北京航空航天大学

2021－2022 学年 第二学期期末

《决策与商务智能系统》

考 试 A 卷

任课教师：姚忠

班 级_____学 号 _____

姓 名_____成 绩 _____

考试日期:2022 年 6 月 24 日

A

班号 _____ 学号_____ 姓名_____ 成绩_____

《决策与商务智能系统》期末考试卷

注意事项：1、所有答题只有在答题纸上有效

2、除选择与正误判断题外，所有文字答题需用中文回答(专业名词可直接用英文单词作答)。

试题：

一、**True or False.** (In each of the following sentences, judging correct or not correct for the statement. If correct, writes a T, otherwise writes an F in the Answer Sheets. Each is 1 mark, total 20 marks). （ 20 分）

1. In today's business environment, creativity, intuition, and interpersonal skills are effective substitutes for analytical decision making.
2. Subject oriented databases for data warehousing are organized by detailed subjects such as disk drives, computers, and networks.
3. The use of statistics in baseball by the Oakland Athletics, as described in the Moneyball case study, is an example of the effectiveness of prescriptive analytics.
4. The data warehousing maturity model consists of six stages: prenatal, infant, child, teenager, adult, and sage.
5. One of the four components of BI systems, business performance management, is a collection of source data in the data warehouse.
6. Because the recession has raised interest in low-cost open source software, it is now set to replace traditional enterprise software.
7. The design phase of decision making is where the decision maker examines reality and identifies and defines the problem. F
8. Web-based collaboration tools (e.g., GSS) can assist in multiple stages of decision making, not just the intelligence phase.
9. Business intelligence systems typically support solving a certain problem or evaluate an opportunity, while decision support systems monitor situations and identify problems and/or opportunities, using analytic methods.
10. There are basic chart types and specialized chart types. A Gantt chart is a specialized chart type.
11. The best key performance indicators are derived independently from the company's strategic goals to enable developers to "think outside of the box.".
12. Data mining can be very useful in detecting patterns such as credit card fraud, but is of little help in improving sales. F
13. During classification in data mining, a false positive is an occurrence classified as true by the algorithm while being false in reality.
14. The k-nearest neighbor algorithm appears well-suited to solving image recognition and categorization problems.
15. In the student retention case study, support vector machines used in prediction had

proportionally more true positives than true negatives.

16. In text mining, inputs to the process include unstructured data such as Word documents, PDF files, text excerpts, e-mail and XML files.
17. In text mining, if an association between two concepts has 7% support, it means that 7% of the documents had both concepts represented in the same document.
18. Social network analysis can help companies divide their customers into market segments by analyzing their interconnections.
19. In order to be effective, analysts must use models to solve problems with no regard to the organizational culture to find optimal results.
20. An advantage of simulation is that it allows model builders to solve problems with minimal interaction with users or managers.

二、**Multiple Choice** (only one correct in each questions, each is 1 marks, total 20)（20分）

1. What data discovery process, whereby objects are categorized into predetermined groups, is used in text mining?

A) classification
B) trend analysis
C) association
D) clustering

2. Which of these applications will derive the LEAST benefit from text mining?

A) sales transaction files
B) patent description files
C) customer comment files
D) patients' medical files

3. What types of documents are BEST suited to semantic labeling and aggregation to determine sentiment orientation?

A) collections of documents
B) medium- to large-sized documents
C) large-sized documents
D) small- to medium-sized documents

4. How does blind search differ from optimization?

A) Blind search cannot result in optimal solutions whereas optimization methods do.
B) Blind search represents a guided approach while optimization is unguided.
C) Blind search usually does not conclude in one step like some optimization methods.
D) Blind search is usually a more efficient problem solving approach than optimization.

5. Environmental scanning is important for all of the following reasons EXCEPT

A) organizational culture is important and affects the model use.
B) environments have greater impact on a model than the organization does.
C) environmental factors may have created the current problem.

D) it is critical to identify key corporate decision makers.

6. What does Web structure mining involve?
A) analyzing the pattern of visits to a Web site
B) analyzing the universal resource locators in Web pages
C) analyzing the PageRank and other metadata of a Web page
D) analyzing the unstructured content of Web pages

7. Search engine optimization (SEO) is a means by which
A) Web site developers can negotiate better deals for paid ads.
B) Web site developers can increase Web site search rankings.
C) Web site developers optimize the artistic features of their Web sites.
D) Web site developers index their Web sites for search engines.

8. What types of documents are BEST suited to semantic labeling and aggregation to determine sentiment orientation?
A) collections of documents
B) medium- to large-sized documents
C) large-sized documents
D) small- to medium-sized documents

9. What do voice of the market (VOM) applications of sentiment analysis do?
A) They examine the "market of ideas" in politics.
B) They examine employee sentiment in the organization.
C) They examine customer sentiment at the aggregate level.
D) They examine the stock market for trends.

10. In sentiment analysis, which of the following is an implicit opinion?
A) The cruise we went on last summer was a disaster.
B) Our new mayor is great for the city.
C) The customer service I got for my TV was laughable.
D) The hotel we stayed in was terrible.

11. Neural networks have been described as "biologically inspired." What does this mean?
A) They are faithful to the entire process of computation in the human brain.
B) They have the power to undertake every task the human brain can.
C) They crudely model the biological makeup of the human brain.
D) They were created to look identical to human brains.

12. Why is sensitivity analysis frequently used for artificial neural networks?
A) because it is generally informative, although it cannot help to identify cause-and-effect relationships among variables
B) because some consequences of mistakes by the network might be fatal, so justification may matter

C) because it provides a complete description of the inner workings of the artificial neural network
D) because it is required by all major artificial neural networks

13. Which data mining process/methodology is thought to be the most comprehensive, according to kdnuggets.com rankings?
A) CRISP-DM
B) SEMMA
C) KDD Process
D) proprietary organizational methodologies

14. Data warehouses provide direct and indirect benefits to using organizations. Which of the following is an indirect benefit of data warehouses?
A) extensive new analyses performed by users
B) simplified access to data
C) better and more timely information
D) improved customer service

15. Which component of a reporting system contains steps detailing how recorded transactions are converted into metrics, scorecards, and dashboards?
A) assurance
B) extract, transform and load
C) data supply
D) business logic

16. Which type of question does visual analytics seeks to answer?
A) What is happening today?
B) What happened yesterday?
C) When did it happen?
D) Why did it happen?

17. Why are analytical decision making skills now viewed as more important than interpersonal skills for an organization's managers?
A) because personable and friendly managers are always the least effective
B) because interpersonal skills are never important in organizations
C) because analytical-oriented managers tend to be flashier and less methodical
D) because analytical-oriented managers produce better results over time

18. Which kind of data warehouse is created separately from the enterprise data warehouse by a department and not reliant on it for updates?
A) sectional data mart
B) volatile data mart
C) public data mart
D) independent data mart

19. In estimating the accuracy of data mining (or other) classification models, the true positive rate is
A) the ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified negatives.
B) the ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified positives.
C) the ratio of correctly classified positives divided by the total positive count.
D) the ratio of correctly classified negatives divided by the total negative count.

20. For DSS, why are semistructured or unstructured decisions the main focus of support?
A) MIS staff prefer to work on solving unstructured and semistructured decisions.
B) There are many more unstructured and semistructured decisions than structured in organizations.
C) Unstructured and semistructured decisions are the easiest to solve.
D) They include human judgment, which is incorporated into DSS.

三、**Interpretation of the concepts, methods or tools** (each is 2 marks, total in 10 marks. In your answers, each is no more than 100 Chinese words).（ 10 分）

1. Web Crawler；2. Sentiment analysis; 3. Data Mart;
4. Data Mining; 5. k-fold cross-validation.

四、**Briefly answering the following 4 questions** (total in 15 marks. In our answers, each is no more than 300 Chinese words).（ 15 分）

1. Identify, with a brief description, each of the four steps in the sentiment analysis process. （10 marks）
2. Describe k-means clustering algorithm. （5 marks）

五、**Computing question**. (12 marks) (12分)

Use the information gain to select which attribute as the branch node is the best for classification customer's loan application, the sample is listed in the following table.

Samples of Customer's Application of Loan

ID	Age	Has_Job	Own_House	Credit_rating	Class
1	young	FALSE	FALSE	Fair	No
2	young	FALSE	FALSE	Excellent	No
3	young	TRUE	FALSE	Good	Yes
4	young	TRUE	TRUE	Good	Yes
5	young	FALSE	FALSE	Fair	No
6	middle	FALSE	FALSE	Fair	No
7	middle	FALSE	FALSE	Good	No
8	middle	FALSE	TRUE	Good	Yes
9	middle	TRUE	TRUE	Excellent	Yes
10	middle	FALSE	TRUE	Excellent	Yes
11	old	FALSE	TRUE	Excellent	Yes
12	old	FALSE	TRUE	Good	Yes
13	old	TRUE	FALSE	Good	Yes
14	old	TRUE	FALSE	Excellent	Yes

15	old	FALSE	FALSE	Fair	No

Hints: $\log_2(1/5)=-2.3219$，$\log_2(2/5)=-1.3219$，$\log_2(3/5)=-0.737$，$\log_2(4/5)=-0.3219$;
$\log_2(1/3)=-1.585$, $\log_2(2/3)=-0.585$；$\log_2(1/6)=-2.5849$，$\log_2(5/6)=-0.2630$;;

The entropy formula is

$$entropy(D) = -\sum_{j=1}^{|C|} Pr(c_j)\, log_2\, Pr(c_j).\ \ s.t.\ \sum_{j=1}^{|C|} Pr(c_j) = 1,$$

Attribute A_i entropy is computed as

$$entropy_{A_i}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times entropy(D_j)$$

Information gained by selecting attribute A_i to branch or to partition the data is

$$gain(D, A_i) = entropy(D) - entropy_{A_i}(D)$$

六、**Computation questions**. （8 分）

Suppose teacher has secretly written down a string of six digits (001001). Each digit is a decision variable that can take the value of either 0 or 1. Using genetic algorithm (assume 001001 is one of chromosome) operators such as crossover and mate, to guess the number that the teacher has written down. Try to guess this number as quickly as possible (with the least number of trials). In the first step, 4 number strings will be given you as follows:

Step 1:

(A) 110100; score = 1 (i.e., one digit guessed correctly)

(B) 111101; score =

(C) 011011; score =

(D) 101100; score =

七、**Mini Case Study**. The following is a case study, one corporate operational description in their routine operational management. Reading it carefully, then answer the following questions. (each question is 5 marks, total 15marks). (15 分)

1. What does Influence Health do? (5 分)
2. What were the company's challenges, proposed solutions, and obtained results? (5 分)
3. How can data mining help companies in the healthcare industry(in ways other than he ones mentioned in this case)？(5 分)

Case Title：Influence Health Uses Advanced Predictive Analytics to Focus on The Factors That Really Influence People's Healthcare Decision.

Influence Health, Inc. provide the healthcare industry's only integrated digital consumer engagement and activation platform. It enables providers, employers, and payers to positively influence consumer decision making and health behaviors well beyond the physical care setting through personalized and interactive multichannel engagement. Since 1996, the Birminghan, Alabama-based company has helped more than 1100 provider organizations influence consumers in a way that transforms financial and quality outcomes. Healthcare is a personal business. Each patient's needs are different and require

an individual response. On the one hand--as the cost of providing healthcare services continues to rise—hospital and health systems increasingly need to harness economies of scale by catering to larger and larger populations. The challenge then becomes to provide a personalized approach while operating on a large scale. Influence Health specializes in helping its healthcare sector clients solve this challenge by getting to know their existing and potential patients better and targeting each individual with the appropriate health services at the right time. Advanced predictive technology from IBM allows Influence Health to help its clients discover the factors that have the most influence on patients' healthcare decision. By assessing the propensity of hundreds of millions of prospects to require specific healthcare services, Influence Health is able to boost revenues and response rates for healthcare campaigns, improving outcomes for its clients and their patients alike.

Targeting the Savvy Consumer

Today's healthcare industry is becoming more competitive than ever before. If the use of an organization's services drops. So do its profits. Rather than simply seeking out the nearest hospital or clinic, consumers are now more likely to make positive choice among healthcare providers. Paralleling efforts that are common in other industries, healthcare organization must make increased efforts to market themselves effectively to both existing and potential patients, building long-term engagement and loyalty.

The key to successful healthcare marketing are timeless and relevance. If you can predict what kind of health services an individual prospect might need, you can engage and influence her or him much more effectively for wellness care.

Venky Ravirala. Chief analytics officer at Influence Hearth, explains, "Healthcare organization risk losing people's attention if they bombard then with irrelevant messaging. We help our clients avoid this risk by using analytics to segment their existing and potential prospects and market to them in a much more personal and relevant way.

Faster and More Flexible Analytics

As its client base has expanded, the total volume of data in Influence Health's analytics systems has grown to include over 195 million patient records with a detailed disease encounter history for several million patients. Ravirala comments, "with so much data to analyze, our existing method of scoring data was becoming too complex and time-consuming. We wanted to be able to extract insights at greater speed and accuracy."

By leveraging predictive analytics software form IBM. Influence Health is now able to develop models that calculate how likely each patient is to require particular services and express this likelihood as a percentage score. Microsegmentation and numerous disease-specific models draw on demographic, socio-economic, geographical, behavioral, disease history., and census data and examine different aspects of each patient's predicted healthcare needs.

"IBM solution allows us to combine all these models using an ensemble technique, which helps to

overcome the limitations of individual models and provide more accurate results." Comments Venky Ravirala, chief analytics officer at Influence Health. "It gives us the flexibility to apply multiple techniques to solve the problem and arrive at the best solution. It also automates much of the analytics process, enabling us to respond to clients' requests faster than before, and often give them a much deeper level of insight into their patient population."

For example, Influence Health decided to find out how disease prevalence and risk vary between different cohorts within the general population. By using very sophisticated cluster analysis techniques, the company was able to discover new comorbidity patterns that improve risk predictability for over 100 common diseases by up to 800 percent.

This helps to reliably differentiate between high-risk and very high-risk patients---making it easier to target campaigns at the patients and prospects who need them most. With insight like these in hand, Influence Health is able to use its healthcare marketing expertise to advise its clients on how best to allocate marketing resources.

"Our clients make significant budgeting decisions based on the guidance we give them," states Ravirala. "We help them maximize the impact of one-off campaign---such as health insurance marketplace campaigns when Obamacare began---as well as their long-term strategic plans and ongoing marketing communication."

Influence Health continues to refine its modeling techniques, gaining an ever-deeper understanding of the critical attributes that influence healthcare decision. With a flexible analytics toolset at its fingertips, the company is well equipped to keep improving its service to clients. Ravirala explains, "In the future, we want to take our understanding of patient and prospect data to the next level, identifying patterns in behavior and incorporating analysis with machine-learning libraries. IBM SPSS has already given us the ability to apply and combine multiple models without writing a single line of code. We're eager to further leverage this IBM solution as we expand our healthcare analytics to support clinical outcomes and population health management services."

"We are achieving analytics on an unprecedented scale. Today, we can analyze 195 million records with 35 different models in less than two days—a task which was simply not possible for us in the past," says Ravirala.