Methodological Issues in Data Science

Wayne Wobcke

School of Computer Science and Engineering
University of New South Wales

e-mail:w.wobcke@unsw.edu.au

© W. Wobcke 2019

Overview

- Data Science
- Bias
- Overfitting
- Human Expertise
- Slicing and Dicing
- Combining Datasets
- Validation

©W. Wobcke 2019

Data Science

Example: Mobile Phone Data includes location of cell towers

- Location is Angkor Wat and time is 1 day $\Rightarrow$ tourist?
- Or, journey "similar to" typical tourist trips $\Rightarrow$ tourist
- Location is shopping centre $\Rightarrow$ shopping (if not home)?
- Most frequent called person $\Rightarrow$ spouse? (if married)
- Spouse $\Rightarrow$ opposite gender (use as a check)
- Location is port and truck driver $\Rightarrow$ shipment
- Destination(s) of truck $\Rightarrow$ type of shipment?

Methodology: Emphasis on dealing with multiple levels of uncertainty

©W. Wobcke 2019

Bias

Bias = Propensity for method to generalize (good or bad)

- Dataset not representative of population
  - Only people in areas with phone towers have phones
  - Only poorer people need “access” to phone credits
- Training data “discriminates” against certain groups
  - Learner trained on white male faces
- Learner generalizes “wrong” features
  - White background (only pictures of snow leopards are in winter)
- Learner “misses” relevant features
  - Seasonal effects of population movement (food shortages)

©W. Wobcke 2019

Overfitting

Overfitting = Fit given data too closely and not work in other contexts

Example: How not to measure wealth index (Blumenstock et al. 2015)

- Mobile phone data with 5088 features and 856 labelled examples
- Choose features based on whole dataset (not training set)
- Don't consider what is Rwanda-specific about this data
- Use non-standard methdology drawn from another paper
- Ignore sensible (human-generated) baselines
- 5-fold cross-validation produces 5 models, not one

Claim: Most neural network/deep learning models overfit

©W. Wobcke 2019

Human Expertise

Essential when data is limited in quality, quantity (most of the time)

- Human suggests relevant features
  - Protest less likely to be violent if venue private
  - AfPak ontology of events of interest to conflict progression
- Human defines useful indicators
  - Village is safe if market is open at night
- Human validates model output
  - Check agreement with model on 15% random sample
  - Verify main features used by the model
  - Define baseline for comparative performance
  - Cross check model output with other datasets

©W. Wobcke 2019

Slicing and Dicing

- Data may only be reliable in certain contexts
  - May be able to determine event occurrence, not details
  - Sentiment analysis notoriously inaccurate
- May want to analyse subgroups by region, status, etc.
  - “Big data” can soon become “small data”
  - Need statistical methods to assess reliability
  - Map quality of data to quality of resulting decision

©W. Wobcke 2019

Combining Datasets

Use of only one type of data is insufficient for many purposes

- Especially social media data (Twitter, Facebook)
- Especially with complex metrics and indicators
  - Population health using images of hospital carpark
  - Rainfall locations and amounts using satellite data
- Need triangulation/corroboration, not increased uncertainty
  - Need to "correlate" independent data sources

©W. Wobcke 2019

Validation

Is data fit for (what) purpose?

- No model is ever perfect (especially learned models)
- Statistical correlations are usually very weak
- Contextualize models to local circumstances
- Cross check model outputs with other datasets
- Express uncertainty associated with conclusions/decisions
- “Big data” methods can provide “early warning” signals
- Complement national statistics with different time scales
- Continually validate models as assumptions vary

©W. Wobcke 2019

National Statistics

- Can define metrics, but still have to map to indicators
- What are the statistics used for (tolerance of error)?
- Statistical methods to assess reliability survey $\leftrightarrow$ data
- Could be possible to also use social media location data
  - Social media users have GPS location
  - Social media users form a social network
  - Social media users subset of all users

©W. Wobcke 2019