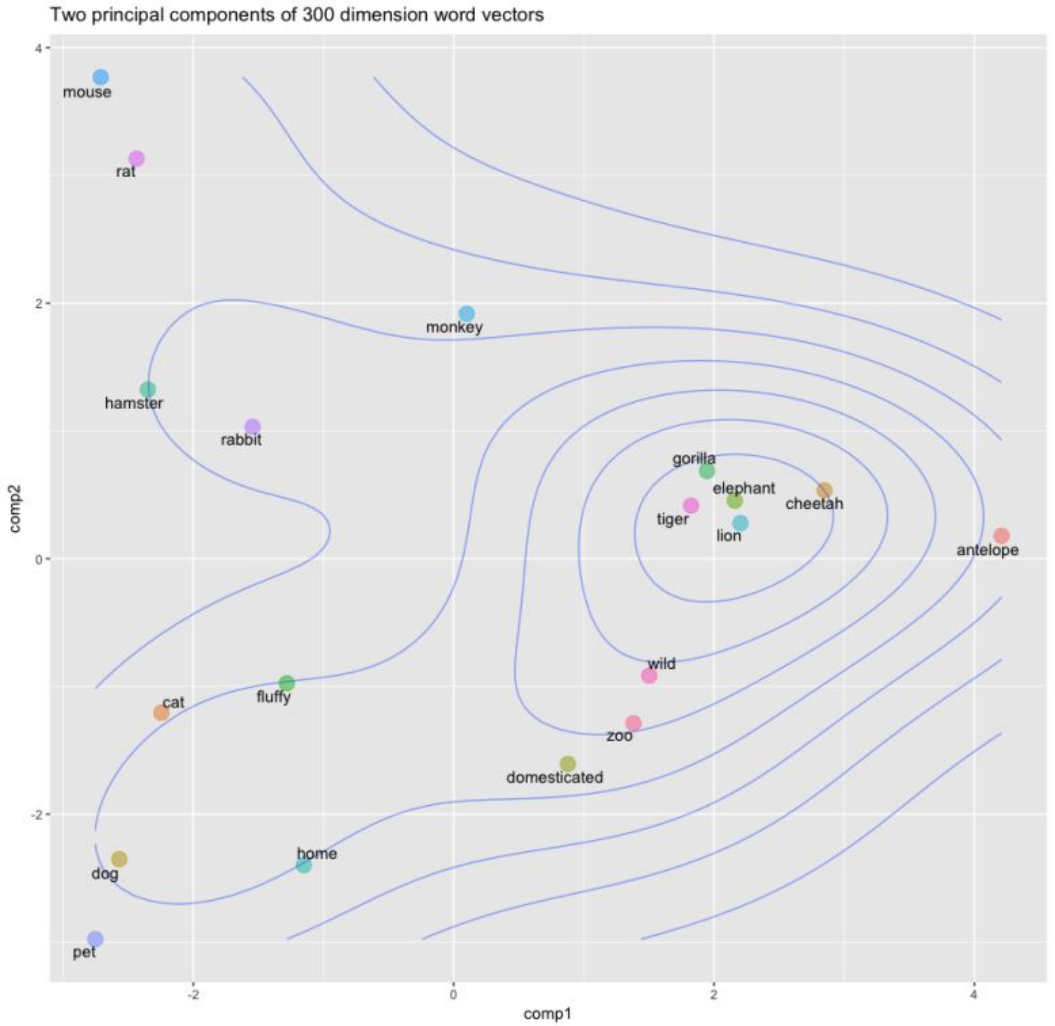
Transforming data to features:
Mapping categories to continuous valued feature vectors



Transforming data to features:
Mapping categories to continuous valued feature vectors

Word vectors	Dimensions			
dog	-0.4	0.37	0.02	-0.34
cat	-0.15	-0.02	-0.23	-0.23
lion	0.19	-0.4	0.35	-0.48
tiger	-0.08	0.31	0.56	0.07
elephant	-0.04	-0.09	0.11	-0.06
cheetah	0.27	-0.28	-0.2	-0.43
monkey	-0.02	-0.67	-0.21	-0.48
rabbit	-0.04	-0.3	-0.18	-0.47
mouse	0.09	-0.46	-0.35	-0.24
rat	0.21	-0.48	-0.56	-0.37

Visualizing Embeddings



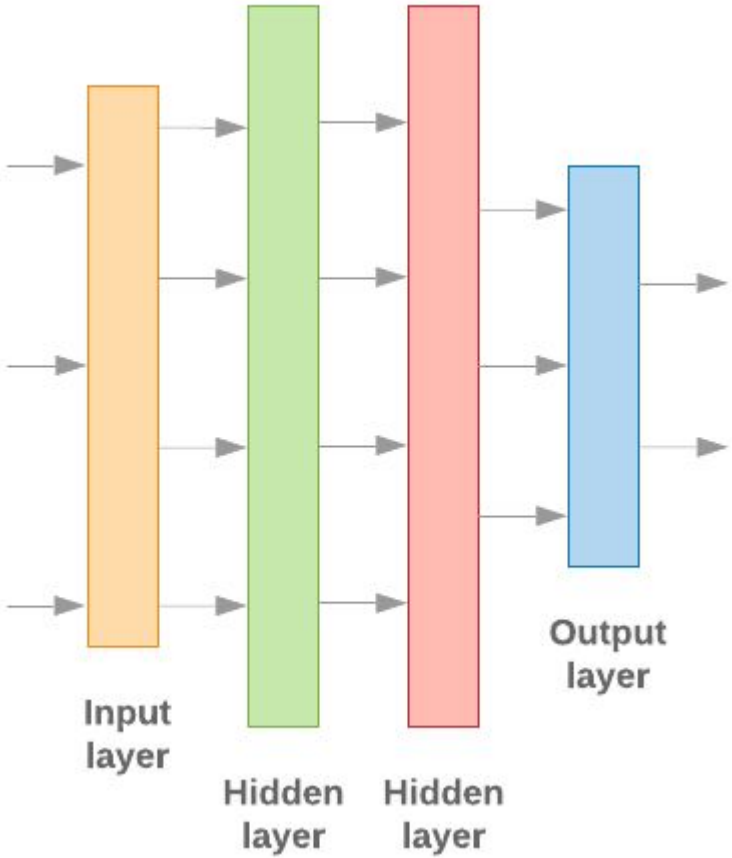
Questions?

Hands On

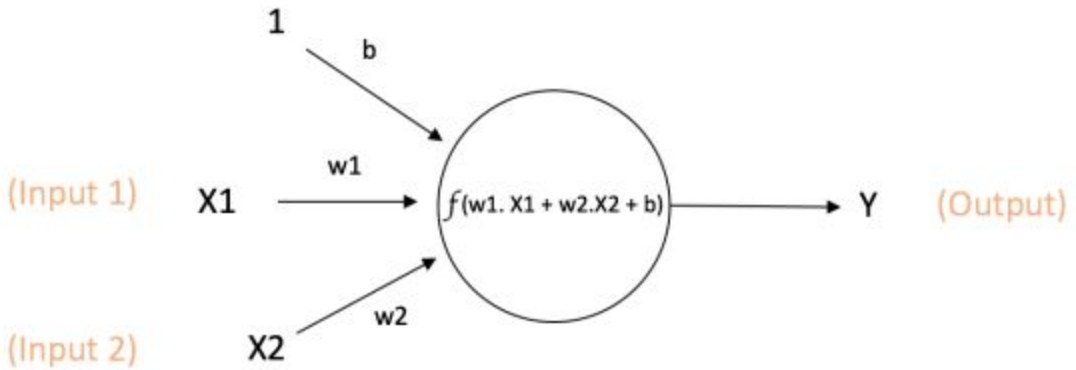
A Text Classification Cookbook

03

A note on neural networks

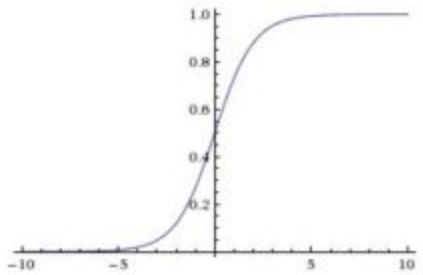


A note on neural networks: A single neuron

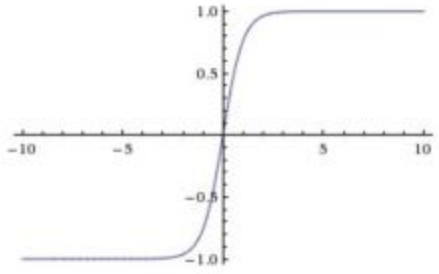


Output of neuron = Y= f(w1. X1 + w2.X2 + b)

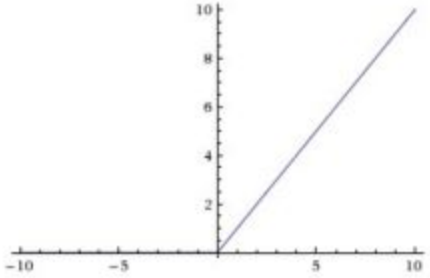
A note on neural networks: The activation function f



Sigmoid

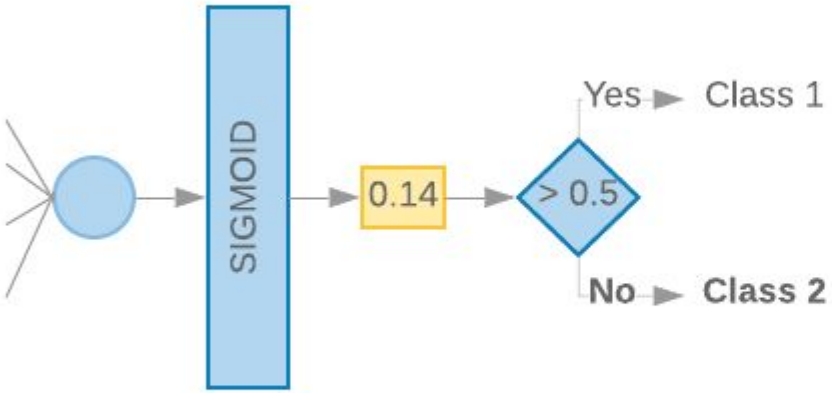


tanh

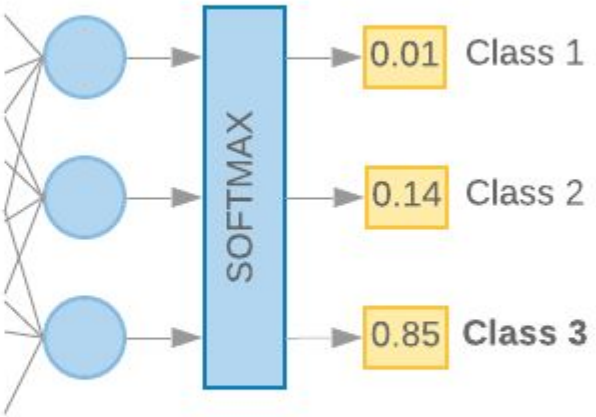


ReLU

A note on neural networks: Activation functions

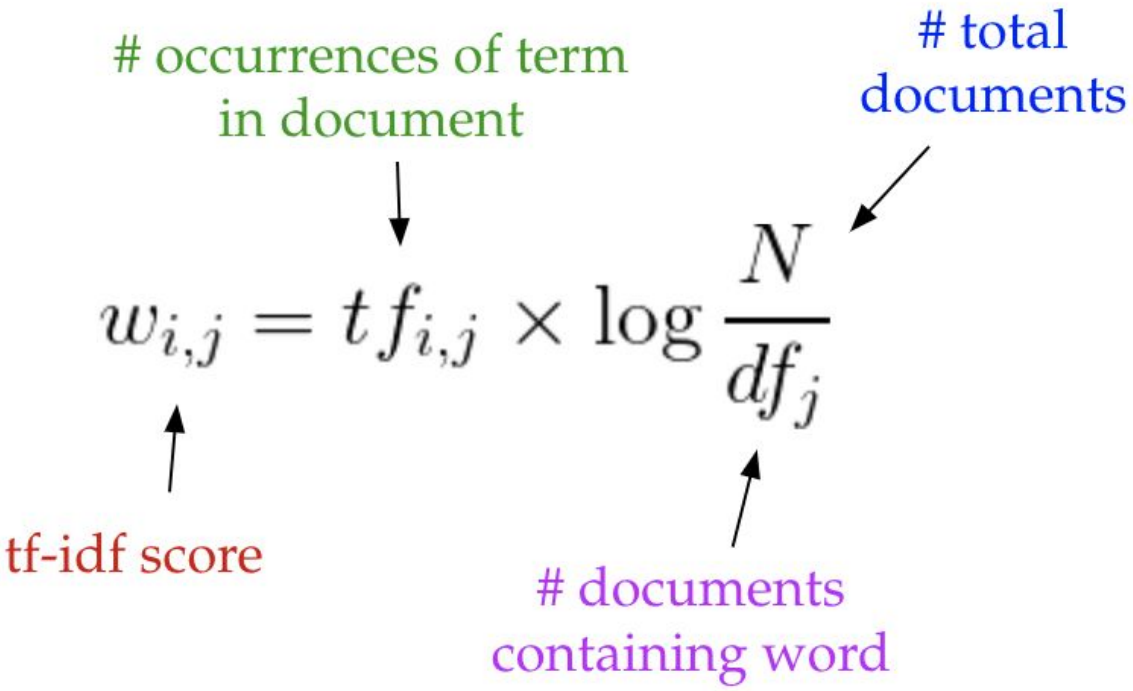


Output layer
Binary classification



Output layer
Multi-class classification

Term Frequency over Inverse Document Frequency: TF-IDF



Wrapping Up

Quirks of Real World AI Development

04

Quirks of Real World AI Development

- From unquantifiable goals to measurable metrics
- Getting a good labelled dataset
- Serving an ML Model
- Developer productivity

From unquantifiable goals to measurable metrics

- Finding the right metric
- Measuring success on real data
- Metrics may be influenced by other changes

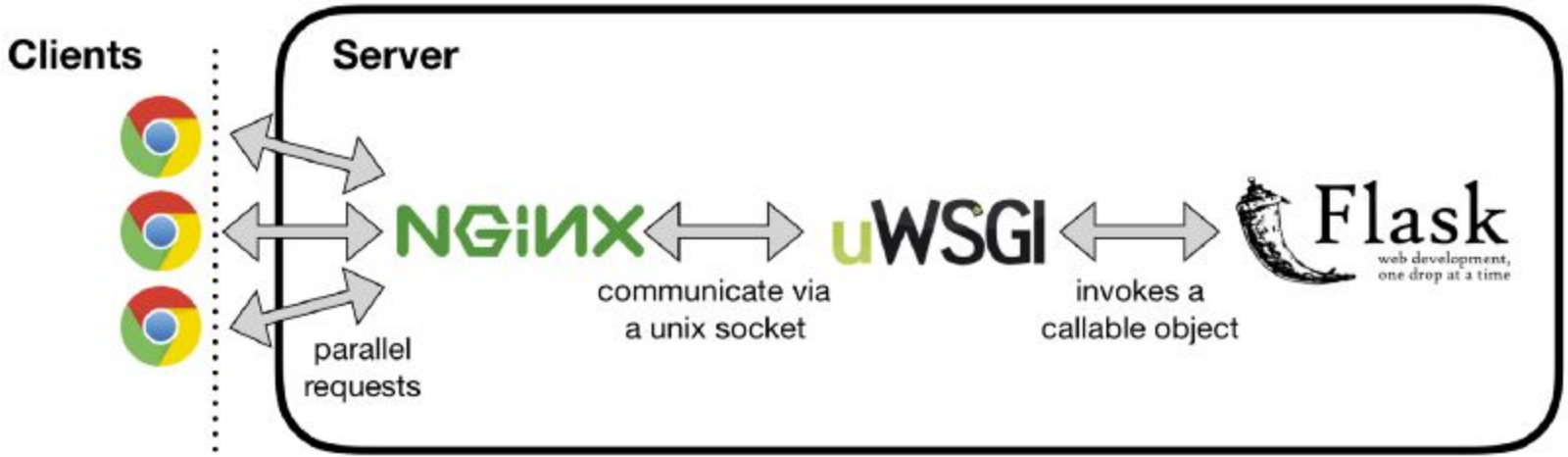
Getting a good labelled dataset

- Direct or Derived labels
- Finding a good sample to label
- Wrong labels in large datasets
- Bias in data

Serving an ML Model

- Kind of model and features to use
  - Offline
  - Online
  - Nearline
  - On device

Serving an ML Model: A simple scenario



Developer Productivity

- Versioning of models and datasets
- Searchable and reproducible experiments
- Monitoring performance, A/B testing, Debugging

Any Questions?

Slides, Code and Links

can be found at github.com/arjun-rao/talks/

Contact Us:

LinkedIn/Twitter: arjunra0

Linkedin:anishamascarenhas

Twitter: anisham197

References & Useful links

- https://ai.google/education
- https://developers.google.com/machine-learning/
- https://research.fb.com/the-facebook-field-guide-to-machine-learning-video-series/
- https://hackernoon.com/a-guide-to-scaling-machine-learning-models-in-production-aa8831163846