School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2019)
Sample Solutions: Week 7

1. What is "link analysis"? Briefly describe the "PageRank" algorithm and the rationale behind it.

   The terms in the document aren't the only factor for determining which documents are relevant. Link analysis is a weighting (or re-ranking) procedure which alters the importance of documents (or terms within documents), based on the link structure of the Web.

   For "popularity"-based approaches, we're trying to build a model of how popular pages are based on which pages are linking to a given document (and, in turn, the popularity of those pages). We then assume that popular pages are more likely to be relevant than unpopular pages. In other words, if a document has many (good) incoming links, we assume that it is more likely to be relevant than a document with few (good) incoming links. "PageRank" is an example of one such algorithm.

2. In the PageRank algorithm:

   (a) What is the mechanism for the "random walk"?

   - We begin at a random page (or probabilistically, we begin at all pages with equal probability).
   - We choose one of the following:
     - We follow one of the outgoing links from this page, with probability (1-λ) – evenly distributed amongst all outgoing links on this page
     - We "teleport" to a page entirely at random, with probability λ – evenly distributed amongst all pages (Note: if there are no out-going links, we do this with probability 1, evenly distributed.)

   (b) In terms of user behaviour, what is the significance of "teleporting"?

   - Following a link from a page is kind of obvious, based on user behaviour. "Teleporting", on the other hand, corresponds to navigating via entering a URL into the address bar (which doesn't appear to be related the content of this page).

---

4. What is data mining/machine learning? What makes this a knowledge task?

   - Data mining: extracting implicit, previously unknown, potentially useful information from data
   - Machine learning: algorithms for acquiring structural descriptions from examples and use it to something about the new/unseen instances
   - Knowledge task: the information/descriptions we produce are unknown and useful to humans

5. Revise the definitions of instances and attributes (or features). For the following problems, identify what the instances and attributes might consist of:

   - An instance is a single exemplar from the data, consisting of a bundle of (possibly unknown) attribute values (feature values) [and in the case of supervised ML a class value].
   - An attribute is a single measurement of some aspect of an instance, for example, the frequency of some event related to this instance, or the label of some meaningful category.
   - Attributes are usually classified as either nominal (labels with no ordering), ordinal (labels with an ordering), or continuous (numbers, even if they perhaps aren't continuous in the mathematical sense).

(i) Building a system that guesses what the weather (temperature, precipitation, etc.) will be like tomorrow

- It seems fairly clear that each instance will be a day; depending on how we construe the problem, various properties could be attributes — the most logical is probably the corresponding data (temperature, precipitation, humidity, wind speed, etc.) from the previous day(s).

(ii) Predicting products that a customer would be interested in buying, based on other purchases that customer has previously made

- It's a little unclear whether each customer is an instance, or each customer-product pairing is an instance (which will in turn lead to different strategies to do the learning). In either way the attributes can be the customer's name, age, address, gender, shopping log, credit card information and more.

(iii) Automatically identifying the author of a given piece of literature

- Again it can be simply inferred that pieces of literature are the logical instances here and the attributes can be the title, year of publish, language, genre and more.

6. What are the main differences between supervised and unsupervised machine learning?

- Generally speaking, supervised techniques in machine learning start from exemplars — labelled with classes — in a set of training data, and use these to classify unknown instances in a set of test data.
- Unsupervised methods are not based on a set of labelled training data: they are often broken down into *weakly unsupervised* methods, where the class set is known, but the system does not have access to labelled training data; and *strongly unsupervised* methods, where even the class set is unknown.
- For example, Naive Bayes, Support Vector Machines, Decision Trees, and k-Nearest Neighbour are all examples of supervised systems.
- Clustering (e.g. k-means, Expectation Maximisation) is an example of an unsupervised methodology.

7. Based on the problems in question 5, identify the "concept" we might attempt to "learn" for each problem, and conjecture whether a typical strategy is likely to use supervised or unsupervised Machine Learning.

(i) The question suggests that there are multiple concepts here — corresponding to the various weather features of the particular day that we are trying to predict; assuming that we can access historical data for the particular location, (supervised) regression seems like the most plausible ML strategy.

(ii) There are a couple of different ways of construing the problem:

- If we attempt to exhaustively label every product for every customer as either "interested" or "not interested", then we have a classification problem, where we might try to predict these labels based on some properties of the product and customer;
- If we instead construe a customer as an instance, we might then try to find a single product (or set of products) that the customer would be interested in. Whether this would be a supervised problem (probably classification) or unsupervised problem (probably association rule mining, or perhaps clustering) would depend on how likely we are to be able to access labelled data for a sufficient number of customers, so that we could build a sensible model

(iii) Again there is some question about the problem domain, for example:

- If we have a single unknown piece of literature and a fixed set of authors who may have written it — and a collection of their previous writing — then this is probably a classification problem, where we might associate each piece of writing with the words (or grammatical structure, and perhaps metadata) contained within it;

- If we have an open–domain problem — that potentially anybody could have written it — then collecting labelled data would be possible (i.e. classification), albeit obnoxious. We might instead prefer to use a clustering approach based on the document's linguistic properties (although this is unlikely to identify a single author);
- We might instead have a situation like plagiarism detection, where we don't have access to very much data for any individual author. In that case, simple classification is unlikely to be very effective (because our model might be insufficient to represent each author), but we could try something like outlier detection