

Data Mining

-Course Revision

INFS4203 / INFS7203 Data Mining

1

Assessment

Assessment Task	Weighting
Final Examination	60%
Individual Assignments	20% (5% x 4 assignments)
Project	20% 5% report, 5% presentation, 10% overall

60 marks

INFS4203 / INFS7203 Data Mining

2

Final Exam

7 Questions / 60 Marks

1	Introduction to Data Mining and Data Issues 1) data mining applications
2	Association Rules Mining 1) Apriori Algorithm 2) FP-tree
3	Classification 1) Decision Tree 2) Bayesian Theorem
4	Clustering 1) K-means Clustering 2) Agglomerative Hierarchical Clustering
5	Text Mining 1) VSM Model
6	Web Mining

INFS4203 / INFS7203 Data Mining

3

Introduction to Data Mining

INFS4203 / INFS7203 Data Mining

4

Major research topics in DM

- Association Rule Mining
 - Frequent patterns, associations, correlations, or causal structures among sets of items or objects
 - Market Basket Analysis
- Classification
 - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute
 - Document topic estimation

INFS4203 / INFS7203 Data Mining

5

Major research topics in DM

- Clustering
 - Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups
 - search result grouping/reduce the information redundancy
- Text Mining
 - Information retrieval and natural language processing
 - Document search
- Web Mining
 - techniques to automatically discover and extract information from Web documents/services
 - Web Community

INFS4203 / INFS7203 Data Mining

6

Some More Applications...

- Data analysis and decision support
 - Market analysis and management
 - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
 - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
 - Text mining (news group, email, documents) and Web mining

INFS4203 / INFS7203 Data Mining

7

A Motivating Example: Market Basket Analysis

Anything interesting?

Bread → Milk (100%)
Diapers → Beer (66%)
Diapers → Milk (100%)

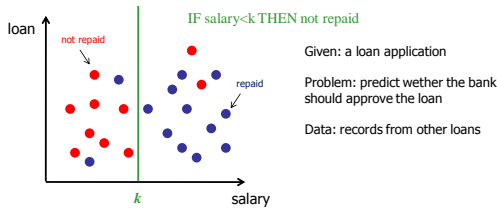
Customers who buy diapers also tend to buy beer

Identify potential cross-selling opportunities among related items

INFS4203 / INFS7203 Data Mining

8

An Example: Credit Risk



INFS4203 / INFS7203 Data Mining

9

An Example: Credit Card Fraud Detection

Monthly Transactions

User info. (age, salary, etc.)

A profile of legitimate transactions

compare

Very different!

FRUSTRATED Potentially fraudulent!

Considering each transaction as a data point, clustering can be applied to detect the outlier of the entire dataset.

INFS4203 / INFS7203 Data Mining

10

Association Rule Mining

Apriori Example

Min_Sup = 2

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Itemset	Support Count
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

Itemset	Support
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

Create tables like these.

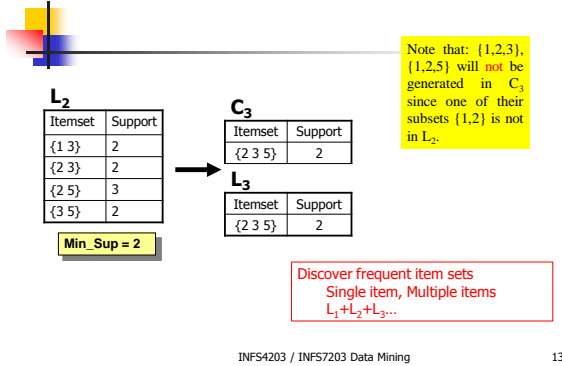
INFS4203 / INFS7203 Data Mining

11

INFS4203 / INFS7203 Data Mining

12

Apriori Example



INFS4203 / INFS7203 Data Mining

13

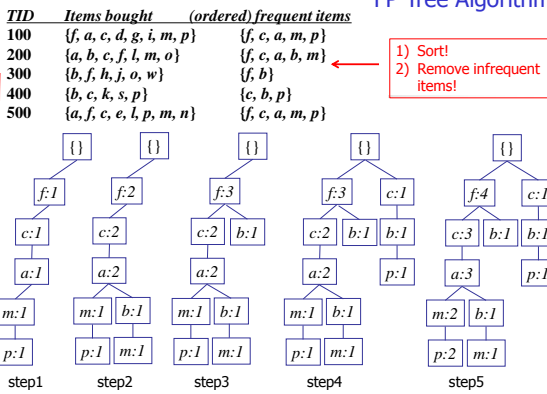
Key points

- Frequent item set
 - Single item, Multiple items
 - $L_1+L_2+L_3...$
- Support and confidence
 - How to calculate: $Confidence(X \Rightarrow Y) = P(Y | X) = P(X \cup Y) / P(X)$
- Strong association rules
 - Generated from $L_1, L_2, L_3...$
 - confidence \geq min_conf
 - $A \rightarrow B, B \rightarrow A, A \rightarrow BC, BC \rightarrow A...$

INFS4203 / INFS7203 Data Mining

14

FP Tree Algorithm



INFS4203 / INFS7203 Data Mining

15

Key points

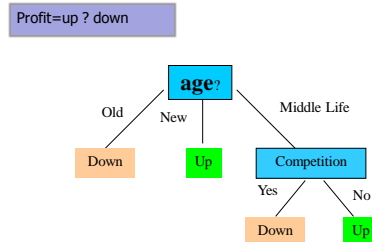
- FP tree construction
 - Step by step
- Ordered frequent items
 - Frequent items only!
- Strong association rules
 - Not required

INFS4203 / INFS7203 Data Mining

16

Classification

Decision tree



INFS4203 / INFS7203 Data Mining

17

INFS4203 / 7203 Data Mining

18

The Rules Derived from the Tree

	IF	THEN
Rule1	age is old	profit is down
Rule2	age is new	profit is up
Rule3	age is midlife AND competition is no	profit is up
Rule4	age is midlife AND competition is yes	profit is down

INFS4203 / 7203 Data Mining

19

- Tree → Rules
- Rules → Tree

INFS4203 / INFS7203 Data Mining

20

Information Entropy

$$H(C) = -\sum_{i=1}^n p(c_i) \times \log_2(p(c_i))$$

One attribute value, e.g.: old

$$H(C | a_j) = -\sum_{i=1}^n p(c_i | a_j) \times \log_2(p(c_i | a_j))$$

One class label, e.g.: up

$$H(C | A) = \sum_{j=1}^m [p(a_j) \times H(C | a_j)]$$

$$\min_{t=1}^n \{H(C | A_t)\}$$

One attribute, e.g.: Age

INFS4203 / 7203 Data Mining

21

Step 1:
H(Profit|Age), H(Profit|Competition), H(Profit|Type)

	Profit	Age	Competition	Type	
Age	old	down	old	no	software
		down	old	no	hardware
		down	old	yes	software
	new	up	new	no	hardware
		up	new	no	software
		up	new	yes	software
	midlife	down	midlife	yes	software
		up	midlife	no	hardware
		up	midlife	no	software
down		midlife	yes	hardware	

INFS4203 / 7203 Data Mining

22

- 1) Calculate H(C|A) at different levels
- 2) Just consider the points in sub-table!

Step 2:
H(Profit|Competition), H(Profit|Type) in sub-table

	Profit	Profit	Competition	Type	
Age	old	down	no	software	
		down	no	hardware	
		down	yes	software	
	new	up	no	hardware	
		up	no	software	
		up	yes	software	
middle life	competition	no	up	no	hardware
		no	up	no	software
	yes	down	yes	hardware	
	yes	down	yes	software	

INFS4203 / 7203 Data Mining

23

Key Points for Decision Tree

- Information Entropy Calculation
- Find the rules from a decision tree
- Tree construction
 - Calculate H(C|A) at different levels
 - The smaller the better!
 - Majority vote or weighted linear combination

INFS4203 / INFS7203 Data Mining

24

Two Simple Combination Techniques

- Majority Vote.
 - Simple voting! This strategy always performs surprisingly good!



P. 25

Two Simple Combination Techniques (cont')

- Weighted Linear Combination.
 - If a classifier is more reliable, then we value its decision higher.
 - We will discuss how to compute the reliability of a classifier shortly.
 - Usually performs even better than Majority Vote.



P. 26

Metrics for Performance Evaluation

- Confusion Matrix:

		Prediction	
		Salmon	Not Salmon
Actual Class	Salmon	A	B
	Not Salmon	C	D

A: TP (true positive) B: FN (false negative)

C: FP (false positive) D: TN (true negative)

$$\text{Accuracy} = \frac{A + D}{A + B + C + D} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision, Recall and F-Measure

- Measuring the quality (effectiveness) of the model:

$$\text{Precision, } p = \frac{A}{A + C}$$

$$\text{Recall, } r = \frac{A}{A + B}$$

$$\text{F-measure} = \frac{2rp}{r + p}$$

		Prediction	
		Salmon	Not Salmon
Actual Class	Salmon	A	B
	Not Salmon	C	D

P. 28

Bayesian Theorem

- Given training data \mathbf{X} , posteriori probability of a hypothesis H , $P(H|\mathbf{X})$, follows the Bayes theorem

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

- Informally, this can be written as
 - posteriori = likelihood x prior/evidence
- Predicts \mathbf{X} belongs to C_i iff the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|\mathbf{X})$ for all the k classes

Naïve Bayesian Classifier: Training Dataset

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Data sample

$X = (\text{age} \leq 30,$

Income = medium,

Student = yes

Credit_rating = Fair)

	Attribute1	Attribute2	Attribute3	Attribute4	
age	income	student	credit_rating		buys_computer
<=30	high	no	fair		no
<=30	high	no	excellent		no
31...40	high	no	fair		yes
>40	medium	no	fair		yes
>40	low	yes	fair		yes
>40	low	yes	excellent		no
31...40	low	yes	excellent		yes
<=30	medium	no	fair		no
<=30	low	yes	fair		yes
>40	medium	yes	fair		yes
<=30	medium	yes	excellent		yes
31...40	medium	no	excellent		yes
31...40	high	yes	fair		yes
>40	medium	no	excellent		no

Naïve Bayesian Classifier: An Example

- $P(C_i)$: $P(\text{buys_computer} = \text{"yes"}) = 9/14 = 0.643$
 $P(\text{buys_computer} = \text{"no"}) = 5/14 = 0.357$
- 2 classes: Yes / No prior**
- Compute $P(X|C_i)$ for each class

Attribute1	$P(\text{age} = \text{"<=30"} \text{buys_computer} = \text{"yes"}) = 2/9 = 0.222$	likelyhood
age	$P(\text{age} = \text{">30"} \text{buys_computer} = \text{"no"}) = 3/5 = 0.6$	
Attribute2	$P(\text{income} = \text{"medium"} \text{buys_computer} = \text{"yes"}) = 4/9 = 0.444$	
income	$P(\text{income} = \text{"medium"} \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$	
Attribute3	$P(\text{student} = \text{"yes"} \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$	
student	$P(\text{student} = \text{"yes"} \text{buys_computer} = \text{"no"}) = 1/5 = 0.2$	
Attribute4	$P(\text{credit_rating} = \text{"fair"} \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$	
Credit	$P(\text{credit_rating} = \text{"fair"} \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$	

$$P(C_i | \mathbf{X}) = P(\mathbf{X} | C_i) P(C_i)$$

$$P(\mathbf{X} | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \dots \times P(x_n | C_i)$$

Naïve Bayesian Classifier: An Example

- $\mathbf{X} = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit_rating} = \text{fair})$
- $P(\mathbf{X} | C_i)$: $P(\mathbf{X} | \text{buys_computer} = \text{"yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$
 $P(\mathbf{X} | \text{buys_computer} = \text{"no"}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$
- posterior** $P(\mathbf{X} | C_i) * P(C_i)$: $P(\mathbf{X} | \text{buys_computer} = \text{"yes"}) * P(\text{buys_computer} = \text{"yes"}) = 0.028$
 $P(\mathbf{X} | \text{buys_computer} = \text{"no"}) * P(\text{buys_computer} = \text{"no"}) = 0.007$
- Therefore, X belongs to class ("buys_computer = yes")**

Clustering

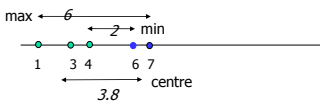
Distance Measures

- L_1, L_2, L_∞ , Jaccard Distance, Cosine Distance, Edit distance
 - How to calculate
- High-D space**
 - 2d: plane
 - H-d: dimensionality > 3
 - Data point representation: E.g., P1(0,5,1,2,4): it is a data point in a 4d space

- No complex calculations
- But work on high-d points

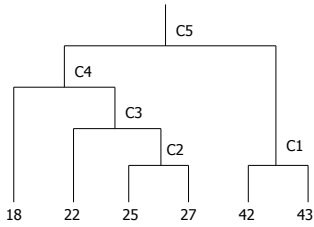
Distance between clusters

- Min
- Max
- Group average
- Centre-distance



Agglomerative Clustering

- Initially, each data point is an individual cluster
- Merge two closest clusters to generate a new cluster
- Repeat
 - Until all data points are merged into one



INFS4203 / 7203 Data Mining

P. 37

K-means Clustering

- Initially, randomly select K centroids
- 1. Calculate **distances** between data points and centroids
 - Distance function: $L_1, L_2 \dots$
- 2. Assign data points to the nearest centroids
- 3. Recalculate centroids
 - Average
- Repeat
 - Until centroids don't change

INFS4203 / INFS7203 Data Mining

38

Initial centroids: 22, 35, 43

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} Lp(m_i, x)$$

Cluster#	Old Centroid	Cluster Elements	new Centroid
1	22	18,22,25,27,28	24
2	35	33,35	34
3	43	42,43,56	47



Can be any distance Measure!
Eg., L1,L2,...

Cluster#	Old Centroid	Cluster Elements	new Centroid
1	24	18,22,25,27,28	24
2	34	33,35	34
3	47	42,43,56	47

INFS4203 / 7203 Data Mining

P. 39

Text Mining

Vector Space Model

- Each word is a dimension
 - If we have M different words. Then, we have a M -dimensional vector space.
- Each document is regarded as a point in this vector space.
 - $d = \{w_1, w_2, \dots, w_m\}$
In term of geometry, w_i is the coordinate of dimension i in d .
Yet, conceptually, w_i denotes the importance of word i in d .

41

TF-IDF Calculation

Term Importance:

$$w(word_i) = TF(word_i) \times IDF(word_i)$$

Term Frequency :

$TF(word_i)$ = number of times $word_i$ appears in the document

Inverse Document Frequency :

$$IDF(word_i) = \log \frac{\text{total documents}}{\text{document frequency}}$$

\log_{10}

42

A Running Example Step 1 – Extract text

This is a data mining course. → This is a data mining course

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mining Text mining is a subfield of data mining

Mining text is interesting, and I am interested in it. → Mining text is interesting and I am interested in it

43

A Running Example Step 2 – Remove stopwords

This is a data mining course. → This is a data mining course Stopwords are provided.

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mining Text mining is a subfield of data mining

Mining text is interesting, and I am interested in it. → Mining text is interesting and I am interested in it

44

A Running Example Step 3 – Convert all words to lowercase

This is a data mining course. → This is a data mining course

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mining Text mining is a subfield of data mining

Mining text is interesting, and I am interested in it. → mining text is interesting and I am interested in it

45

A Running Example Step 4 – Stemming

This is a data mining course. → This is a data mine course

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mine text mine is a subfield of data mine mine

Mining text is interesting, and I am interested in it. → mine interest Mining text is interesting and I am interested in it interest

46

A Running Example Step 5 – Count the word frequencies

This is a data mining course. → This is a data mine course
course:1, data:1, mine:1

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mine text mine is a subfield of data mine mine
data:1, mine:2, study:1, subfield:1, text:2

Mining text is interesting, and I am interested in it. → mine interest Mining text is interesting and I am interested in it interest
Interest:2, mine:1, text:1

47

A Running Example Step 6 – Create an indexing file

This is a data mining course. → This is a data mine course
course:1, data:1, mine:1

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mine text mine is a subfield of data mine mine
data:1, mine:2, study:1, subfield:1, text:2

Mining text is interesting, and I am interested in it. → mine interest Mining text is interesting and I am interested in it interest
interest:2, mine:1, text:1

Create a table like this.

ID	word	document frequency
1	course	1
2	data	2
3	interest	1
4	mine	3
5	study	1
6	subfield	1
7	text	2

48

A Running Example Step 7 – Create the vector space model

This is a data mining course. → This is a data mining course
 mine course
 course:1, data:1, mine:1
 (1, 1, 0, 1, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mining. Text mining is a subfield of data mining.
 study mine text mine
 data:1, mine:3, study:1, subfield:1, text:2
 (0, 1, 0, 3, 1, 1, 2)

Mining text is interesting, and I am interested in it. → Mining text is interesting and I am interested in it.
 mine interest
 interest:2, mine:1, text:1
 (0, 0, 2, 1, 0, 0, 1)

ID	word	document frequency
1	course	1
2	data	2
3	interest	1
4	mine	3
5	study	1
6	subfield	1
7	text	2

49

A Running Example Step 8 – Compute the inverse document frequency

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 1, 0, 3, 1, 1, 2)

Mining text is interesting, and I am interested in it. → (0, 0, 2, 1, 0, 0, 1)

$IDF(word) = \log \frac{\text{total documents}}{\text{document frequency}}$

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

Create a table like this.

50

A Running Example Step 9 – Compute the weights of the words

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)
 (0.477, 0.176, 0, 0, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 1, 0, 3, 1, 1, 2)
 (0, 0.176, 0, 0, 0.477, 0.477, 0.352)

Mining text is interesting, and I am interested in it. → (0, 0, 2, 1, 0, 0, 1)
 (0, 0, 0.954, 0, 0, 0, 0.176)

$w(word_i) = TF(word_i) \times IDF(word_i)$
 $TF(word_i)$ = number of times $word_i$ appears in the document

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

51

A Running Example Step 10 – Normalize all documents to unit length

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)
 (0.938, 0.346, 0, 0, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 1, 0, 3, 1, 1, 2)
 (0, 0.225, 0, 0, 0.611, 0.611, 0.450)

Mining text is interesting, and I am interested in it. → (0, 0, 2, 1, 0, 0, 1)
 (0, 0, 0.983, 0, 0, 0, 0.181)

$w(word_i) = \frac{w(word_i)}{\sqrt{w^2(word_1) + w^2(word_2) + \dots + w^2(word_n)}}$

Represent each document as a weighted vector by using TF/IDF weight scheme.

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

52

Normalization

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)
 (0.477, 0.176, 0, 0, 0, 0, 0)

$w(course) = \frac{0.477}{\sqrt{0.477^2 + 0.176^2 + 0 + 0 + 0 + 0 + 0}} = 0.938$

(0.938, 0.346, 0, 0, 0, 0, 0)

A Running Example

Everything become structural!

- We can perform classification, clustering, etc!!!!

This is a data mining course. → (0.938, 0.346, 0, 0, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 0.225, 0, 0, 0.611, 0.611, 0.450)

Mining text is interesting, and I am interested in it. → (0, 0, 0.983, 0, 0, 0, 0.181)

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

54

Similarity Measure

Vector similarity (dot product):

$$sim(Q, D) = \sum_{k=1}^r w_{qk} \cdot w_{dk}$$

Use this simple one, if you have done the normalization.

Cosine vector similarity:

$$sim(Q, D) = \frac{\sum_{k=1}^r w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^r (w_{qk})^2} \cdot \sqrt{\sum_{k=1}^r (w_{dk})^2}}$$

Otherwise ...

55

INFS4203 / INFS7203 Data Mining

56

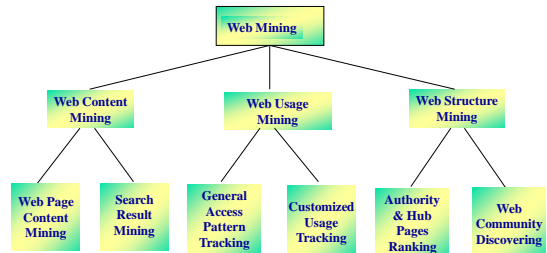
Web Mining

What is Web Mining?

- Web data mining - techniques to automatically discover and extract information from Web documents/services

57

A taxonomy of Web Mining



58

Well-known methods

- HITS (Topic distillation)
- PageRank (Ranking web pages used by Google)
- Algorithms in cyber-community

Focus on the assignment 4

59

Final Exam Timetable:

DATE	SESSION	VENUE
10/11 (Thur)	02:30pm	50-S201

Consultation Time: (or just email to us)

Fri 4/11	2pm-4pm	Yang
Thur 3/11	1pm-2:30pm	Helen



INFS4203 / INFS7203 Data Mining

60