errorfaktor

spatial machine learning
information error mitigation
(ML-IEM)

faiz ikramulla

mano ramireddy

Maryville University Data Science, September 2020

Influence – neither good nor bad

- Fundamental concept in organization behavior

"the power or capacity of causing an effect in <u>indirect</u> or <u>intangible</u> ways" (mw)

-Is social media influence spatially random?

-Is social media influence spatially unbiased?

-Is social media influence spatially proprietary? – NO!

- Social media data is MASSIVE, yet (some?) is free and open source !
- We can model and locate the "error", and optimize or "fix" it

Machine Learning

- FAST(ER) computation of MASSIVE information to achieve explainable human & machine usable solutions

    -"all models are wrong, some models are useful"

    -error modeling - global and local optimizations possible

    -quantization, discretization, transformation

        -3D (lat,lon,z) – possible for spatial

        -4D (x,y,z + t) (time-series) – good for temporal spatial

- Patterns of "error" in space and time in dimensions not perceivable by humans alone

Machine Learning

- Indirect to direct
- Intangible to tangible
- Invisible to visible…
- … Influence to neutrality

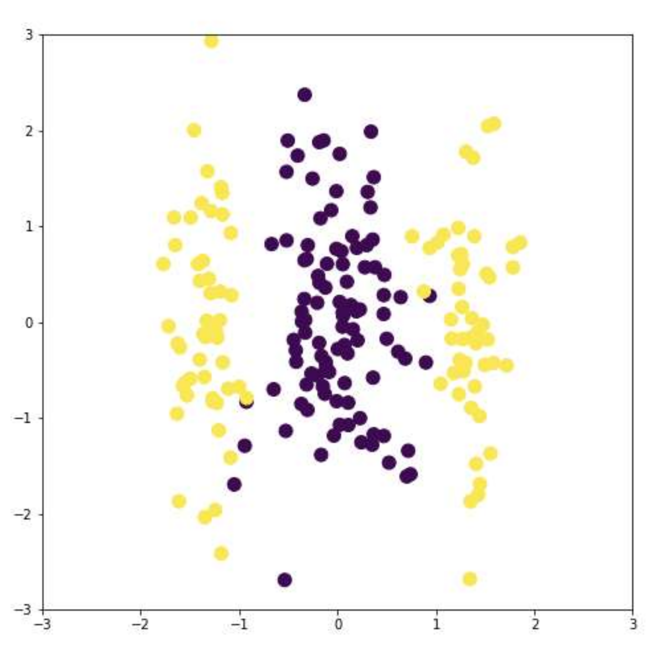
-SORT , FILTER , SELECT , JOIN , REMOVE

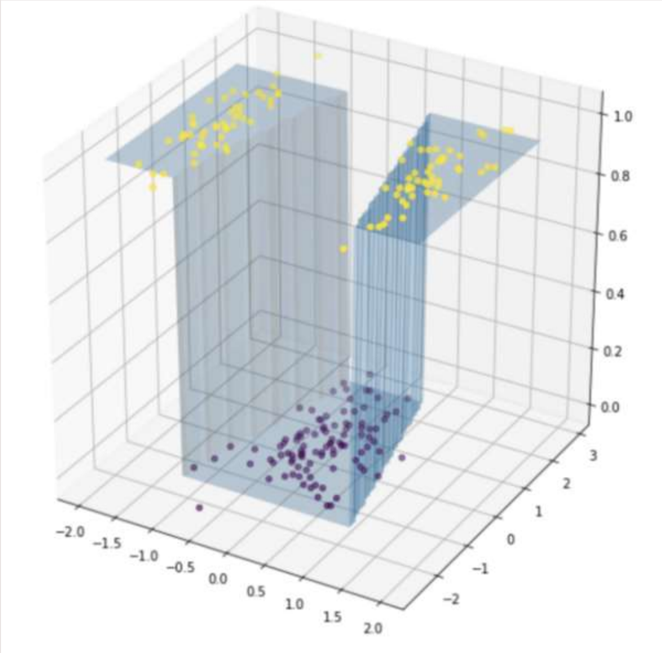
-NLP (Text, Audio, Locational, Other Behaviors)

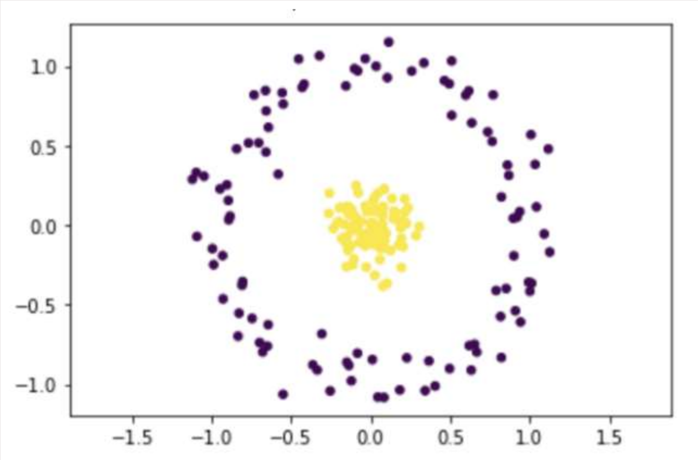
-Clustering / Vector Machines

Machine Learning – Ex.

https://towardsdatascience.com/animations-of-neural-networks-transforming-data-42005e8fffd9







Machine Learning – Feature Engineering

- Word source (original, re-post, human generated, machine generated)
- Word stats (static) – min, max, mean, standard deviation, histos
- Word stats (dynamic) – time series min, max, mean, std.
- Word association per time or per location or per time&location
- Same as above for emoji / symbols / digital images / digital video / etc.

- ESTIMATION: only about 10-20% of the social media has an original source. The rest is reposts or "shares".

Machine Learning – Label Determination

- Word association to other words or other events
- Unsupervised Learning – we are agnostic on influence – based on our features, we just want the machine to differentiate in ways we can not (easily) perceive
- We can then look for correlations between the differentiators (or their transformations) and apply optimization methods to reduce the error

Solution Proposal (ML-IEM)

Like an FMEA

- Mode – normal, abnormal, incorrect
- Severity – catastrophic, critical, marginal, negligible
- Likelihood – frequent, probable, occasional, remote, improbable

FUNCTIONS		POTENTIAL FAILURE MODES				POTENTIAL CAUSES			
Item / Function	Sub-Item / Sub-Function	Potential Failure Mode	Potential Effects of Failure	SEVERITY	CLASS	Potential Causes / Mechanisms of Failure	OCCURRENCE	Current Design Controls Prevention	Current Desig Detect
				1			1		
				1			1		
				1			1		
				1			1		
				1			1		
				1			1		

Dataset - Twitter Hydroxychloroquine

https://digital.library.unt.edu/ark:/67531/metadc1706013/

Twitter Hydroxychloroquine (UNT)

- Mode – normal, abnormal, incorrect
- Severity – catastrophic, critical, marginal, negligible
- Likelihood – frequent, probable, occasional, remote, improbable

- FACT: only 13-14% of this tweet data consisted of original posts.

Machine Learning – Algo Development

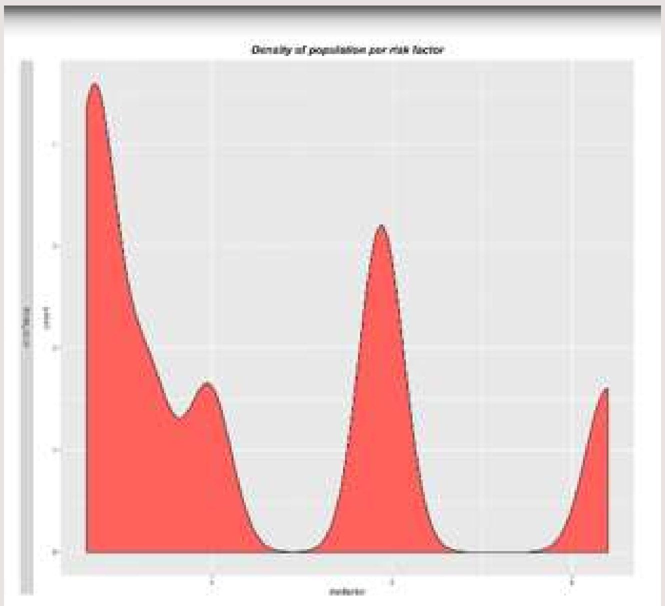
Main Flow:

- STEP 1: Model with all tweets
- STEP 2: Model with retweets/reposts FILTERed OUT / REMOVEd
- STEP 3: Compare - Are the differentiators now more or less or similarly evident?

REPEAT 1-3 (SORT , FILTER , REMOVE , etc.):

- Repetitive twitter handle sources, repetitive associate words
- Human vs. machine source
- Paid vs. unpaid source – need independent data set

Machine Learning + Geospatial



Example:

Probability Dist. Function of model “error” computation based on geocoded input data

Not normalized / not random

3 possible anomalies

Where AND when are they?

Events in that timeframe / location?

How to normalize / reduce?

Solution Proposal - reduce “error”

- Identify factors that when removed trend towards normalization.
- Spatially correlate these to specific locations – with statistical power
- Highlight these as geocoded “influencers” to users

  -words, handles, expressions, linguistics, behaviors + correlations

  -make non-personal
- Present space-time alternative view “FMEA dashboard” - per metro region

  -as-is

  -with “influencers” removed

  -with alternate “influencer” weighting (frequency* + TBD, pending model)

Workflow

!!! “An ounce of SQL is a worth a pound of python” !!!

- SQL - Query / aggregate / filter / sort / combine / join / transform geocoded data with built-in integrity **@Quality**
- R/Python – tidy, transform, process, analyze, model, visualize with automated math, stats, and logic **@Speed**
- General – structured data, ethical data **@Performance**
- Operation/Deployment – anywhere you like! **@Interoperable**

~70% of the time should be spent on preparing data to ensure reliable results.

Open Source Geospatial Tools

PostgreSQL (PostGIS)

R (rgdal, raster, sf, sp, leaflet)

Python (gdal, rasterio)

OSM

Open data everywhere!

Many other applications… election, census,

References

https://en.wikipedia.org/wiki/Failure_mode,_effects,_and_criticality_analysis

https://github.com/mona-kay/odsc-sql-for-data-science

https://www.merriam-webster.com/

https://towardsdatascience.com/animations-of-neural-networks-transforming-data-42005e8fffd9

https://digital.library.unt.edu/ark:/67531/metadc1706013/