Fundamentals of Data Science

Semester B 20-21

Tutorial 1

1. Discuss whether or not each of the following activities is a data mining task.

   a. Computing the total sales of a company.
   b. Sorting a student database based on student identification numbers.
   c. Predicting the future stock price of a company using historical records.

2. We consider a collection of news articles shown in the following table. Each article is represented as a set of word-frequency pairs (*w*,*c*), where *w* is a word and *c* is the number of times the word appears in the article.

Article	Words
1	dollar:1, industry:4, country:2, loan:3, deal:2, government:2
2	machinery:2, labor: 3, market:4, industry:2, work:3, country:1
3	job:5, inflation:3, rise:2, jobless:2, market:3, country:2, index:3
4	domestic:3, forecast:2, gain:1, market:2, sale:3, price:2
5	patient:4, symptom:2, drug:3, health:2, clinic:2, doctor:2
6	pharmaceutical:2, company:3, drug:2, vaccine:1, flu:3
7	death:2, cancer:4, drug:3, public:4, health:3, director:2
8	medical:2, cost:3, increase:2, patient:2, health:3, care:1

   Suppose we apply cluster analysis to group the set of articles based on their respective topics. Suggest a suitable representation format for each article such that the degree of similarity between two articles can be readily compared.

3. We consider the problem of predicting whether a loan applicant will default on his/her loan. A data set for this problem is shown in the table in the next page. Each record contains the personal information of a borrower, along with a class label indicating whether the borrower has defaulted on loan payments.

Applicant No.	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	Yes
4	Yes	Married	120K	No
5	No	Single	95K	No
6	No	Married	60K	No
7	Yes	Single	220K	No
8	No	Single	80K	Yes
9	No	Married	75K	No
10	No	Single	90K	No

a. Suppose we first focus on the attribute Home Owner. Describe in what way this attribute can help us to determine whether a borrower will default on his/her loan payment or not.

b. Suppose we next focus on the subset of non-home owners. Describe in what way the attribute Marital Status can help us to determine whether a borrower will default or not.

c. Finally, determine how we can make use of the attribute Annual Income to complete the prediction process.