Data Mining

– Midterm, April 21, 2016 –

1. (30%) Answer the following questions with respect to matrix decomposition.
   (a) Compare/contrast the SVD and NFM methods.
   (b) Decompose a matrix [[5,3,0,1],[4,0,0,1],[1,1,0,5],[1,0,0,4]] into [[2.21, 0.13], [1.76, 0.18],[0.32,1.94],[0.36,1.55]]*[[2.24,1.33,0.56,0.31],[0.13,0.3,0.12,2.5]] by the NFM. What are the meanings of the rows or columns in these three matrices?
   (c) Map both vectors [1,1,0,0] and [1,0,1,0] into the transformed space by using the matrices shown in (b)?

2. (30%) Given the following transactions T_1={a,b,c,d,e}, T_2={b,c,d,e}, T_3={a,b,c,d}, T_4={a,b,c} and T_5={a,b}, where Sup_{min}=40% and $Conf_{min}$=60%,
   (a) Mine all frequent patterns.
   (b) Find all closed frequent patterns.
   (c) Find all association rules whose confidence levels are between 60% and 80%.

3. (20%) Answer the following questions with respect to discretization.
   (a) What is the entropy-based discretization method? What is the natural partitioning method?
   (b) Compare/contrast both methods.

4. (20%) Predict if there is a game by the Bayesian classification method by using the following dataset.
   (a) when temperature is hot, humidity is low and wind is light.
   (b) when temperature is cold and wind is strong.

Temperature	Humidity	Wind	Rain	Game
hot	low	light	yes	no
mild	low	strong	no	yes
cold	low	light	no	yes
cold	high	strong	yes	no
hot	high	light	no	yes

1. (i.

ⓐ

$S = (1, -1, 2, -2, -3, 3, -4, 4)$ $L[i] = \frac{S[2i]+S[2i+1]}{\sqrt{2}}$ $H[i] = \frac{S[2i]-S[2i+1]}{\sqrt{2}}$

$\Rightarrow (0, 0, 0, 0, \sqrt{2}, 2\sqrt{2}, -3\sqrt{2}, -4\sqrt{2})$

15

ⓑ

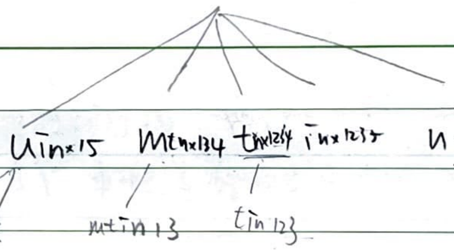
我們可以觀察到 x'為x, x''為x'的近似 但是當做一次轉換時 ①因為是將每兩項相加除√2, 有它取平均的效果 會使得局部的變化縮小 像ⓐ小題的原本可能有(1,-1), (2,-2)的振幅 透過 Haar wavelet transform 後 只剩下0. 少了震盪的幅度 ③而因為他是每兩個一組 每兩個一組 所以做HWT後的近似值 LPF 仍可以保有 sequence 的變化 (振盪平衡矣 振盪的情況), 轉換後 長度 距離 內積 均相同 (1/2)倍長度) ~~較數學的關係~~

趨勢的線性會保留 (縮減長度)

ⓐ

2

n	t	u	i	m
1	1	1	1	1
2	2	5	2	3
3	3		3	4
4	4		5	
5				



min_support = 40% ×5 = 2.

confidence = 60%

18

closed patterns: n, in, tn, tin, mtn, uin, mtin.

ⓑ

n → i 4/5 = 80%   i → n 4/4 = 100%   t → n 4/4 = 100%   m → tn 3/3 = 100%   u → in 2/2 = 100%

n → t 4/5 = 80%   i → tn 3/4 = 75%   t → in 3/4 = 75%   m → tin 2/3 = 67%

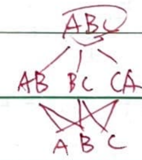
n → mt 3/5 = 60%   ~~i → un 2/4 = 50%~~   t → mn 3/4 = 75%   ~~mt → n 3/3 = 100%~~

~~n → ui 2/5 = 40%~~   ~~i → mtn 2/4 = 50%~~   ~~t → min 2/4 = 50%~~

n → ti 3/5 = 60%   ~~i → u 2/4 = 50%~~

~~n → mti~~ 2/5 = 40%   ~~t → m~~ 2/4 = 50%

~~n i → u~~ 2/4 = 50%   ~~it → mn = 2/3 = 67%~~ it → mn = 2/3 67%



~~A→BC~~ ⇒ B→CA C→AB

Association rules

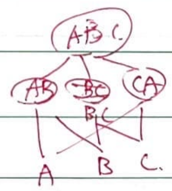
n → i   i → n   t → n   m → tn   u → in.

n → t   i → tn   t → in   m → tin

n → mt   t → mn

n → ti

it → mn
im → tn
it → n



⇒ ~~A→B~~ A→B B→A C→A BC→A
A→C B→AC C→AB
A→BC

① 第一層只做 closed.

② 往上做. X→Y, X只找非 closed pattern.

3. Apriori 是一個來找 frequent pattern 常用的 algorithm，這個演算法的最大原則是 "如果一個 itemset 不是 frequent，則它的 super set 也不會是 frequent"，所以從最小的長度看起，慢慢往上推疊，可以避免掉必須 traverse 2^n-1 種組合的情況，e.g. AB 不是 frequent，則 ABC, ABD ... 都不用考慮。

15

Partition 的做法是首先將 data set 分塊，分別找出在各個 partition 裡的 local frequent pattern，再將這些 local frequent pattern 拼湊驗證為 global frequent。其原理是 "當一個 set 為 global frequent，則它必為某個 partition 的 local pattern"，可用反證法得證，假設全部有 N 筆，$sup_{min}=a\%$，假設拆 10 塊每塊 $\frac{N}{10}$ 筆，假設存在一個 set 皆為 local，但為 global 則 $\left(\frac{N}{10}\times a\%-1\right)\times 10 \geq N\times a\%$ 需成立，但明顯不行。而檢查 local 是否為 global 的好處也是指 partition 後合起，將某種 set 都分在一起成 local 但其實 global 時，不為 frequent。

至於這兩者的最大差別，在於計算上的速度，Apriori 從最小長度一直往上算，每次計算就要 traverse 整個 data set，計算成本很高，然而 partition 只要一開始 partition 時 traverse 一次，之後再做一次 traverse 檢查 local 是否為 global，總共只需兩次。

計算 global 時 candidate 量會暴增。

4. Decision tree：可以利用 Entropy 或是 Gini Index 來找到最代表性的 attribute 以此 attribute 做為樹幹分支的節點，一直做下去，直到將資料分離的夠仔細。 divid and conquer, top-down.

naive Bayesian：利用机率模型計算各個結果的可能性，最後根據最大的机率來指定分類。

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}$$，比較各個 class 的 $P(X|C_i)P(C_i)$

12

naive Bayesian 利用机率模型計算，有理論基礎，計算也簡易，但是他的前提是假設是各 attribute 之間是獨立的，e.g. 年紀、薪資，這個假設其實是有問題的，因為 attribute 間可能仍有有 dependency，而像 decision tree 雖然做法有點像 greedy 找最好，但是沒有強的假設前提。

Decision tree 需要看過資料，Naive 不需要資料，處理 sparse。

Decision tree 可以產生 tree rule。

Decision tree 快 entropy，time complexity 小

Naive，Incremental

記分欄 轉頁從此開始寫起。

5. 公司可以收集兩大類型的數據加以分析. 第一大類是每個 user 對於各 video 的評分. 第二大類是收集 user 的 information 及 user 所瀏覽過的 video list, 以下分別對這兩類資料提供 mining method 及可應用的方式.

① 評分. 我們可以在收集到評分(0~5分)時. 利用 NMF 的就, 拆成兩個矩陣這樣會得到兩種結果. ① 使用者對看過的 video 的預期評價. ② 使用者對於新 feature (video 類型 e.g. 搞笑. 動作..) 的偏好程度。第個結果可以應用在推薦系統上面, 藉由推薦 user 可能喜歡的 video 來留住 user. 第二個可以輔助, 當公司製作自產劇時 能夠觀察 user 的取向. 來針對取向設計劇情內容.

② user information 及瀏覽紀錄. 可以利用 naive Bayesian 机率模型的分析, 得到各族背景的使用者 (e.g. 年紀. 性別) 會偏好各族類型的劇. 這個分析結果 可以運用在廣告引入系統時. 能延在使用者方面. 公司可能要提供免費. 或是降低費用門檻. 但是因為可以知道觀看此 video 的使用者族群背景. 所以可能向外招廣告投放. 因為有精準的 TA. 更可以吸引廣告主在此平台上投放廣告.

太注重在分析階段. 可以多描述一些在 pre-process 時的工作.

Concept hierarchy 來做資料分群.

資料離散化.

Data Mining

– Midterm, April 26, 2018 –

1. (30%) Consider a book review matrix R, where each entry represents the review score for each book, and the review score of each user is denoted as one row. Decompose R into three matrices by SVD, U={{-0.4, 0.6, -0.2, 0.1, -0.7}, {-0.4, -0.2, -0.8, 0.3, 0.3}, {-0.5, 0.4, 0.2, -0.5, 0.5}, {-0.6, -0.6, 0.2, -0.3, -0.4}, {-0.4, 0, 0.5, 0.7, 0.2}}, Λ={{2.5, 0, 0, 0, 0}, {0, 0.8, 0, 0, 0}, {0, 0, 0.7, 0, 0}, {0, 0, 0, 0.4, 0}, {0, 0, 0, 0, 0.25}}, V^T={{-0.4, -0.2, -0.5, -0.7, -0.3}, {-0.6, 0, -0.2, 0.2, 0.7}, {-0.4, -0.5, 0.8, -0.2, 0}, {-0.5, 0, 0, 0.6, -0.6}, {-0.2, 0.9, 0.3, -0.3, -0.1}}
   (a) What are the axes of the new data space if the dataset is transformed into a three-dimensional space?
   (b) How/what is the original data transformed into a three-dimensional space?
   (c) Transform (0, 1, 1, 0, 1) into a three-dimensional space.

2. (20%) Given the following closed patterns J:5, JU:4, JUN:3 and JUNE:2, where the number after the colon denotes the support of the closed pattern,
   (a) restore all frequent patterns and their supports.
   (b) Find all association rules, where the minimum confidence threshold is 80%.

3. (20%) Given the following sequence database {<MAY>, <M(MA)(MAY)A>, <M(MAY)A>},
   (a) find all sequential patterns, where the minimum support threshold is 60%.
   (b) Find all association rules, where the minimum confidence threshold is 80%.

4. (30%) Answer the following questions with respect to an online food ordering and delivery platform (such as Grubhub, UberEATS), which provides an online and mobile platform for restaurant pick-up and delivery orders. Describe the content of data that might be collected. Then, design a framework to mine some business insights from the dataset, and discuss the expected results and possible value-added services.

Data Mining

– Midterm, April 25, 2019 –

1. (20%) How does SVD work? How is it used to reduce the number of dimensions? What are the axes of the reduced space? How is a new data record transformed into the reduced space?

2. (20%) Design a neural network to classify customers into 3 classes, where each customer profile contains 20 attributes. What will you concern when choosing the number of hidden layers and the number of nodes in each hidden layer?

3. (20%) Why are the sigmoid and hyperbolic tangent functions often used as the activation function in a neural network? Compare/contrast both functions.

4. (20%) What is LSTM? What is each function in LSTM used for?

5. (20%) What is the philosophy of devising SVM? How does SVM resolve a non-linear classification problem?

Data Mining

– Midterm, April 29, 2021 –

1. (30%) Answer the following questions with respect to the non-negative matrix factorization (NMF) method, where an $n \times m$ matrix V is decomposed into WH, each column of V denotes a record of data, W is an $n \times k$ matrix and H is a $k \times m$ matrix.
   (a) What do the columns (or rows) of W and H stand for?
   (b) How do you apply the NMF to deal with the data sparsity?
   (c) How do you obtain the representation of a new vector $X^T=(x_1, x_2, \ldots, x_n)$ in the transformed space?

2. (30%) Answer the following questions with respect to the decision trees.
   (a) Why is the ID3 decision tree method biased towards a multi-level attribute in attribute selection?
   (b) How can the C4.5 decision tree method resolve the problem described in 2(a)?
   (c) Why does the C4.5 prefer the attribute with unbalanced splits?

3. (40%) Answer the following questions with respect to the neural networks.
   (a) (10%) Why is the sigmoid or hyperbolic tangent function frequently used in a neural network model?
   (b) (10%) When will the ReLU activation function be used? Why?
   (c) (20%) How are the biases and weights updated in each iteration of the training process of a multi-layer feed forward model if the hyperbolic tangent function is used?

SVD

$$
\underset{5\times 3}{U}
\begin{bmatrix}
-0.4 & 0.6 & -0.2 & 0.1 & -0.7 \\
-0.4 & -0.2 & -0.8 & 0.3 & 0.3 \\
0.5 & 0.4 & 0.2 & -0.5 & 0.5 \\
0.6 & -0.6 & 0.2 & -0.3 & -0.4 \\
-0.4 & 0 & 0.5 & 0.7 & 0.2
\end{bmatrix}
\underset{3\times 3}{\Lambda}
\begin{bmatrix}
2.5 & 0 & 0 & 0 & 0 \\
0 & 0.8 & 0 & 0 & 0 \\
0 & 0 & 0.7 & 0 & 0 \\
0 & 0 & 0 & 0.4 & 0 \\
0 & 0 & 0 & 0 & 0.3
\end{bmatrix}
\underset{5\times 5}{V^T}
\begin{bmatrix}
-0.4 & -0.2 & -0.5 & -0.7 & -0.3 \\
-0.6 & 0 & -0.2 & 0.2 & 0.7 \\
-0.4 & -0.5 & 0.8 & -0.2 & 0 \\
-0.5 & 0 & 0 & 0.6 & -0.6 \\
-0.2 & 0.9 & 0.3 & -0.3 & -0.1
\end{bmatrix}
$$