Workshop Week 9 - COMP20008 2018

Background

Read Sections 1,2, 4.1,4.2 from *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection, Peter Christen, Springer, 2012.* Available as an e-book for download by University Library

Technical questions

1. Suppose you are conducting data linkage between two databases, one with m records and the other with n records (assume $m < n$). Under a basic approach, $m \times n$ record comparisons will be needed.

   - What is the maximum number of record matches? What is the corresponding number of non-matches in this circumstance?

   Now suppose a blocking methodology using b blocks is employed.

   - How many comparisons will be needed in the best case?
   - How many comparisons will be needed in the worst case?
   - How many comparisons will be needed on average? (state any assumptions)
   - What is the advantage of using large b? What is the advantage of using small b?

2. (From 2017 exam) Max is having a conversation about data integration. He says "Using blocking for record linkage between two datasets (dataset A and dataset B) is a bad idea. It is too time consuming to assign the records to blocks. It is much better instead to directly compare the records in A against the records in B without using any blocking step". Argue why Max's statement is incorrect.

3. Explain how the data duplicate detection problem relates to the data linkage problem. What are the major differences?

4. One may evaluate the output of a data linkage system according to how many records are linked correctly and how many records are linked incorrectly.

   - What are the reasons two records could be linked incorrectly?
   - Suppose a false positive (FP) is two records that are linked by the system, which a human believes should not have been linked. Suppose a false negative (FN) is when two records are not linked by the system, which a human believes should have been linked. A true positive (TP) is two records linked by the system which

a human believes should have been linked and a true negative (TN) is two records not linked by the system which a human believes should not have been linked.

- What are the relative sizes of the categories TP, TN, FP, FN? In practice, how might one calculate these sizes?
- It is desirable to minimise both FP and FN, but it may be difficult to minimise both simultaneously. Give an example application where minimising FP is more important than minimising FN. Give an example application where minimising FN is more important than minimising FP.

Unstructured

The remainder of the workshop is available for discussions about issues related to:

- Phase 3 of project (expectations, hints, difficulties, ...)
- General questions about lecture materials, previous workshops (code/exercises, ...)

Previous Exam Questions

Exam 2018 - Question 2

Given two instances represented by the tuples $(3, 2, 2, 3, 1, ?, 7)$ and $(?, 4, 6, 5, 9, ?, ?)$

a) (1 mark) Write an expression for the Euclidean distance between these two instances using mean imputation (Method 1 in lectures).

b) (1 mark) Write an expression for the Euclidean distance between these two instances using scaling (Method 2 in lectures).

In both 2a) and 2b), you may leave any square root terms unsimplified.

Exam 2018 - Question 8

Given a dataset with three classes, A, B and C, suppose the root node of a decision tree has 50 instances of class A, 50 instances of class B and 100 instances of class C. We are evaluating a candidate split of this root node into three children using a categorical feature F that has three possible values. The first child has distribution (25 class A, 25 class B, 25 class C), the second has (20 class A, 5 class B and 0 class C), the third has (5 class A, 20 class B, 75 class C).

a) Based on these numbers, write an expression for computing the utility of this candidate split using the information gain criterion. The expression may be complex and you do not need to simplify it to a single number.

b) Explain the relationship between the split utility and mutual information.

Exam 2018 - Question 10

Consider the following steps of the k-NN algorithm to classify a single test instance

```
1. Compute distance of the test instance to each of the instances
in the training set and store these distances
2. Sort the calculated distances
3. Store the K nearest points
```

```
4. Calculate the proportions of each class
5. Assign the class with the highest proportion
```

Step 1 may be very slow if the training set is large.
Suggest three possible strategies that might be used to speed up this step and describe a disadvantage of each.

Exam 2018 - Question 13

Suppose you are conducting data linkage between two databases, one with m records and the other with n records (assume $m < n$). You are using a blocking strategy.

a) Assuming that each record is assigned to exactly one block and that the blocking strategy uniformly distributes records across blocks, show that a total of mn/b record comparisons will be needed on average.

b) Explain why this formula (mn/b record comparisons) would not be applicable if a record could be assigned to multiple blocks.

c) Explain an advantage of assigning a record to multiple blocks, instead of to a single block. Explain a disadvantage.