Midterm 1 - Review

AW

Midterm Overview

1. **Description**
2. PCA
3. Choose Binary Split
4. Multi-outcome Split
5. Graduate

Midterm Description

- Total comes to approximately 90 pts. However, max you can earn is capped at 70.
- On one sheet exam for graduate and for undergraduate – however questions are marked as follows:
  - UG (e.g., 'problem 1UG') – for both graduate and undergraduate
  - G – (e.g. 'problem 2G') for graduate

  Please pay attention and avoid confusion.
- 3 questions UG + 1 question G
- Computationally intensive - you should use Wolfram alpha or calculator.
- You MUST give ALL intermediate results whenever asked for. No intermediate results no credit

Midterm Description - continued

- Credit for each question clearly marked. Partial credit possible, also marked.
  - Credit is based on undergraduate credit
  - Undergrad credit for a UG question is given in brackets e.g. [30] means that a question gives undergraduate student 30 pts.
  - Graduate credit is given in curly brackets. For UG questions it is given as a multiplier to apply to undergrad credit, e.g. [30]{4/5} means that grad students get for this question $30 \times \frac{4}{5} = 24$ points. Same multiplier applies to all partial points
  - Grad credit for grad only questions (G) credit is given in curly brackets {20} means grad students get 20 pts for this question
  - Undergrads can try grad question for extra credit **but only after answering all undergrad questions.**

Midterm Description - continued

1. [30]{4/5} - PCA. The problem is very similar to exercises 1-6 in HW 1A. Show intermediate results:
   - mean vector, mean-centered matrix, covariance matrix, characteristic polynomial, eigenvalues, transformation matrix
2. [35pts]{4/5} - Given a data table. Compute gain using a pre-specified measure for a multi-valued nominal attribute, when we are looking to evaluate only binary splits. Problem is similar to ex HW 2A: ch 4 #2. Intermediate results for alternatives are expected.

Midterm Description - continued

3. [24]{5/6} - Given a simple data table; using a pre-specified measure construct the first level of DT. Intermediate results for each attribute expected. Problem is similar to exercise in HW 2A: Ch 4 #5,6
4. {16 }Graduate only. Explain what happens when certain condition of correlation covariance holds. The problem is similar to Hw1G problem 1 Ch. 2 #22

Midterm Overview

1. Description
2. PCA
3. Choose Binary Split
4. Multi-outcome Split
5. Graduate

PCA

- Data Frame:

Attr /rec	A1	A2	A3
1	1	-1	4
2	2	1	3
3	1	3	-1
4	4	-1	3

- Compute multivariate mean: column mean

{{1,-1,4},{2,1,3},{1,3,-1},{4,-1,3}} = {2, 1/2, 9/4}

- Center the matrix

{{1,-1,4},{2,1,3},{1,3,-1},{4,-1,3}}-{{2, 1/2, 9/4}, {2, 1/2, 9/4},{2, 1/2, 9/4},{2, 1/2, 9/4}}=

{{-1, -3/2, 7/4}, {0, 1/2, 3/4}, {-1, 5/2, -13/4}, {2, -3/2, 3/4}}

- Compute covariance matrix
  1/(4-1)({{-1, -3/2, 7/4}, {0, 1/2, 3/4}, {-1, 5/2, -13/4}, {2, -3/2, 3/4}}^T*{{-1, -3/2, 7/4}, {0, 1/2, 3/4}, {-1, 5/2, -13/4}, {2, -3/2, 3/4}}
  ={{2,-4/3,1},{-4/3,11/3,-23/6}{1,-23/6,59/12}}
- Characteristic polynomial
  {{2,-4/3,1},{-4/3,11/3,-23/6},{1,-23/6,59/12}}
- Solve 121/27 - (319 λ)/18 + (127 λ^2)/12 - λ^3=0

PCA (continued)

- Eigenvalues of characteristic polynomial are λ_1=8.5783; λ_2=1.6972; λ_3=0.30781;
- Eigenvectors of centered matrix

eigenvectors {{2,-4/3,1},{-4/3,11/3,-23/6},{1,-23/6,59/12}}

= for λ_1 (0.328266, -0.869573, 1.); for λ_2 (-2.65908, 0.14618, 1.); for λ_3 (0.448599, 1.31934, 1.);

- Are eigenvectors orthogonal?
  - Yes, because it is orthogonal diagonalization
- Are eigenvectors normal?
  - obviously not (i.e. $||v_i|| \neq 1$); need to be normalized

PCA (continued)

normalize (0.328266, -0.869573, 1)=(0.240443,-0.636932,0.732465)

normalize (-2.65908, 0.14618, 1)=(-0.934763,0.0513876,0.351536)

normalize (0.448599, 1.31934, 1)=(0.261544,0.769206,0.583024)

- Rotation matrix:

{{0.240443,-0.636932,0.732465},{-0.934763,0.0513876,0.351536},{0.261544,0.769206,0.583024}}^T

- Rotated raw data:

{{1,-1,4},{2,1,3},{1,3,-1},{4,-1,3}}*{{0.240443,-0.636932,0.732465},{-0.934763,0.0513876,0.351536},{0.261544,0.769206,0.583024}}^T={{3.81,0.41,1.82},{2.04,-0.76,3.04},{-2.40,-1.13,-1.99},{3.79,-2.73,2.02}}

Midterm Overview

1. Description
2. PCA
3. Choose Binary Split
4. Multi-outcome Split
5. Graduate

Choose Binary Split with Gini

Customer ID	Shirt Size	Class
1	Small	C0
2	Medium	C0
3	Medium	C0
4	Large	C0
5	Extra Large	C0
6	Extra Large	C0
7	Small	C0
8	Small	C0
9	Medium	C0
10	Large	C0
11	Large	C1
12	Extra Large	C1
13	Medium	C1
14	Extra Large	C1
15	Small	C1
16	Small	C1
17	Medium	C1
18	Medium	C1
19	Medium	C1
20	Large	C1

It is an ordinal attribute sm<med<large<xlarge, so if class depends on this attribute then splits could be

1. S vs. M+L+XL,
2. S+M vs. L+XL
3. S+M+L vs. XL

For #1

- Left child: 3-C0,2C1
- Right child: 7-C0, 8C1

For #2

- Left child: 6C0, 6C1
- Right child: 4C0,4C1

For #3

- Left child: 8C0, 8C1
- Right child: 2C0,2C1

Gini Index

- Parent GINI index:

$Gini(P)=1-(10/20)^2-(10/20)^2=1-0.25-0.25=0.5$

- Gini index for #1

$Gini(L^{\#1})=1-(3/5)^2-(2/5)^2=1-0.36-0.16=0.48$

$Gini(R^{\#1})=1-(7/15)^2-(8/15)^2=1-0.218-0.284=0.498$

Combined $Gini(\#1)=5/20*0.48+15/20*0.498=0.493$

- Gini index for #2

$Gini(L^{\#2})=1-(6/12)^2-(6/12)^2=1-0.25-0.25=0.5$

$Gini(R^{\#2})=1-(4/8)^2-(4/8)^2=1-0.25-0.25=0.5$

Combined $Gini(\#2)=(12/20)*0.5+(8/20)*0.5=0.5$

- Gini for #3 actually obvious without computation

$Gini(L^{\#3})=0.5$; $Gini(R^{\#3})=0.5$; Combined $Gini(\#3)=0.5$

- Winner #1, Gini gain =Gini(parent)-Gini(#1)=0.5-0.493=0.07

Midterm Overview

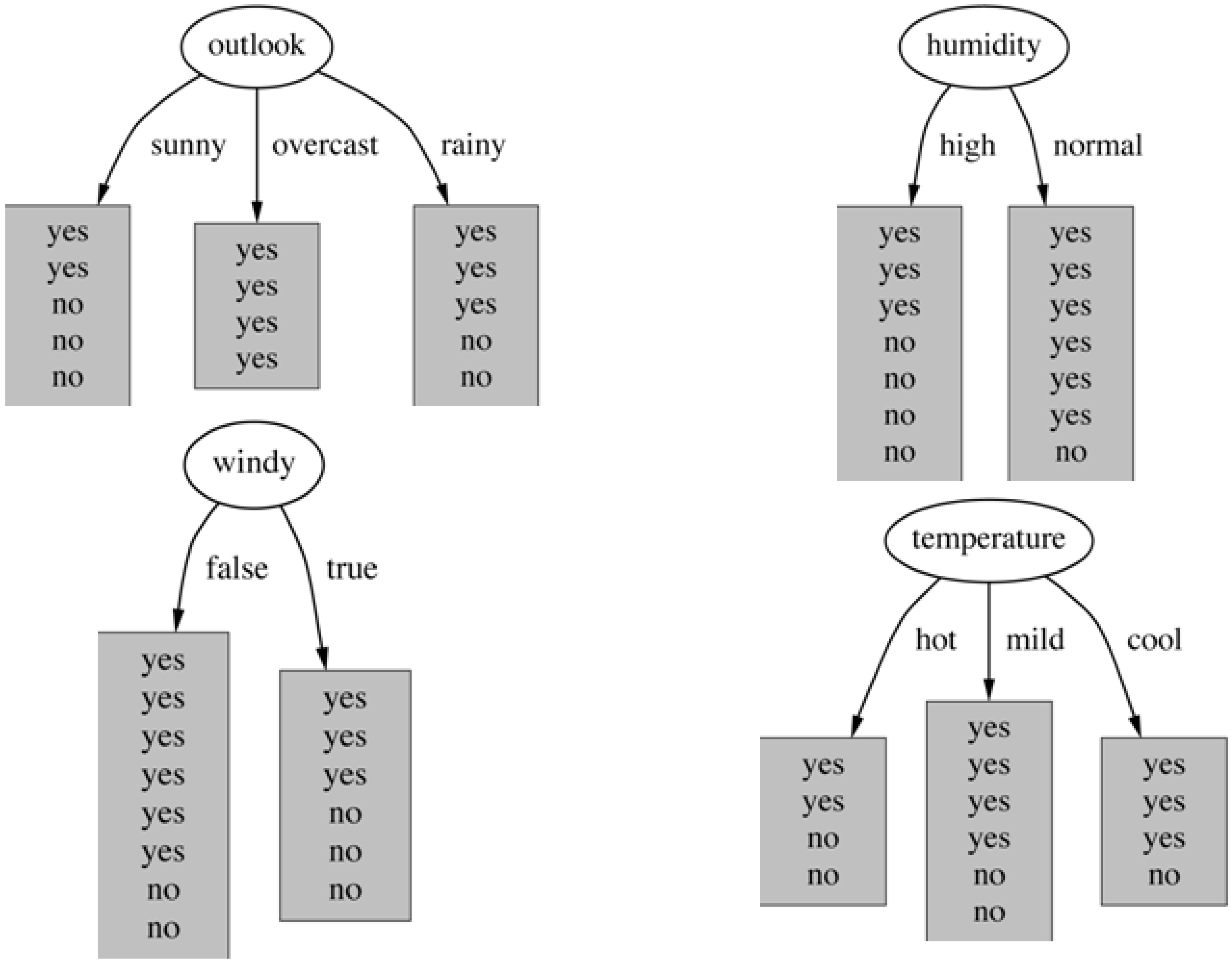
1. Description
2. PCA
3. Choose Binary Split
4. **Multi-outcome Split**
5. Graduate

Information Gains for Multi-brunch Outcome Split

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Parent Entropy:

$$H(p) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$


outlook
sunny
overcast
rainy
yes yes no no no
yes yes yes yes
yes yes yes no no
humidity
high
normal
yes yes yes no no no no
yes yes yes yes yes yes no
windy
false
true
yes yes yes yes yes yes no no
yes yes yes no no no
temperature
hot
mild
cool
yes yes no no
yes yes yes yes no no
yes yes yes no

Information Gain for "Outlook"

$$H(2/5,3/5) = -2/5\log(2/5) - 3/5\log(3/5) = 0.971\,\text{bits}$$

- Outlook = Sunny:

$$H\left(\frac{2}{5},\frac{3}{5}\right) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.97$$

- Outlook = Overcast:

$$H(1,0\ ) = 0$$

- Outlook = Rainy:

$$H\left(\frac{2}{5},\frac{3}{5}\right) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.97$$

- Expected information for attribute:

$$H(outlook) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot .0971 = 0.693$$

- Information Gain

$$\text{Gain(Outlook)} = \text{H(parent) - H(Outlook)} = 0.940 - 0.693 = 0.247\text{bits}$$

Information Gains for Multi-brunch Outcome Splitt

- Information gain for attributes from weather data:

$$Gain(outlook) = 0.247$$
$$Gain(Windy) = 0.048$$
$$Gain(Temperature) = 0.029$$
$$Gain(Humidity) = 0.152$$

Outlook wins!

Midterm Overview

1. Description
2. PCA
3. Choose Binary Split
4. Multi-outcome Split
5. Graduate

Use of Correlation/ Variance/Covariance

- Discuss how you might map correlation values from the interval $[-1, 1]$ to the interval $[0, 1]$. Note that the type of transformation that you use might depend on the application that you have in mind. Thus, consider two applications: clustering time series and predicting the behavior (magnitude of change) of one time series given another.

Use of Correlation/ Variance/Covariance

- For time series clustering:
  - Interesting is only high positive correlation – negative can be disregarded. Then define similarity as
    $$sim(A_i, A_j) = \begin{cases} corr(A_i, A_j) \ if \ corr(A_i, A_j) > 0 \\ 0 \ \ otherwise \end{cases}$$
- For magnitude change:
  - Sign is unimportant only magnitude is of interest so similarity is $sim(A_i, A_j) = |corr(A_i, A_j)|$