NỘI DUNG BÁO CÁO KẾT THÚC
HỌC PHẦN KHAI PHÁ DỮ LIỆU
TS. NGUYỄN THỊ NGỌC ANH

PHẦN I. LÝ THUYẾT

I. Khai phá dữ liệu

1. Khái niệm khai phá dữ liệu

2. Các lĩnh vực liên quan tới Khai phá dữ liệu: giải thích sự liên hệ các lĩnh vực này có liên quan như thế nào tới Khai phá dữ liệu

3. Các nhiệm vụ của Khai phá dữ liệu

4. Các ứng dụng thực tế của Khai phá dữ liệu: kể tên ít nhất 3 ứng dụng, giải thích

5. Vai trò của Khai phá dữ liệu trong nền công nghiệp 4.0

6. Phân biệt được Khai phá dữ liệu với tìm kiếm thông thường. Cho ví dụ

II. Quy trình khai phá dữ liệu

1. Vẽ mô hình tổng quan Quy trình Khai phá dữ liệu

2. Trình bày nhiệm vụ của mỗi bước trong Quy trình khai phá dữ liệu. Cho ví dụ minh họa tại mỗi bước.

3. Tại bước tiền xử lí dữ liệu, làm rõ nhiệm vụ của tiền xử lí dữ liệu giải quyết những vấn đề gì? Trình bày cụ thể cho vấn đề feature extraction gồm Feature selection và Feature reduction. Khái niệm Feature selection và Feature reduction. Cho ví dụ minh họa. Trình bày thuật toán SVD/PCA cho feature reduction. Định nghĩa và nhiệm vụ của SVC/PCA là gì? Cho ví dụ minh họa cụ thể để giảm số chiều dữ liệu như thế nào cho thuật toán PCA?

4. Dịch tài liệu FeatureExtraction-Selection-Reduction từ trang 1 - trang 6. Chạy code và show dữ bài tập 3.2.1 bằng việc áp dụng PCA để minh họa. Thực hiện chạy code cho dữ liệu cô Ngọc Anh ấn định cho bài tập 3.2.1.

III. Kho dữ liệu

1. Khái niệm Kho dữ liệu.

2. Phân biệt được Kho dữ liệu với Cơ sở dữ liệu tác nghiệp. Cho ví dụ

3. Trình bày các đặc tính của Kho dữ liệu

4. Phát biểu khái niệm OLAP. Phân biệt OLTP và OLAP. Cho ví dụ

5. Trình bày kiến trúc của Kho dữ liệu

6. Trình bày quy trình xây dựng Kho dữ liệu

7. Phát biểu khái niệm mô hình hóa Kho dữ liệu. Phân biệt được sơ đồ ngôi sao, bông tuyết và chòm sao. Cho ví dụ minh họa.

IV. Học máy

1. Khái niệm học máy

2. Các cơ chế học máy: định nghĩa và làm rõ từng cơ chế này

3. Sự liên hệ của Học máy với Khai phá dữ liệu

4. Ứng dụng của Học máy

V. Phân lớp (Classification)

1. Khái niệm

2. Mục đích của Phân lớp

3. Giải thích rõ các thuật ngữ liên quan tới bài toán Phân lớp: Instance/Sample, Label, Training dataset, Testing dataset, Feature extraction, Feature Selection, Feature Reduction, Ground truth.

4. Quy trình phân lớp: Vẽ mô hình và trình bày các bước

5. Cho ví dụ và vẽ mô hình tổng quan cho bài toán ví dụ gồm:

5.1. Mô tả bài toán;

5.2. Mô tả dữ liệu;

5.3. Nhiệm vụ của bài toán phân lớp của ví dụ là làm gì?

VI. Thuật toán K-NN

1. Khái niệm/ Ý tưởng

2. Quy trình thực hiện Phân lớp dựa trên thuật toán KNN

3. Viết mã giả

4. Ví dụ minh họa: Có thể lấy bài tập ở Phần II để thực hiện phần này.

4.1. Mô tả bài toán;

4.2. Mô tả dữ liệu

4.3. Thực hiện tính tay từng bước theo Quy trình được trình bày ở phần 2. Trình bày kết quả, phân tích và kết luận.

5. Viết chương trình (source code) cho bài toán được lấy minh họa ở phần 4.

6. Trình bày Ưu/Nhược điểm của thuật toán KNN

7. Ứng dụng thực tế: Nghiên cứu ứng dụng thuật toán KNN trong việc chẩn đoán bệnh tim của bệnh nhân. Thực hiện theo yêu cầu như sau:

7.1. Tự viết code:

7.1.1 Xác định tình trạng bệnh cho một bệnh nhân chưa biết dựa trên bộ dữ liệu huấn luyện cho sẵn.

7.1.2 Chia dữ liệu đã được cho sẵn: trong đó 80% dữ liệu là dùng cho để huấn luyện và 20% còn lại để kiểm thử. Dùng phương pháp confusion matrix để tính hiệu năng của thuật toán KNN khi áp dụng cho dữ liệu này. Trình bày kết quả, phân tích và kết luận.

7.2 Dùng tool Weka

7.2.1 Xác định tình trạng bệnh cho một bệnh nhân chưa biết dựa trên bộ dữ liệu huấn luyện cho sẵn.

7.1.2 Chia dữ liệu đã được cho sẵn, trong đó 80% dữ liệu là dùng cho để huấn luyện và 20% còn lại để kiểm thử. Show kết quả thông qua bảng confusion matrix để tính hiệu năng của thuật toán KNN khi áp dụng cho dữ liệu này. Trình bày kết quả, phân tích và kết luận.

7.2.3. Biểu diễn dữ liệu (visulization) thông qua scatter plot.

VII. Thuật toán Naïve Bayes: Thực hiện yêu cầu giống KNN từ phần 1-6. Tuy nhiên, phần ví dụ thì lấy yêu cầu bài tập ở Phần II để thực hiện phần này.

VIII. Thuật toán Decision tree: C4.5 và ID3: Thực hiện yêu cầu giống KNN từ phần 1-6. Tuy nhiên, phần ví dụ thì lấy yêu cầu bài tập ở Phần II để thực hiện phần này.

IX. Phân cụm

1. Khái niệm

2. Mục đích của Phân cụm

3. Trình bày sự khác nhau giữa Phân cụm và Phân lớp: khái niệm, cơ chế học, mục đích.

X. Thuật toán K-means: Thực hiện yêu cầu giống KNN từ phần 1-6. Tuy nhiên, phần ví dụ thì lấy yêu cầu bài tập ở Phần II để thực hiện phần này.

XI. Luật kết hợp

1. Khái niệm

2. Các miền ứng dụng

3. Thuật toán APRIORI:

3.1. Trình bày ý tưởng

3.2. Cho ví dụ minh họa: Phát biểu bài toán, mô tả dữ liệu và trình bày cách thực hiện.

3.3. Ưu/Nhược điểm của thuật toán

4. Thuật toán FP-GROWTH:

4.1. Trình bày ý tưởng

4.2. Cho ví dụ minh họa: Phát biểu bài toán, mô tả dữ liệu và trình bày cách thực hiện.

4.3. Ưu/Nhược điểm của thuật toán

XII. Đánh giá mô hình

1. Các phương thức để xác định hiệu năng của mộ hình Phân lớp/Phân cụm

2. Phương pháp Confusion matrix: định nghĩa, giải thích các tham số

3. Các tính độ chính xác mô hình thông qua confusion matrix

PHẦN II: THỰC HÀNH

Yêu cầu chung:

1. Thực hiện tính tay và trình bày vào file báo cáo .pptx từng bước thực hiện cụ thể. Nếu tính kết quả trên excel, nộp file báo cáo này kèm.

2. Viết chương trình (soure code).

3. Sử dụng tool WEKA (không bắt buộc).

Ex1: Given a collection dataset as following: Using KNN method

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

Apply KNN to classify the data point z = 5.0 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).

Ex2: Given a collection dataset as following: Using KNN method

X1	X2	Class
4	3	+
3	7	+
7	***4***	**+**
4	1	+
6	5	+
5	6	+
3	7	+
6	2	+
4	6	-
4	4	-
5	8	-
7	8	-
7	6	-
4	10	-
9	7	-
5	4	-
8	5	-
6	6	-
7	4	-
8	8	-
7	5	?????

Tranning data

query point

(a) Classify the query point s_{query}=(7,5) according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).

(b) Using the training dataset in Ex2, dividing 80% for training and 20% for testing. Using kNN method for this dataset, how about classification accuracy with k = 1-9 (odd number) (using majority vote). Applying the confusion matrix to evaluate the model performance.

Ex3: Applying KNN on real Iris dataset, show experimental results for classification problem on the two strongest features and full features. Discuss and conclude on those results. You can download dataset directly from

https://archive.ics.uci.edu/ml/datasets/iris

Note: To choose the two strongest features, you can visualize a couple of features in 2D via using scatter plot and get the best one.

Ex4: Using Naïve Bayes:

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

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

(b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample $(A = 0, B = 1, C = 0)$ using the naïve Bayes approach.

Ex5: Using Naïve Bayes

ID	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes

ID	Outlook	Temp	Humidity	Windy	Play
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No
15	**Sunny**	**Cool**	**High**	**True**	**???**

Hint:

Outlook			Temp			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	**2/9**	**3/5**	Hot	**2/9**	**2/5**	**High**	**3/9**	**4/5**	**False**	**6/9**	**2/5**	**9/14**	**5/14**
Overcast	**4/9**	**0/5**	Mild	**4/9**	**2/5**	**Normal**	**6/9**	**1/5**	**True**	**3/9**	**3/5**		
Rainy	**3/9**	**2/5**	Cool	**3/9**	**1/5**								

Ex 6: Decision tree

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

(a) What is the entropy of this collection of training examples with respect to the positive class?

(b) What are the information gains of a_1 and a_2 relative to these training examples?

(f) What is the best split (between a_1 and a_2) according to the Gini index?

Ex7. Decision tree

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Consider the training examples shown in Table for a binary classification problem

a) Compute the Gini index for the overall collection of training examples.

b) Compute the Gini index for the Customer ID attribute.

c) Compute the Gini index for the Gender attribute.

d) Compute the Gini index for the Car type attribute using multiway split.

e) Compute the Gini index for the Shirt Size using multiway split.

f) Which attribute is better, Gender, Car Type or Shirt size.

Ex.8: Decision tree

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Consider the training examples shown in Table for a binary classification problem
a) What is the entropy of this collection of training examples with respect to the positive class?
b) What is the best split (among a_1, a_2, a_3) according to the information gain?
c) What is the best split (between a_1 and a_2) according to the Gini index?

Ex 7: K-means

Use the k-means algorithm and Euclidean distance to cluster the following 8 samples with k = 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Visulization after clustering.

Ex 8: Association rules

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

$$s(\{e\})$$

$$s(\{b, d\})$$

$$s(\{b, d, e\})$$

(b) Use the results in part (a) to compute the confidence for the association rules $\{b, d\} \longrightarrow \{e\}$ and $\{e\} \longrightarrow \{b, d\}$. Is confidence a symmetric measure?

Ex 9: Association rules

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

a) compute the support for itemsets {e}, {b,d}, and {b, d, e} by treating each transaction ID as a market basket.

b) Use the results in part (a) to compute the confidence for the association rules {b,d}→ {e} and {e} → {b,d}.

Ex 10: Association rules

Consider the following set of frequent 3-itemsets:

$$\{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}, \{2, 3, 5\}, \{3, 4, 5\}.$$

Assume that there are only five items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.

(c) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

PHẦN 3: CHƯƠNG TRÌNH ỨNG DỤNG THỰC TẾ

Ứng dụng cài đặt các thuật toán toán đã học vào ứng dụng thực tế với bộ dữ liệu cho sẵn (dữ liệu cô Ngọc Anh cung cấp hoặc chọn dữ liệu từ trang UCI), yêu cầu báo cáo sản phẩm như sau:

1. Mô tả bài toán
2. Mô tả dữ liệu
3. Mô tả hệ thống được cài đặt
4. Kết quả và phân tích
5. Kết luận
6. Tài liệu tham khảo