

Statistical Data Mining

ORIE 4740
Prof. Dawn Woodard
Cornell University

1

INTRODUCE TAS

2

Outline

- 1 Motivation / Examples**
- 2 Course summary**
- 3 Overview of Learning and Prediction**
- 4 Case: Heart Disease Detection**

3

Data Mining

“[Data mining is] the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data...[it] employs pattern recognition technologies, as well as statistical and mathematical techniques.”

- the Gartner Group

5

Data mining

- Finding hidden, meaningful, and often unsuspected **information in data**
- Data mining often involves **large data sets** with many records (e.g. customers) and many variables (attributes). Desirable approaches are both meaningful and computationally tractable on these large data sets.
- **Desirable approaches have few assumptions** or are robust to the violation of those assumptions.
 - Contrast w/ intro stats class: assume normal distribution, linear relationship, etc.

6

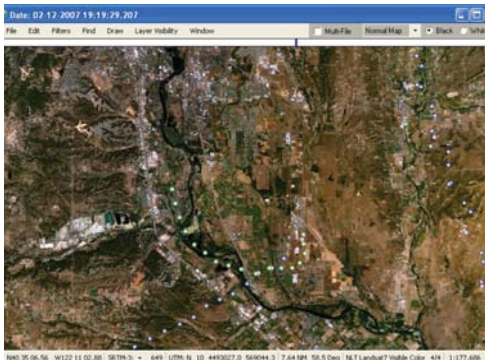
Business Applications

- Market segmentation
 - Finding groups of customers with similar purchasing habits
- Evaluating risk of credit card applicants
- Targeted marketing
 - Identifying likely purchasers
- Home valuation

7

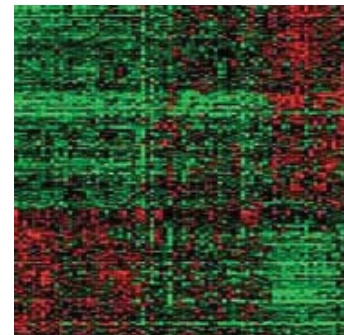
Aircraft Identification

Estimating the type of aircraft based on its location at a series of time points.



8

Identification of Cancer Genes



Finding genes that show very high or low expression in individuals with a particular cancer

9

Information Tech. Applications

- Search engine technology
- Identifying spam
- Handwriting and voice recognition
- Smart web browsers that identify ads on web pages

10

Course goals

Course Goals:

To be able to:

- Take a large data set, [decide on a set of data mining techniques](#) to address the question of interest, [apply and compare](#) those techniques, and [draw conclusions](#).
- [handle large data sets](#)
 - lack of knowledge about the relationships between the variables.
 - computational complexity
 - numerical issues
- [modify or extend an existing implementation](#) of a data mining technique.

12

Syllabus

- Review syllabus.

13

Syllabus

Are you well-versed in R, S-PLUS, Matlab, C or Java?

- A. Yes
- B. No

14

Syllabus

Have you had a course that included multiple linear regression and logistic regression?

- A. Yes
- B. No

15

Syllabus

Reading & Assignment for 1st week: Chap. 1 & 2 in SPB. Read & complete tutorial in Sections 1.7-3.4 of “Introduction to R”

16

Software

- Questions welcomed in class!
- Class participation is part of grade

17

- R statistical software (open-source).
- Has a proprietary relative, [S-PLUS](#), frequently used for statistical analysis and graphics in industry and government
- Everything we will do also works in S-PLUS.

18

Supervised Learning

Supervised Learning:

- Learn a rule for predicting the value of an outcome variable based on the value of some set of predictor variables.
 - e.g. predicting house sale price using square footage, number of BRs, etc.
- Have a set of “training” samples for which the predictors and outcome values are known.
 - <fill in>
- Example: classification
 - Outcome variable is a class (category).
 - E.g., predicting whether or not a Skype account is fraudulent based on the number of contact requests issued, number of contact requests accepted, and number of contact requests received.
 - Learn a good “classification rule” from the training data
 - Apply it to new data to predict the class.

20

Supervised Learning: Evaluating Accuracy

- How would you evaluate the predictive accuracy of a classification rule on new (test) data?

21

Unsupervised Learning

Unsupervised Learning:

- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
 - The training data is the purchasing data of a set of customers
 - Learn a decision rule for categorizing new customers
- How might one evaluate accuracy for this type of decision rule?

22

Supervised or Unsupervised?

Say we want to estimate the amount of \$ an individual spends annually on fabric softener, based on their demographics (age, income, location of residence, etc.). The training data consist of measurements of the demographic variables and self-reported amount of \$ spent annually on fabric softener, for 6,000 individuals.

Choose one:

- A. Supervised learning
- B. Unsupervised learning

23

Supervised or Unsupervised?

We want to reduce the dimension of our data, while preserving as much information as possible. E.g., want to create a single numeric summary of financial status, based on the variables: income, savings, potential future earnings, and credit rating (reduces dimension from 4 to 1). The training data consist of these 4 financial measurements for 200,000 customers.

Choose one:

- A. Supervised learning
- B. Unsupervised learning

24

Supervised or Unsupervised?

We want to group power generation facilities into groups of similar facilities, using a set of operational variables:

- power output per year
- generation capacity
- % time fully utilized
- type of fuel.

The training data consist of these operational measurements, for 1,700 facilities.

Choose one:

- A. Supervised learning
- B. Unsupervised learning

25

Model-based vs. Heuristic Approaches

■ Heuristic approaches to learning

- employ decision rules that have been observed to often work well in practice
- Are not based on an explicit model and may not have other formal justification

■ Model-based approaches

- use an explicit model of the system
- use the data to [learn the parameters](#) of that model
- [predict in an optimal manner](#) based on the model

■ What would be the advantages of each?

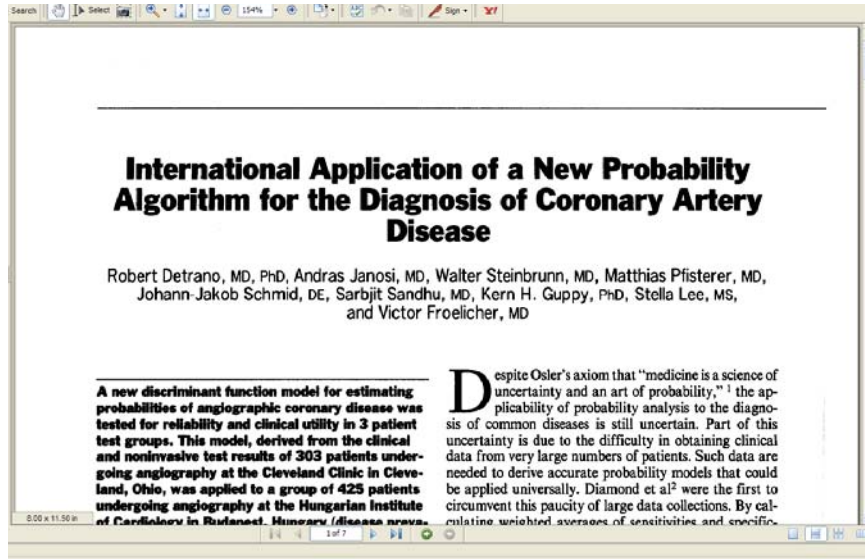
26

Heart Disease Detection

- Can patients be effectively screened for the presence of heart disease (CAD) without the use of angiography?
- Angiography is an invasive and expensive procedure where a tube is inserted into the artery of concern.
- Rather than using angiography on all patients to detect CAD, it is better to use it on high-risk patients.

28

Heart Disease Detection



29

Heart Disease Detection

- The authors use data from the Cleveland Clinic. The data has non-invasive clinical test results as well as angiography results (CAD / no CAD) for 303 patients.
- They learn a classification rule for predicting CAD based on the non-invasive test results.
- This classification rule uses a model-based approach (logistic regression).
- They check the predictive accuracy of their classification rule on data from patients in Hungary (74%) and California (77%).

30

Heart Disease Detection

- What are the training data here?
- What are the test data?
- Is the test data set a good choice or not?

31

Heart Disease Detection

What type of learning is this?

- A. Supervised
- B. Unsupervised

32

Heart Disease Detection

The Cleveland data:

```
70.0 1.0 4.0 130.0 322.0 0.0 2.0 109.0 0.0 2.4 2.0 3.0 3.0 2
67.0 0.0 3.0 115.0 564.0 0.0 2.0 160.0 0.0 1.6 2.0 0.0 7.0 1
57.0 1.0 2.0 124.0 261.0 0.0 0.0 141.0 0.0 0.3 1.0 0.0 7.0 2
64.0 1.0 4.0 128.0 263.0 0.0 0.0 105.0 1.0 0.2 2.0 1.0 7.0 1
74.0 0.0 2.0 120.0 269.0 0.0 2.0 121.0 1.0 0.2 1.0 1.0 3.0 1
65.0 1.0 4.0 120.0 177.0 0.0 0.0 140.0 0.0 0.4 1.0 0.0 7.0 1
56.0 1.0 3.0 130.0 256.0 1.0 2.0 142.0 1.0 0.6 2.0 1.0 6.0 2
59.0 1.0 4.0 110.0 239.0 0.0 2.0 142.0 1.0 1.2 2.0 1.0 7.0 2
60.0 1.0 4.0 140.0 293.0 0.0 2.0 170.0 0.0 1.2 2.0 2.0 7.0 2
63.0 0.0 4.0 150.0 407.0 0.0 2.0 154.0 0.0 4.0 2.0 3.0 7.0 2
59.0 1.0 4.0 135.0 234.0 0.0 0.0 161.0 0.0 0.5 2.0 0.0 7.0 1
53.0 1.0 4.0 142.0 226.0 0.0 2.0 111.0 1.0 0.0 1.0 0.0 7.0 1
44.0 1.0 3.0 140.0 235.0 0.0 2.0 180.0 0.0 0.0 1.0 0.0 3.0 1
61.0 1.0 1.0 134.0 234.0 0.0 0.0 145.0 0.0 2.6 2.0 2.0 3.0 2
57.0 0.0 4.0 128.0 303.0 0.0 2.0 159.0 0.0 0.0 1.0 1.0 3.0 1
71.0 0.0 4.0 112.0 149.0 0.0 0.0 125.0 0.0 1.6 2.0 0.0 3.0 1
46.0 1.0 4.0 140.0 311.0 0.0 0.0 120.0 1.0 1.8 2.0 2.0 7.0 2
53.0 1.0 4.0 140.0 203.0 1.0 2.0 155.0 1.0 3.1 3.0 0.0 7.0 2
64.0 1.0 1.0 110.0 211.0 0.0 2.0 144.0 1.0 1.8 2.0 0.0 3.0 1
40.0 1.0 1.0 140.0 199.0 0.0 0.0 178.0 1.0 1.4 1.0 0.0 7.0 1
```

33

Heart Disease Detection

The meta-data:

```
-----
-- 1. age
-- 2. sex
-- 3. chest pain type (4 values) |
-- 4. resting blood pressure
-- 5. serum cholestoral in mg/dl
-- 6. fasting blood sugar > 120 mg/dl
-- 7. resting electrocardiographic results (values 0,1,2)
-- 8. maximum heart rate achieved
-- 9. exercise induced angina
-- 10. oldpeak = ST depression induced by exercise relative to re:
-- 11. the slope of the peak exercise ST segment
-- 12. number of major vessels (0-3) colored by flourosopy
-- 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
```

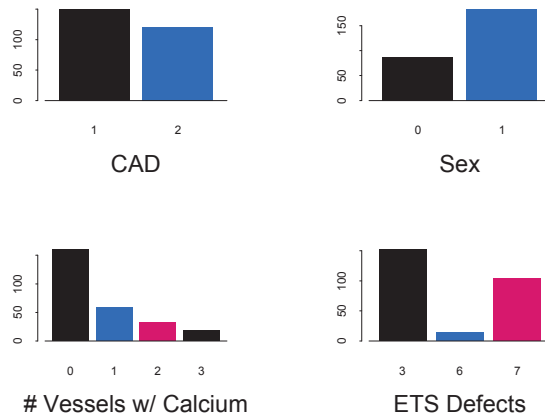
Attributes types

```
-----
Real: 1,4,5,8,10,12
Ordered:11,
Binary: 2,6,9
Nominal: 7,3,13
```

34

Heart Disease Detection

Frequencies of Variables:



35

Heart Disease Detection

- The goal is to learn a good **classification rule** to predict the **presence / absence of CAD** from the 13 predictors in the data set:

- age
- sex
- chest pain type
- blood pressure
- Number of vessels showing calcium on fluoroscopy
- exercise thallium scintigraphic defects (fixed, reversible, none)
- electrocardiogram results
- exercise-induced angina (presence / absence)
- etc.

36

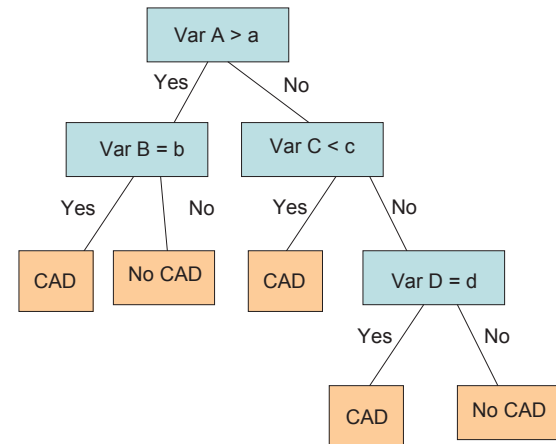
Heart Disease Detection

- We'll come up with our own classification rule for this problem

37

Heart Disease Detection

We'll use a [classification tree](#):



38