CS-E4650 Methods of Data mining
Home assignments 2

Deadline Sun 11.10.2020 23:59

Version 2: Corrected the thresholds in task 3 (were given for $n \cdot MI$) and clarified equations of mutual information (2-based logarithm). Added missing negation to task 3d.

Each task has maximum 20 points.

1. (20 p) Consider the `spiral.txt` dataset. The first two columns are the data features, and the third contains the ground-truth labels. Remember to discard the label before running the clustering algorithms.

   It is recommended to plot the data set.

   Cluster the data into 3 clusters using the following methods:

   - k-means,
   - Spectral clustering, using a Gaussian kernel and a Laplacian matrix of your choice. You can try different values of the kernel parameter, to see if it has any effect. Note: if you are using a software package, try to figure out which Laplacian matrix it uses.

   Compute the quality of the resulting clusters using the following indices:

   - Silhouette,
   - Davies-Bouldin,
   - Normalized Mutual Information, where the first variable is the ground-truth labels and the second is the labels obtained by the clustering algorithm.

   Rank the algorithms in terms of their performance with respect to each of these indices and discuss the results. Which one captures the performance of the algorithm more accurately?

2. (20 p) Let us consider a different validation method. Let K be the kernel matrix, that is, $K_{ij} = \kappa(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)$, where x_i is

the i-th data point. Given a clustering of the data, define

$$c_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

For a given clustering, we define

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j \neq i} c_{ij} K_{ij}}{\sum_{j \neq i} K_{ij}},$$

where n is the number of data points.

(a) Compute τ for the previously obtained clusterings and compare the results with the ones indicated by Silhouette and Davies-Bouldin. Discuss the results. Is this index a better choice in this case?

(b) Do you see any clear disadvantage in using τ? Hint: consider the definition of the Gaussian kernel.

(c) Propose an alternative validation index. Discuss the motivation, and possible disadvantages. Try it and discuss the results. Hint: try to combine the idea of τ with other similarity matrices, such as the k-nearest-neighbour graph adjacency matrix, etc.

3. (20 p) Let us consider an imaginary database consisting of 1000 patients (50% female, 50% male), 30% of them with heart disease. The database contains information on patients and their life style like smoking status, drinking coffee or tea, having stress, going for sports, and using natural products. Table 1 lists some candidate rules related to heart disease. The required equations for mutual information are given in Appendix 1.

a) Calculate leverage or lift values for all rules. Prune out rules that do not express positive statistical dependence.

b) Evaluate mutual information MI of remaining rules and prune out rules where $n \cdot MI < 1.5$ (i.e., $MI < 0.0015$).

c) Evaluate overfitting among remaining rules using value-based interpretation and conditional mutual information MI_C: Rule $\mathbf{X} \rightarrow C{=}c$ is pruned out if there exists some $\mathbf{Y} \subsetneq \mathbf{X}$, such that for $\mathbf{X} \rightarrow C{=}c$ either $P(C{=}c|\mathbf{Y}) \geq P(C{=}c|\mathbf{X})$ or the improvement is not sufficient, $n \times MI_C < 0.5$ (i.e., $MI_C < 0.0005$).

d) How would you judge rule *tea*→ ¬*heart disease*, given $fr(tea)$=390, $fr(tea, \neg\ heart\ disease)$=283, and $fr(tea,\ smoking)$=40? Hint: calculate expected frequency of *tea,* ¬*heart disease* assuming conditional independence given *smoking* and ¬*smoking*. You can derive all required probabilities from these frequencies and the table.

Table 1: Candidate rules $\mathbf{X} \to C{=}c$ related to C = heart disease. $fr_X = fr(\mathbf{X})$, $fr_{XC} = fr(\mathbf{X}C)$.

num	rule	fr_X	fr_{XC}
1	smoking → heart disease	300	125
2	stress → heart disease	500	150
3	sports → ¬ heart disease	500	400
4	coffee → ¬ heart disease	342	240
5	natural product → ¬ heart disease	2	2
6	female → ¬ heart disease	500	352
7	female, stress → heart disease	260	100
8	chocolate, bananas → heart disease	120	32
9	smoking, coffee → heart disease	240	100
10	smoking, sports → heart disease	80	32
11	stress, smoking → heart disease	200	100
12	female, sports → ¬ heart disease	251	203

4. (20 p) Make experiments with Kingfisher program (see exercise session 2 instructions) using data worlddiskr.names. The data is a discretized and transaction-formed data from the earlier worldstat.csv data.

   a) Transform item names to integer codes, search positive and negative rules using $\ln(p_F)$ measure and transform number codes back to item names so that you can interpret the rules. How do you interpret (summarize) the main message of best rules? Can you find significant rules related to ex-colonies, corruption, length of compulsory education or oil?

   b) The data contains a lot of specious associations, including rules $\mathbf{Q} \to C{=}c$ that are specious generalizations of some $\mathbf{X} \to C{=}c$, $\mathbf{Q} \subsetneq \mathbf{X}$. Make a program that prunes out some of these by calculating conditional leverage $\delta_2 = \delta_C(\mathbf{Q} \to C{=}c \mid \neg\mathbf{X})$. If $\delta_2 \leq 0$, the rule is specious. (Note that $\delta_1 = 0$ always for this type of

specious rules.) How many rules can you prune among 100 best rules?

5. (20 p) Searching only maximal, closed or free sets are popular techniques for reducing the number of frequent patterns. (See definitions in the Appendix 2.)

   a) (10 p) Can you find all positive statistical associations, if you construct rules only from these sets (i.e., without searching anything else from data)? Consider separately maximal, closed and free sets and different min_{fr}s and justify your answer (give proofs or counter-examples)!

   b) (10 p) Can you detect overfitted (over-specialized or redundant) rules, if you are given only maximal, closed or free sets? Justify your answer!

Appendix 1: Required equations of mutual information

Mutual information of rule $\mathbf{X} \rightarrow C{=}c$ is

$$MI{=}\log \frac{P(\mathbf{X}C)^{P(\mathbf{X}C)}P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)}P(\neg\mathbf{X}C)^{P(\neg\mathbf{X}C)}P(\neg\mathbf{X}\neg C)^{P(\neg\mathbf{X}\neg C)}}{P(\mathbf{X})^{P(\mathbf{X})}P(\neg\mathbf{X})^{P(\neg\mathbf{X})}P(C)^{P(C)}P(\neg C)^{P(\neg C)}}$$

Conditional mutual information for evaluating rule $\mathbf{XQ} \rightarrow C{=}c$ given $\mathbf{X}$ in the value-based interpretation is

$$MI_C{=}\log \frac{P(\mathbf{X})^{P(\mathbf{X})}P(\mathbf{XQ}C)^{P(\mathbf{XQ}C)}P(\mathbf{XQ}\neg C)^{P(\mathbf{XQ}\neg C)}P(\mathbf{X}\neg\mathbf{Q}C)^{P(\mathbf{X}\neg\mathbf{Q}C)}P(\mathbf{X}\neg\mathbf{Q}\neg C)^{P(\mathbf{X}\neg\mathbf{Q}\neg C)}}{P(\mathbf{XQ})^{P(\mathbf{XQ})}P(\mathbf{X}\neg\mathbf{Q})^{P(\mathbf{X}\neg\mathbf{Q})}P(\mathbf{X}C)^{P(\mathbf{X}C)}P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)}}$$

Here log **is the 2-based logarithm. Note that in the task the thresholds are given for** $n \cdot MI$**, where** n**=data size.**
Note that mutual information doesn't differentiate between positive and negative dependencies. Therefore you need other means to find out if (conditional) dependence is positive or negative.

Appendix 2: Definitions of maximal, closed and free frequent sets

Let $\mathbf{X}$ be a frequent itemset, i.e., $P(\mathbf{X}) \geq min_{fr}$. $\mathbf{X}$ is

- maximal, if for all $\mathbf{Y} \supsetneq \mathbf{X}$ $P(\mathbf{Y}) < min_{fr}$;
- closed, if for all $\mathbf{Y} \supsetneq \mathbf{X}$ $P(\mathbf{Y}) < P(\mathbf{X})$;
- free, if for all $\mathbf{Y} \subsetneq \mathbf{X}$ $P(\mathbf{Y}) > P(\mathbf{X})$.

Note: $\mathbf{X} \subsetneq \mathbf{Y}$ means that $\mathbf{X}$ is a proper subset of $\mathbf{Y}$ (excludes $\mathbf{X} = \mathbf{Y}$).