Exam — Answers — Business Intelligence and Process Modelling
Universiteit Leiden — Informatica & Economie, Friday June 10, 2016, 14:00–17:00

1. The term *codeless reporting* refers to making data accessible such that one does not have to do any programming (or SQL-querying) to view and understand this data. An example is inspecting data visualized through a dashboard.

2. If $x < y$, then there are redundant or unused visual attributes, which may result in an unnecessary cluttered visualization. If $x = y$, then all data properties can be mapped to visual attributes, which assuming that the actual value of x is not too high, results in a representative visualization. If $x > y$, then some data properties are not visualized, which results in a poor visualization, assuming the non-visualized attributes are important for understanding the data.

3. A "full" man implies 100%, but the percentages sum to more than 100%. A two-dimensional visualization is used for a one-dimensional value, which results in visually misleading proportions.

4. Transactional system vs. data warehouse: volatile data vs. nonvolatile data, oriented on daily operations vs. oriented on data analytics, data storage focus vs. information acquisition focus (and many more).

5. If the input data is of poor quality, then likely the output of an analysis of this data is also of poor quality. This is why data quality assurance is so important.

6. Overfitting means that the model also describes noise and random errors, and not only the real patterns in the data. Underfitting means that the model is too generic and does not capture all the patterns present in the data. A decision tree with a a very low (close to 0) depth d is more likely to be underfitted. Similarly a decision tree with d very close to or (although prohibited in most learners) larger than n is more likely to be overfitted.

7. In *supervised* outlier detection, the task is to distinguish between items that have been labeled as 'normal' or 'abnormal' using some classification technique. *Unsupervised* outlier detection deals with finding instances that are least similar to all the other instances, for example instances that fall into their own cluster after some clustering algorithm is applied to the data. In *semi-supervised* outlier detection, first normal instances are characterized using a training data set, and then unlabeled examples are judged according to whether or not they fit the trained data.

8. Advantages: 1) lower variable costs, 2) lower investment costs, 3) immediate access to specialized experts and 4) possible benefit of their experience in analyzing similar data. Disadvantages: 1) possibly valuable data may leave the company which may raise privacy concerns, 2) albeit anonymized, the third party may use the data indirectly for other (competing) business with similar data, 3) the knowledge of the analysis of the data will not become "in-house" and 4) it results in a dependency on the provider.

9. Business intelligence is about automatically analyzing and mining business data in order to make better business decisions. Process modelling is about (often manually) formalizing processes within an organization. Business process intelligence is then about analyzing and mining process data in order to (automatically) derive and improve the business processes.

10. It tells us that all derived features are somewhat unrelated to each other, as there are no strong correlations between them. It does not say anything about the usefulness, as it might be that the the grade actually correlates with none of the attributes, with a combination of some of the attributes, or with all attributes. However, given the fact that the features all seem somewhat independent, but were created with the domain knowledge in mind, they may work together quite well in a classifier.

11. Using the 20 features and the grade as a 21st feature, an supervised learning algorithm can be devised. Overfitting can be prevented by using separate training set and test set and/or k-fold cross-validation.

12. A perceptron is not just more simple than a neural network, it is also less powerful in a sense that it can only linearly separate instances, whereas more complicated patterns may need to be detected. The Minimal Description Length (MDL) principle instead says that the best description for a dataset is the one that results in the largest compression of the data.

13. 1) The density, measuring how sparse or dense the network is by dividing the number of edges by the maximum number of edges. 2) The degree distribution (for the indegree and outdegree), the distribution of how often each degree value occurs.

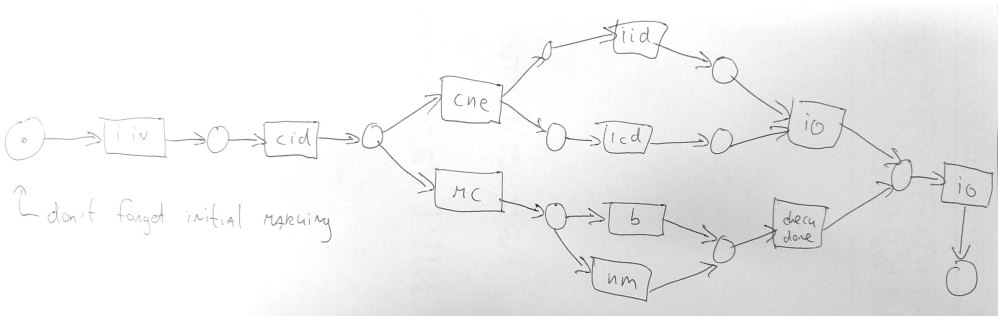
14. 1) Indegree centrality, indicating the number of nodes pointing to the considered node, meaning that we are counting the number of shareholders of that firm. 2) Outdegree centrality, indicating the number of nodes in which the considered node holds shares, meaning that we measure the number of firms on which the considered firm is dependent.

15. The emergence of new nodes indicates that new companies being founded, whereas the appearance of new links may indicate firm acquisitions.

16. In its descriptive role, it indicates what the current business processes look like. It can be prescriptive in the sense that it indicates what the processes should look like (and how the actual process differs and should perhaps be changed). It can be explanatory in a sense that using the model it can explain why certain business activities are for example efficient, slow or crucial.

17. `https://www.informatik.uni-hamburg.de/TGI/PetriNets/introductions/aalst/elevator2.swf` (credits to Wil van der Aalst).

18. In the figure below, the transitions have shortened labels corresponding to the ones in the BPMN figure in the exam.



19. 1) Fitness (the discovered model should allow for the behavior seen in the event log), 2) precision (the discovered model should not allow for behavior completely unrelated to what was seen in the event log, 3) generalization (the discovered model should generalize the example behavior seen in the event log) and 4) simplicity (the discovered model should be as simple as possible).