INFO411 Data Mining and Knowledge Discovery

Project 4

Instructions:

This task is a real-world data mining problem. You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) your data mining approach and the methodologies; (4) the strengths and weaknesses of your proposed approach; (5) your results, a comparison, and an analysis of the results; (6) a brief discussion and a conclusion. Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Document Classification (Web Spam Detection)

Background:

The term *web spam* refers to the results of activities with an intention to mislead search engines into believing that a particular web page has a high authority value on a particular query, while in fact that particular web page may contain little or no relevant information. Search engines sort the URLs returned in response to a user query on the basis of a score that is usually composed of two parts: a measure of the relevance of the page content with respect to the query (see for example Manning, C., Raghavan, P., & Schütze, H. (2008), "An introduction to information retrieval", Cambridge University Press) and a measure of the popularity of the page (see e.g. Brin, S., & Page, L. (1998). The anatomy of a large–scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 107–117.).

In general spam techniques can be classified into content-based and link-based according to whether their focus is on the former or on the latter measure, respectively (see i.e. Gyöngyi, Z., & Garcia-Molina, H. (2005). "Web spam taxonomy". Adversarial Information Retrieval on the Web).

In a link-based spam, a web page is linked by a large number of links from other web sites. This will increase the popularity of the spam web page as its popularity is due largely to the number of web pages which are linked to it. Thus it is possible to create many web sites, and have each of these web sites linked to a particular spam web page. This is commonly called a "link farm" where links to other web pages can be automatically generated. On the other hand, in content-based spam, a web page is automatically provided with terms that are visually hidden from users and irrelevant to the actual content of the web page. But, these terms are indexable by search engines, so that whatever query that a user issues to a search engine, there is a high possibility that the spam web page will be returned.

We will make use of the benchmark dataset known as the UK2007 benchmark problem. This is a large collection of annotated spam/nonspam hosts labeled by a group of volunteers. The base data is a set of 105,896,555 pages in 114,529 hosts in the .UK domain. The collection was tagged at the host level by a group of volunteers.

The Spam/nonspam labels are available from: https://chato.cl/webspam/datasets/uk2007/

The dataset comes with different feature sets such as direct features, link-based features, and content based features. They can be downloaded at https://chato.cl/webspam/datasets/uk2007/features/

Definition of the task:
The UK2007 spam detection is a classification learning problem. You are to identify the value of each of the three types of features (which one of these feature sets helps to create a model with the best predictive power).
We introduced a number of classification methods in the lectures.
1. Deploy the most suitable of these classification methods to each of the features sets and fully justify your choice of method.
2. Rank feature sets by the quality of results (first list the feature set that produced the best result, the feature set that produced the poorest result is listed last).
3. Main objective of this project: Fully analyse and compare the results. Use AUC (Area under the ROC curve) as a basis for the comparisons. Fully explain your findings.

Requirements:
1. Present a general description of the dataset and present the general properties of the dataset.
2. Deploy the classification methods to each of the feature sets and present the results. Analyse and compare the results then discuss the strengths and weaknesses of the classification method used (in the context of this learning problem).
3. Deploy the classification method to combinations of the feature sets. Present the results and offer a qualitative comparison.
4. Summarize: What new and interesting things did you discover while working on this project?