Data Science as 9 problems

DSTA

A gentle-yet-focussed introduction

Figure 1: Ch. 2

---

1. (was 2) Regression/value estimation

Instance:

- a collection (dataset) of numerical $< \mathbf{x}, y >$ datapoints
- a regressor (independent) value $\mathbf{x}$

. . .

Solution: a regressand (dependent) value y

that complements $\mathbf{x}$

Measure: error over the collection

---

2. (was 1) Classification and class probability

Instance:

- a collection (dataset) of datapoints from $\mathbf{X}$
- a classification system $C = \{c_1, c_2, \ldots c_k\}$

. . .

Solution: classification function $\gamma : \mathbf{X} \to C$

Measure: misclassification

. . .

[PF] "classification predicts whether something will happen, whereas regr. predicts how much something will happen."

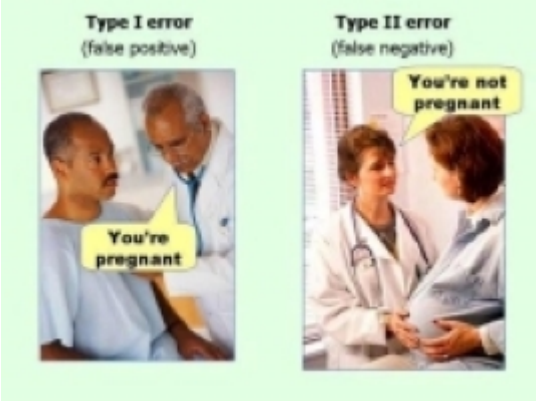


Figure 2: Type I and II errors

---

3. Similarity

Identify similar individuals based on data known about them.

Instance:

- a collection (dataset) of datapoints from $\mathbf{X}$, e.g., $\mathbb{R}^n$
- (distance functions for some of the dimensions)

. . .

Solution: similarity function $\sigma : \mathbf{X} \to \mathbb{R}$

[**Measure:** error]

---

4. Clustering (segmentation)

group individuals in a population together by their similarity (but not driven by any specific purpose)

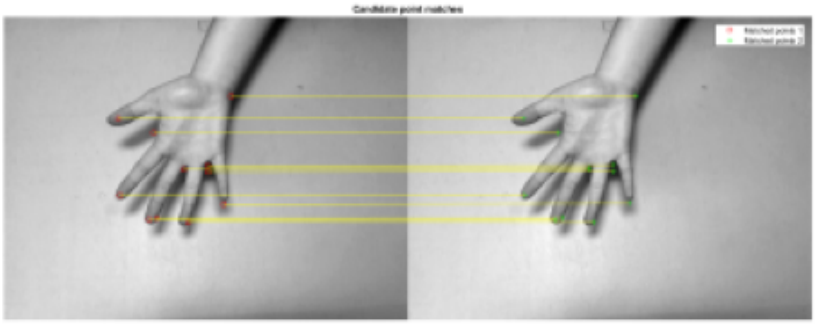


Figure 3: Ch. 2

Instance:

- a collection (dataset) $\mathbf{D}$ of datapoints from $\mathbf{X}$, e.g., $\mathbb{R}^n$
- a relational structure on $\mathbf{X}$ (a graph)
- a small integer k

. . .

Solution: a partition of $\mathbf{D}$ into $\mathcal{C}_\infty, \ldots \mathcal{C}_\|$

Measure: network modularity Q: proportion of the relational structure that *respects* the clusters.

Detection version: k is part of the output.

See an example research work (from yours truly)

---

5. Co-occurence (frequent itemset mining)

similarity of objects based on their appearing together in transactions.

Instance:

- a collection (dataset) $\mathbf{T}$ of itemsets (subsets of $\mathbf{X}$) or sequences
- a theshold τ

. . .

Solution: All *frequent patterns:* subsets that appear in $\mathbf{T}$ above τ

. . .

Detection version: τ is part of the output.

Market-basket analysis, (some) recommendation systems

---

6. Profiling (behaviour description)

Instance:

- a user description **u** drawn from a **D** collection
- a stimulus $a \in \mathbf{A}$
- a set of possible responses **R**

. . .

Solution: a functional reaction of **u** to **a**, i.e., $\rho : \mathbf{U} \times \mathbf{A} \rightarrow \mathbf{R}$

. . .

Application: anomaly/fraud detection.

Example research work on Social media profiling

---

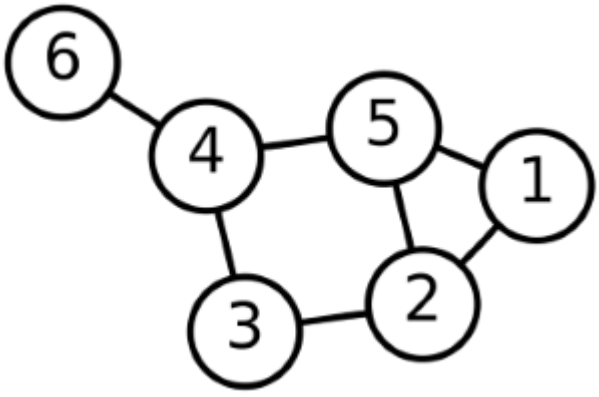
7. Link prediction

Instance: a dynamical graph (network) **G** , i.e., a sequence

$< V, E >$,

$< V, E' = E + \{(u, v)\} >$,

$< V, E'' = E' + \{(r, s)\} > \ldots$



Question: what is the next link to be created?

What YouTube video will you watch next?

Alternatives: predict the **strength** of the new link; link deletion.

8. Data reduction

Instance:

- a collection (dataset) **D** of datapoints from **X**, e.g., $\mathbb{R}^m$
- [a distinct independent variable x_i]

. . .

Solution: a projection of **D** onto $\mathbb{R}^n$, $n < m$

Measure: error in the estimation of x_i

Example: genre identification in consumer behaviour analysis

---

9. Causal modelling

Instance:

- a collection (dataset) **D** of datapoints from **X**, e.g., $\mathbb{R}^m$
- a distinct dependent variable x_i

. . .

Solution: a variable x_j of **D** that controls x_i

Measure: effectiveness of x_j *tuning* to *tune* x_i in turn.

. . .

Example: Exactly What food causes you to put on weight?

Controlled clinical trials, A/B testing.

[Un]Supervision

Supervised Data Science

- obtain a dataset of examples, inc. the "target" dimension, called *label*
- split it in training and test data
- run a. on the test data, find a putative solution
- test the quality/pred. power against test data

. . .

Regression involves a numeric target while classification involves a categorical/binary one

---

Supervised

1: Regression

2: Classification

9: Causal Modelling

Could be either

3: Similarity matching,

7: link prediction,

8: data reduction

(mostly) unsupervised

4: Clustering

5: co-occurrence grouping

6: profiling