First mid-term: Oct 21 9:15am - 11:45am

In-person, open-book, open-note, your own program. More details if you need to use a computer/cacluator WITHOUT internet connection

Review for the mid-term: Oct 19

No class: Oct 28

Location: SB B141

Discuss latest start time for the mid-term on Oct 19.

Topics covered that may be in the 1st mid-term:

- Chapter 1: Concept of information usefulness; e.g., ESP experiment in chapter 1

- Chapter 2: Concept of data warehouse vs DB

- Chapter 3: Mathematical pattern such as given a sequence, derive the missing value; e.g., 2 4 ? 8 10 usin LaRange mulitplier (page 5?)

- Chapter 3: Decision tree such as the 12-item problem or the tree leaf data set (you can use your own program with extra credits too --- only if it works)

- Chapter 8: Statistically significant associaton pattern (you can use your own program with extra credits too --- only if it works)

Remark:

1. If you use your own program, there will be an one-on-one interview to finalize the grade for those questions.

2. General rule: Your final answer typically counts for 30% or less, your steps demonstrating your understanding counts at least 60%

Topics covered that may be in the 2nd mid-term (Nov 18? in the handout):

- Chapter 4: Self-study Information theory (Q&A on OCt 26)

- Chapter 10: Bayesian network

Outside 2nd mid-term (toward end of semester one-on-one interview)

- Chapter 9: Self-study Q&A on chapter 9 about model discovery (mandatory for grad student and optional for undergrad for extra credits)

Type Overall Type Arrangement Margin Margin sub-category Count

Dogwood Simple Opposite Entire X           2

Maple Simple Opposite Lobed X   2

Ash Compound Opposite X X 40

Hickey Compound X Toothed X   2

Locust Compound Alternate Entire X   2

Cherry Simple Alternate Toothed X   2

White Oak Simple Alternate Lobed Rounded 100

Red Oak Simple Alternate Lobed Pointed 100 ------

                   2 3 4 3   250

{0..7} {0,1} {0,1,2} {0,1,2,3} {0,1,2}

Variable

A B C D E

Type OverallType Arrangement Margin Margin sub-category Count

0 0 0 0 0         2

1 0 0 1 0 2

2 1 0 2 0 40

3 1 1 3 0 2

4 1 2 0 0 2

5 0 2 3 0 2

6 0 2 1 1 100

7 0 2 1 2 100

------

2 3 4 3 250

{0..7} {0: Simple,1}   {0:opposite, {Entire:0

    1:X,2}     1:lobed,2:X,3} {0:X,1:Rounded,2}

Is (Overall type:Simple, Arrangement:Opposite)=(B:0, C:0) a statistically significant association pattern given threshold = 1%?

1. Support measure

Pr(B:0, C:0)= \sum\_A,D,E, Type Pr(A, B:0, C:0, D, E, Type)= 4/250=0.016 > 0.01? yes

Pr(B:0)= 206/250

Pr(C:0)=  44/250

2. Level of dependency in terms of mutual information measure

L.H.S. = MI(B:0, C:0) = Log\_2 Pr(B:0 C:0)/Pr(B:0)Pr(C:0)

  = Log\_2 (4/250)/(206/250)(44/250)

  = Log\_2 4\*250/206\*44= Log\_2 0.11 < 0

R.H.S. = CHi-square/2N

N= 250

Chi-square = (Oi - ei)^2/ei

Oi = observed count of (B:0 C:0) = 4

ei = expected count of (B:0 C:0) under the assumption of independence

   = N\*Pr(B:0 \*)Pr(C:0)

   = 250(206/250)(44/250)

   = 36.256

chi-square = (4 - 36.256)^2/(36.256) > 0

chi-square/2N > 0

Therefore, (B:0 C:0) is NOT statistically significant because LHS < RHS even it passes the threshold test.