Workshop Week 10 - COMP20008 2020SM2

1. Consider the following data set for a binary class problem:

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
F	F	-

We wish to select the feature that best predicts the class label using the χ^2 method.

- Write down the observed and expected contingency tables for feature A
- Calculate the $\chi^2(A, Class)$ value.
- Using the table below, conclude whether feature A is independent of the class label for $p = 0.05$.

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46

- Repeat the process for feature B and decide which feature could be best used for predicting the class label.

2. 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.

- Assume there are no duplicates. What is the maximum number of record matches? What is the corresponding number of non-matching comparisons required in this circumstance?

Now suppose a blocking method is employed, where each record is assigned to exactly one block. Assume this method results in b number of blocks.

- What is the smallest possible number of comparisons?
- What is the smallest possible number of comparisons that will return all matches? What is the value of b?
- What is the largest possible number of comparisons?
- If records are evenly allocated to b blocks, how many comparisons will be needed?
- What is the advantage with a large b? What is the advantage with a small b?
- In practice, a record is assigned to more than one block and records are not evenly allocated to blocks. How would this affect your analysis of large b vs small b?
- What is the advantage of records being assigned to multiple blocks?

3. 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.

4. Open Jupyter Notebook and implement a coded solution to Question 1. Ensure that your code gives the same answer as your calculation in Question 1.