Data and Machine Learning

Adam Gibson – Zipfian Academy

Kinds of Data

- Unstructured – text, images, audio, time series
- Structured – relational, event

Structured data

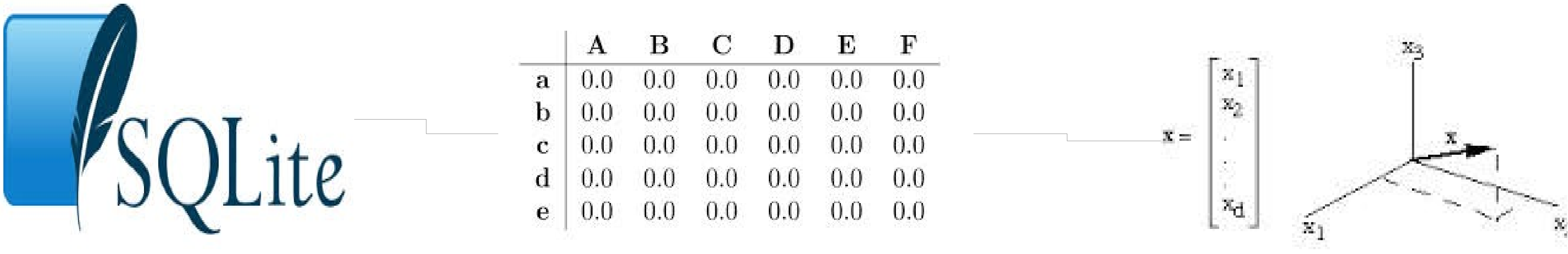
- Event Data (Clicks,referrals,actions)
- Entity (User Data)

Data and ML Algorithms

- Machines understand numbers – different kinds of data needs to be transformed to fit this format
- We call this process vectorization
- The “vector” part means a list of numbers
- The ultimate output is called a feature vector

The pipeline

- Structured data is stored in a database – in our case we will use SQLLite
- Structured databases use sql to get data out of them
- For our purposes, we will use pandas dataframes for quick analysis
- We will need to load pandas dataframes from sql
- From there, we vectorize the dataframes



	A	B	C	D	E	F
a	0.0	0.0	0.0	0.0	0.0	0.0
b	0.0	0.0	0.0	0.0	0.0	0.0
c	0.0	0.0	0.0	0.0	0.0	0.0
d	0.0	0.0	0.0	0.0	0.0	0.0
e	0.0	0.0	0.0	0.0	0.0	0.0

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

Extract,Transform,Load

- SQL Queries with SqlAlchemy
- Convert high level DSL to dataframe
- Group and transform data leveraging data frame
- DataFrame has operations for grouping, aggregation, statistics, and transformation of data

Vectorization

- Data frames handling transformations allows for very quick transformations of data
- One common example is transforming string data in to nominal data where a set of strings is mapped to a set of integers which can be used in feature vectors
- This allows plotting

Vectorization (cont)

- Time stamp data can be converted in to milliseconds
- Continuous data maybe discretized (rounded)

Pre processing the data (after vectorization)

- Normalization of data can allow for mapping the data all in to one vector space.
- Homogenization allows for easier learning and pattern recognition
- Different kinds of data warrant different transforms

Common Transforms

- Subtract mean and divide by standard deviation
- Row wise divide by the max element in the row
- Subtract from min divide by min – max
- Binarization
- Scaling (multiply by a decimal)
- Log transform (numbers could be really big)

Why all these different transforms?

- Depending on the kind of data and classifier, each one will be optimal for certain circumstances
- Images are good when scaled in the 0,1 domain with logistic regression
- Certain kinds of algorithms only accept binary inputs

An example pipeline

- Loading data from a database, create an sql query
- Using pandas load the data in to a dataframe
- Normalize each column relative to the kind of data via the pandas dataframe.map()
- Pass the data to a machine learning algo via dataframe.tomatrix()
- You may need to add train/test splits and the like for various algorithms