CSC662 Data Mining, Data Warehouse and Visualization

บทที่ 9 การทำเหมืองข้อมูลแบบจัดจำแนกประเภทแบบเบย์

เตรียมโดย ผศ. ดร. กรุง สินอภิรมย์สราญ

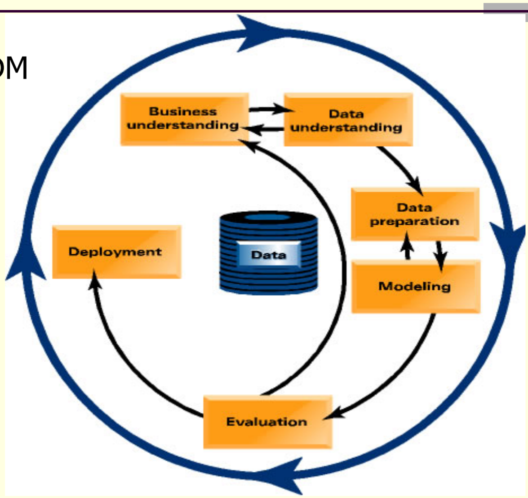
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

โครงร่าง

- การจำแนกประเภทโดยใช้ความน่าจะเป็น
- การจำแนกประเภทโดยใช้กฎของเบย์
- ทฤษฎีของเบย์
- การแก้ปัญหาโดยใช้ Laplace estimator
- สมมติฐานการเป็นอิสระต่อกัน
- การจำแนกประเภทโดยใช้เครือข่ายของเบย์
- การประเมินความถูกต้องของผลลัพธ์

กระบวนการค้นหาความรู้

CRISP-DM



สถิติกับการจัดจำแนกประเภท

- ปัญหาการจัดจำแนกประเภทเป็นปัญหาที่นักสถิติสนใจมานานในรูปที่ตัวแปรตาม (Response variable) มีค่าที่ไม่ต่อเนื่อง
- แนวคิดในการจัดการกับตัวแปรตามหรือคลาสเป้าหมายที่มีค่าไม่ต่อเนื่องทำได้โดยใช้การแจกแจง กล่าวคือสำหรับตัวอย่างที่ไม่ทราบค่าคลาสเป้าหมายจะให้ผลลัพธ์เป็นความน่าจะเป็นที่ตัวอย่างดังกล่าวจะอยู่ในกลุ่มของคลาสเป้าหมายกลุ่มใด กลุ่มหนึ่ง ซึ่งการประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์สามารถใช้ความถี่สัมพัทธ์ได้ในกรณีที่ข้อมูลมีปริมาณมากพอ
- วิธีการนี้มักถูกใช้เป็นมาตรฐานสำหรับเปรียบเทียบกับวิธีการอื่น

การจัดจำแนกประเภทโดยใช้กฎของเบย์

- เป็นการเรียนรู้โดยใช้ความน่าจะเป็น: คำนวณการแจกแจงค่าความน่าจะเป็นตามสมมติฐานที่ตั้งให้กับข้อมูล
- เป็นการเรียนรู้เพิ่มได้: ตัวอย่างใหม่ที่ได้มาถูกนำมาปรับเปลี่ยนการแจกแจงซึ่งมีผลต่อการเพิ่ม/ลดความน่าจะเป็น ทำให้มีการเรียนรู้ที่เปลี่ยนไป วิธีการนี้ตัวแบบจะถูกปรับเปลี่ยนไปตามตัวอย่างใหม่ที่ได้โดยผนวกกับความรู้เดิมที่มี
- การทำนายค่าคลาสเป้าหมายของตัวอย่างใหม่ใช้ความน่าจะเป็นที่มากที่สุดของทุกสมมติฐาน
- เป็นตัวกลางในการเปรียบเทียบกับขั้นตอนวิธีอื่น: วิธีการของเบย์ (Bayes) ถูกใช้เป็นมาตรฐานในการเปรียบเทียบกับวิธีการอื่น

ทฤษฎีของเบย์ (Bayesian theorem)

- ให้ D แทนข้อมูลที่นำมาใช้ในการคำนวณการแจกแจงความน่าจะเป็น *posteriori probability* ของสมมติฐาน h คือ $\Pr(h|D)$ ตามทฤษฎีของเบย์

$$\Pr(h|D) = \frac{\Pr(D|h)\Pr(h)}{\Pr(D)}$$

- ใช้หลักการ MAP (maximum posteriori) hypothesis

$$h_{MAP} \equiv \underset{h\in H}{\arg\max} \Pr(h|D) = \underset{h\in H}{\arg\max} \Pr(D|h)\Pr(h).$$

- สังเกตุว่าการทำนายได้จากการคำนวณความน่าจะเป็นแบบมีเงื่อนไขในทุกสมมติฐาน ซึ่งขึ้นกับจำนวนค่าที่เป็นไปได้ทั้งหมดของตัวแปรตาม

ตัวอย่างการคำนวณค่าความน่าจะเป็นจาก weather

- จากข้อมูลใน weather.arff คลาสเป้าหมายคือ play ซึ่งมีค่าเป็น yes หรือ no การแจกแจงความน่าจะเป็นสำหรับค่าคลาสเป้าหมายตามตัวแปรอิสระคือ

Outlook	yes	no
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Temperature	yes	no
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	yes	no
high	3/9	4/5
normal	6/9	1/5
Windy	yes	no
true	3/9	3/5
false	6/9	2/5

Bayesian classification

- Bayesian classification คือปัญหาการจำแนกข้อมูลตาม a-posteriori probabilities:

  $\Pr(\text{class}=c_i|\ X)$ = ความน่าจะเป็นที่ $X = (x_1,\ldots,x_k)$ จะเป็น c_i

  เช่น

  Pr(class = N | outlook = sunny, windy = true,…)
- วิธีคิด: เรากล่าวว่าตัวอย่าง X เป็น c_i ถ้าค่า $\Pr(\text{class} = c_i|X)$ มีค่าสูงสุดเมื่อเทียบกับค่า c_j อื่น ๆ

การประมาณ a-posteriori probability

- ทฤษฎีของ Bayes:

  Pr(class = c|X) = Pr(X|class = c)·Pr(class = c)/Pr(X)

  ข้อสังเกตุว่า Pr(X) ไม่เปลี่ยนสำหรับค่าคลาส c ที่เปลี่ยนไป
- Pr(class = c) ≈ ความถี่สัมพัทธ์ของตัวอย่างในคลาส c

  Pr(class=c|X) สูงสุดก็ต่อเมื่อ Pr(X|class=c)·Pr(class=c) สูงสุด
- ดังนั้น เราต้องการประมาณ Pr(X | class = c) ที่เหมาะสมที่สุด เพื่อตัดสินการทำนายค่าคลาสของตัวอย่าง X

Naïve Bayesian Classification

- วิธีการของ Naïve Bayesian คือการใช้วิธีการของเบย์พร้อมสมมติฐานของการเป็นอิสระต่อกันของตัวแปรอิสระทุกตัว

  Pr(x_1,…,x_k|class = c) = Pr(x_1|class = c)·…·Pr(x_k|class = c)
- ถ้าลักษณะประจำ i เป็น categorical: Pr(x_i|class = c) ประมาณด้วยความถี่สัมพัทธ์ของตัวอย่างที่มีค่า x_i ในคลาส c
- ถ้าลักษณะประจำ i เป็น continuous: Pr(x_i|class = c) ประมาณด้วยฟังก์ชันความหนาแน่น Gaussian
- การคำนวณทั้งสองลักษณะจากข้อมูลปริมาณมากไม่มีความซับซ้อนและทำงานได้เร็ว

ตัวอย่าง Pr(x_i|play=yes), Pr(x_i|play=no)

Outlook	Temperature	Humidity	Windy	Class
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

Pr(yes) = 9/14	
Pr(no) = 5/14	

outlook	
Pr(sunny|yes) = 2/9	Pr(sunny|no) = 3/5
Pr(overcast|yes)=4/9	Pr(overcast|no) = 0
Pr(rain|yes) = 3/9	Pr(rain|no) = 2/5
Temperature	
Pr(hot|yes) = 2/9	Pr(hot|no) = 2/5
Pr(mild|yes) = 4/9	Pr(mild|no) = 2/5
Pr(cool|yes) = 3/9	Pr(cool|no) = 1/5
Humidity	
Pr(high|yes) = 3/9	Pr(high|no) = 4/5
Pr(normal|yes) = 6/9	Pr(normal|no) = 2/5
Windy	
Pr(true|yes) = 3/9	Pr(true|no) = 3/5
Pr(false|yes) = 6/9	Pr(false|no) = 2/5

14 13

วิธีการจำแนกตัวอย่างโดยใช้ Naïve Bayes

- ตัวอย่างใหม่ที่ไม่ทราบมาก่อน X = (outlook = rain, temperature = hot, humidity = high, windy = false)

Pr(X|play=yes)·Pr(play=yes) = Pr(outlook=rain|play=yes) ·Pr(temperature=hot|play=yes)· Pr(humidity=high|play=yes) ·Pr(windy=false|play=yes)· Pr(play = yes)= 3/9·2/9·3/9·6/9·9/14 = 0.010582

Pr(X|play=no)·Pr(play=no) = Pr(outlook=rain|play=no) ·Pr(temperature=hot|play=no)· Pr(humidity=high|play=no) ·Pr(windy=false|play=no)·Pr(play = no) = 2/5·2/5·4/5·2/5·5/14 =0.018286

- ตัวอย่าง X ถูกจัดอยู่ใน play = no (ไม่เล่น) เพราะมีค่าความน่าจะเป็นใน play = no มากที่สุด

ปัญหาของค่าความน่าจะเป็นที่เป็นศูนย์

- วิธีการประมาณค่าความน่าจะเป็นโดยใช้การคำนวณจากตารางความถี่ เราพบว่าถ้าเราสุ่มตัวอย่างมาไม่ดี เราอาจไม่พบตัวอย่างที่สอดคล้องกับเงื่อนไขที่สนใจ เช่นในตาราง Pr(outlook=overcast | no) = 0 ค่าดังกล่าวทำให้การคำนวณค่าความน่าจะเป็นแบบเบย์เป็นศูนย์เสมอ ซึ่งอาจไม่ใช่สิ่งที่เราต้องการ
- กล่าวคือค่าความน่าจะเป็นที่เป็นศูนย์ทำให้วิธีการดังกล่าว ทำให้ค่าความน่าจะเป็นรวมเป็นศูนย์เสมอ
- กลยุทธในการแก้ปัญหานี้คือ การใช้ตัวประมาณค่าใหม่ที่เรียกว่า Laplace estimator กล่าวคือการบวกหนึ่งหน่วยเข้ากับตารางความถี่ก่อนเริ่มคำนวณ

11

ตัวอย่าง Pr(x_i|play=yes), Pr(x_i|play=no)

Outlook	Temperature	Humidity	Windy	Class
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

Pr(yes) = 10/16	
Pr(no) = 6/16	

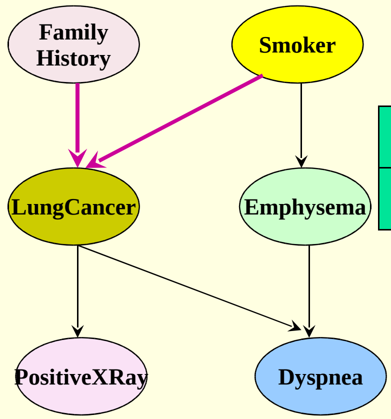
outlook	
Pr(sunny|yes) = 3/12	Pr(sunny|no) = 4/8
Pr(overcast|yes)=5/12	Pr(overcast|no) = 1/8
Pr(rain|yes) = 4/12	Pr(rain|no) = 3/8
Temperature	
Pr(hot|yes) = 3/12	Pr(hot|no) = 3/8
Pr(mild|yes) = 5/12	Pr(mild|no) = 3/8
Pr(cool|yes) = 4/12	Pr(cool|no) = 2/8
Humidity	
Pr(high|yes) = 4/11	Pr(high|no) = 5/7
Pr(normal|yes) = 7/11	Pr(normal|no) = 2/7
Windy	
Pr(true|yes) = 4/11	Pr(true|no) = 4/7
Pr(false|yes) = 7/11	Pr(false|no) = 3/7

11

การเป็นอิสระต่อกันของลักษณะประจำ

- สมมติฐานการเป็นอิสระต่อกันของลักษณะประจำหรือตัวแปรอิสระ ทำให้การคำนวณเป็นไปได้ง่ายและสะดวก
- ผลลัพธ์ที่ได้จะเชื่อถือได้ถ้าสมมติฐานของความเป็นอิสระเป็นจริง
- ในทางปฏิบัติสมมติฐานดังกล่าวยอมรับได้ยาก กล่าวคือลักษณะประจำมักไม่เป็นอิสระต่อกัน
- หลักการแก้คือ
  - ใช้เครือข่าย Bayesian, รวมการให้เหตุผลแบบ Bayesian กับการโยงลักษณะประจำที่มีความน่าจะเป็นที่ไม่เป็นอิสระต่อกัน
  - ใช้ต้นไม้การตัดสินใจ, เลือกลักษณะประจำที่พิจารณาตามความสำคัญ

Bayesian Networks



	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

ตารางความน่าจะเป็นแบบมีเงื่อนไข
ของตัวแปร **LungCancer**

Bayesian Networks

หลักการ Bayesian Networks

- Bayesian network ยอมให้สับเซตของตัวแปรเป็นอิสระต่อกันอย่างมีเงื่อนไข
- ใช้รูปกราฟแสดงความเกี่ยวพันระหว่างตัวแปร โดยที่จุดยอดคือตัวแปร และเส้นเชื่อมระหว่างจุดยอดคือการไม่เป็นอิสระต่อกันของตัวแปรเหล่านั้น
- การกำหนดเครือข่าย Bayesian จะมีผลต่อการคำนวณกล่าวคือ
  - ถ้าผู้ใช้กำหนดเครือข่ายทั้งหมด โปรแกรมเพียงคำนวณค่าเติมในเครือข่าย
  - ถ้าผู้ใช้กำหนดเครือข่ายบางส่วน โปรแกรมต้องมีการทดสอบการเป็นอิสระ ต่อกันของแต่ละส่วนในเครือข่าย
  - ถ้าผู้ใช้ไม่กำหนดโครงสร้างเครือข่ายเลย โปรแกรมต้องทดลองทุกการจัดหมู่ที่เป็นไปได้ซึ่งใช้เวลาคำนวณนาน

ความน่าเชื่อถือของผลลัพธ์การจัดจำแนกประเภท

- ปรกติผู้ใช้ต้องการตัวแบบที่ทำนายตัวอย่างที่ไม่เคยพบได้ถูกต้องมากที่สุด
- กล่าวคือผลลัพธ์ที่ได้จากตัวแบบจะมีสองคำตอบ
  - ทำนายได้ถูกต้องตรงกับค่าคลาสเป้าหมาย
  - หรือทำนายได้ไม่ถูกต้องกับค่าคลาสเป้าหมายที่สนใจ
- เราเรียกตัววัดดังกล่าวว่า 0-1 loss function
- แต่ในกรณีที่ตัวแบบให้ค่าเป็นความน่าจะเป็น เรานิยามใช้ quadratic loss function แทนการประเมิน

Quadratic loss function

- สมมติว่าตัวอย่างหนึ่งตัวให้ค่าคลาสเป้าหมายได้ k แบบ
- จากข้อมูล ตัวแบบอาจประมาณค่าความน่าจะเป็นของคลาสเป้าหมายเป็น (p_1, p_2, ..., p_k)
- แต่ค่าที่แท้จริงของแต่ละตัวอย่างจะเป็น 1 หรือ 0 เขียนแทนด้วย (a_1, a_2, ..., a_k)
- ค่าของ quadratic loss function คำนวณได้จาก

$$\sum_j (p_j - a_j)^2$$

- ค่าน้อยแสดงถึงความถูกต้องที่สูงกว่า

ค่าประเมินความถูกต้อง TP, TN, FP, FN

- ค่าสี่ตัววัดสำหรับคลาสเป้าหมายสองค่า yes กับ no คือ
  - TP (True positive) คือจำนวนตัวอย่างที่ตัวแบบทำนายว่า yes และค่าคลาสเป้าหมายเป็น yes
  - TN (True negative) คือจำนวนตัวอย่างที่ตัวแบบทำนายว่า no และค่าคลาสเป้าหมายเป็น no
  - FP (False positive) คือจำนวนตัวอย่างที่ตัวแบบทำนายว่า yes และค่าคลาสเป้าหมายเป็น no
  - FN (False negative) คือจำนวนตัวอย่างที่ตัวแบบทำนายว่า no และค่าคลาสเป้าหมายเป็น yes

True positive rate, False negative rate

- ค่าประเมินความถูกต้องอีกสองตัวสำหรับคลาสเป้าหมาย yes กับ no คือ
  - True positive rate = TP/(TP + FN) คือสัดส่วนของจำนวนตัวอย่างที่ตัวแบบทำนายว่า yes และค่าคลาสเป้าหมายเป็น yes เทียบกับค่าคลาสเป้าหมายที่เป็น yes ทั้งหมด
  - False negative rate = FP/(FP + TN) คือสัดส่วนของจำนวนตัวอย่างที่ตัวแบบทำนายว่า yes และค่าคลาสเป้าหมายเป็น no เทียบกับค่าคลาสเป้าหมายที่เป็น no ทั้งหมด

Success rate and error rate

- ค่าอัตราการวัดความสำเร็จอีกสองตัวสำหรับคลาสเป้าหมาย yes กับ no คือ
  - Overall success rate คือสัดส่วนของจำนวนตัวอย่างที่ตัวแบบทำนายถูกทั้งหมดหารด้วยจำนวนตัวอย่างทั้งหมด
  - $$\frac{TP+TN}{TP+TN+FP+FN}$$
  - Error rate = 1 – Success rate คือสัดส่วนของจำนวนตัวอย่างที่ตัวแบบทำนายผิดทั้งหมด หารด้วยตัวอย่างทั้งหมด
- สำหรับการวัดความถูกต้องเมื่อค่าคลาสเป้าหมายมีมากกว่าสองค่า เราใช้เมทริกซ์สับสน (Confusion matrix) ในการแสดงผล

Confusion matrix

- เมทริกซ์สับสน คือเมทริกซ์ที่แต่ละเซลเป็นจำนวนตัวอย่างที่สอดคล้องกับสิ่งเงื่อนไขคือ TP, FP, TN, FN
- ในซอฟต์แวร์ Weka แนวแถวแสดงถึงค่าคลาสเป้าหมายจริงจากตัวอย่าง ในขณะที่แนวหลักแสดงถึงค่าคลาสเป้าหมายที่ทำนายจากตัวแบบ
- ความน่าเชื่อถือของตัวแบบได้จาก จำนวนที่ปรากฏในแนวทแยง กล่าวคือตัวแบบมีความน่าเชื่อถือสูงถ้าค่าในแนวทแยงมีค่าสูง ในค่านอกแนวทแยงมีค่าต่

```
=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 1 4 | b = no
```


ตัวอย่างเมทริกซ์สับสน

```
Correctly Classified Instances          13               92.8571 %
Incorrectly Classified Instances         1                7.1429 %
Kappa statistic                          0.8372
Mean absolute error                      0.2917
Root mean squared error                  0.3392
Relative absolute error                 62.8233 %
Root relative squared error             70.7422 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  1         0.2        0.9       1         0.947      0.911    yes
  0.8       0          1         0.8       0.889      0.922    no

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 1 4 | b = no
```

- จากเมทริกซ์สับสน เราได้ว่าตัวอย่างที่เป็น a มีความถูกต้องทั้งหมด แต่ตัวอย่างที่เป็น b มีการทำนายผิด 1 ตัวอย่าง
- ดังนั้น Overall success rate คือ 13/14 = 92.8571%

สรุป

- การจำแนกประเภทเป็น การศึกษาการแบ่งกลุ่มข้อมูล พบในสาขาวิชาทั้งทางด้านสถิติ เทคโนโลยีสารสนเทศ และการเรียนรู้ด้วยเครื่อง
- การใช้ความน่าจะเป็นในการจำแนกประเภทจะใช้ทฤษฎีของเบย์ ซึ่งถ้าตัวแปรเป็นอิสระต่อกัน เราใช้วิธีการ naïve bayesian
- แต่ถ้าตัวแปรไม่เป็นอิสระต่อกัน เราใช้ bayesian network
- สำหรับการประเมินความถูกต้องของตัวแบบ เรากำหนดตัววัดหลายชนิดได้แก่ TP, TN, FP, FN, true positive rate, false negative rate เป็นต้น

เอกสารอ้างอิง

- R. O. Duda, P. E. Hart and D. G. Stork. Pattern Classification, second edition. John Wiley, New York, 2001.
- P. Langley, W. Iba and K. Thompson. An analysis of Bayesian classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA. Menlo Park, CA: AAAI Press, pp. 223-228, 1992.
- P. Langley and S. Sage. Induction of selective Bayesian classifiers. Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, pp. 399-406, 1994.
- D. Heckerman, D. Geiger and D. M. Chickering. Learning Bayesian networks:The combination of knowledge and statistical data. Machine Learning 20(3):197-243. 1995.
- J. Shafer, R. Agrawal, and M. Mehta. SPRINT : A scalable parallel classifier for data mining. In Proc. 1996 Int. Conf. Very Large Data Bases, 544-555, Bombay, India, Sept. 1996.
- S. M. Weiss and C. A. Kulikowski. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufman, 1991.