Introduction to Data Mining
Assignment Due Time: 23:55 December 31, 2022

November 28, 2022

1 Number of Association Rules

The total number of possible rules that can be extracted from a transaction set that contains d items is,

$$R = 3^d - 2^{d+1} + 1 \tag{1}$$

Please prove the above equation.

2 FP-Growth and Pattern Evaluation

Table 1: Example of Transactions

Transaction ID	Items
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

a) Construct the FP-Tree for the transactions in Table 1.

b) Apply the FP-growth algorithm to find frequent itemsets ending in e (minimum support: 0.2)

c) Draw a contingency table for the following rules: $\{b\} \rightarrow \{c\}, \{a\} \rightarrow \{d\}, \{c\} \rightarrow \{a\}$

d) Using the contingency tables in (c) to compute the following evaluation measures.

i) Support and Confidence.

ii) Interest$(X \rightarrow Y) = \frac{P(X,Y)}{P(X)P(Y)}$

iii) Correlation and odds ratio.

3 Pattern Evaluations

Read the following paper and write summarization for the paper by answering the following questions:

- Wu, J., Zhu, S., Xiong, H., Chen, J., & Zhu, J. (2012). Adapting the right measures for pattern discovery: A unified view. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(4), 1203-1214.

a) What is the data mining problem the paper solve? b) What are the main contributions of the paper? How they differ from the prior work? Are they significant?

c) Do you think the proposed method is applicable? Where might be the application scenario?

d) What insights have you obtained from the paper? What future work might be done along the line?

4 Decision Tree

Table 2 summarizes a dataset with three attributes A, B, C and two class labels +,-. Each attribute is a binary variable with values T and F. Build a two-level decision tree.

Table 2: Dataset

A	**B**	**C**	**Number of +**	**Number of -**
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

a) According to the misclassification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in misclassification error. How about the gains in entropy and Gini index?

b) Repeat for the two children of the root node.

c) How many instances are misclassified by the resulting decision tree?

5 SVM

Derive the dual Lagrangian for the liner SVM with non-separable data where the objective function is,

$$f(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + C(\sum_{i=1}^{N} \xi_i)^2 \tag{2}$$

also write down the KKT conditions for this objective function.

6 Naive Bayes

Use Naive Bayes to classify the dataset in Table 2.

a) Estimate the conditional probabilities for $P(A = T|+)$, $P(B = T|+)$, $P(C = T|+)$, $P(A = T|-)$, $P(B = T|-)$, $P(C = T|-)$

b) Given a test sample $(A = T, B = T, C = T)$, predict the class sample using the Naive Bayes approach.

7 Case Study I: Association Rules

Exeter, Inc. is a catalog firm that sells products in a number of different catalogs that it owns. The catalogs number in the dozens, but fall into nine basic categories: Clothing, House wares, Health, Automotive, Personal electronics, Computers, Garden, Novelty gift, and Jewelry. The costs of printing and distributing catalogs are high. By far the biggest cost of operation is the cost of promoting products to people who buy nothing. Having invested so much in the production of artwork and printing of catalogs, Exter wants to take every opportunity to use them effectively. One such opportunity is in cross selling - once a customer has "taken the bait" and purchases one product, try to sell them another while you have their attention.

Such cross promotion might take the form of enclosing a catalog in the shipment of a purchased product, together with a discount coupon to induce a purchase from that catalog. Or it might take the form of a similar coupon sent by e-mail, with a link to the Web version of that catalog. But which catalog should be enclosed in the box or included as a link in the e-mail with the discount coupon? Exter would like it to be an informed choice - a catalog that has a higher probability of inducing a purchase than simply choosing a catalog at random.

8 Case Study II: Document Classification

In this case, we have two categories of emails, in which one category is about hockey and the other is about baseball. The data is in the folder **classification**.

a) Firstly preprocess the documents into numerical data (Record data). The preprocessing guidelines can be found in the **introduction slides (SMO)**, consider using $tf - idf$ (referring to Question 1).

b) Use SVMs to classify the documents and test the classification results with 5-fold cross validation. You should report the precision, recall, and F1-measure of each fold and the average values. (*Recommend LIBSVM to implement SVMs. You can refer to the tutorial slides in evaluating the results.*)

c) **Bonus (5 extra points)**. Implement Sequential Minimal Optimization (SMO) by following the introductive slides.

9 Case Study III: Clustering for Amyotrophic Lateral Sclerosis (ALS)

This case-study examines the patterns, symmetries, associations and causality in a rare but devastating disease, amyotrophic lateral sclerosis (ALS). ALS demands conducting clinical trials and collecting big, multi-source and heterogeneous datasets that can be interrogated to derive potential biomarkers. Overcoming many scientific, technical and infrastructure barriers is required to establish complete, efficient, and reproducible protocols (pipelines/workflows) starting with acquiring raw data, preprocessing, aggregation, harmonization, analysis, visualization and result interpretation. The dataset is in the folder **clustering**.

The clinical data shows that the rate of ALS progression varies significantly among patients. Majority of the patients die within 3 to 5 years after ALS onset, however, a few are able survive for over 10 years. This heterogeneity of disease course hinders demonstration of its biological mechanism and development of effective treatment. We need to develop reliable predictive models of ALS progression to understand the pathophysiology of the disease.

Now perform the clustering analysis according to the following procedures for the ALS dataset.

- Load and prepare the data.
- Perform summary and preliminary visualization.
- Train a K-Means model on the data, select an appropriate K by using the Elbow method. (Please refer to the following procedures: *1. Compute clustering algorithm (e.g., K-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters. 2. For each k, calculate the total within-cluster sum of squared errors (SSE). 3. Plot the curve of SSE according to the number of clusters k. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.*)
- Evaluating the model performance by report the center of clusters and explain details. You can also interpret the clustering results by inspecting the data near the center.

Reference: Tang, M., Gao, C, Goutman, SA, Kalinin, A, Mukherjee, B, Guan, Y, and Dinov, ID. (2018) Model-Based and Model-Free Techniques for Amyotrophic Lateral Sclerosis Diagnostic Prediction and Patient Clustering, Neuroinformatics, 1-15, DOI: 10.1007/s12021-018-9406-9.

Using the dataset to perform an association rules analysis, and comment on the results. Your discussion should provide interpretations the meanings of the various output statistics (lift ratio, confidence, and support) and include a very rough estimate of the extent to which this will help Exeter make an informed choice about which catalog to cross-promote to a purchaser.