



King Mongkut's University of Technology Thonburi

CSS 341 Introduction to Data Science

Examination 3, Semester 1, Academic year 2565

Start: noon of Friday, December 2, 2565 BE

คำแนะนำและคำสั่ง

ข้อสอบชุดนี้มี 3 หน้า มี 3 ข้อ คะแนนเต็ม 30 คะแนน ให้ทำทุกข้อ โดยที่....

1. เขียน Python 3.8 หรือ 3.9 ในรูปแบบ ipynb **ไฟล์เดียวต่อเนื่อง**ทุกข้อ
2. Data files ทุกไฟล์อยู่ใน folder ที่อยู่ข้างๆ คือระดับเดียวกับ folder ที่บรรจุ ipynb ของท่าน ดังนั้น เวลาอ่าน data file ให้กำหนด path เป็น `'../data/???.'???'` เมื่อ `???.'???'` คือชื่อไฟล์ที่อ่านเข้ามา หากผิดกติกานี้จะถูกหัก 1 แต้มต่อการอ่าน data file 1 ครั้ง
3. เขียน Markdown ให้เหมาะสมชัดเจน นั่นคือ สำหรับแต่ละข้อให้ ควรมีใจทย์ (ย่อมาสั้นๆ) ตามด้วยโค้ด และตามด้วยการวิเคราะห์ผลลัพธ์เพื่อสรุปตอบ เรียงลำดับข้อไป
4. การส่ง ให้ส่งขึ้น LEB2 ด้วย file ipynb เท่านั้น. ไม่ต้องส่ง data file มา เพราะมีอยู่แล้ว กลุ่มที่ส่งสายเกินกำหนดจะถูกหัก 3 นาทีละ 1 คะแนน
5. ไฟล์ ipynb ที่ส่ง ให้ตั้งชื่อไฟล์ในรูปแบบ YY-aa-bb-cc-dd-ee เมื่อ YY บอกชั้นปี 63 หรือ 64 ส่วน aa, bb, ... เป็น student ID สองหลักท้าย (เรียงลำดับมาด้วย) ของสมาชิกทุกคน รวมทั้งการตั้งชื่อกลุ่ม (ถ้ามี) ตอนส่งงานเข้า LEB2 ก็ตั้งชื่อแบบเดียวกัน
6. **ส่วนต้นสุด**ของ ipynb file ที่ส่ง**จะต้องมี** student ID และตามด้วยรายชื่อของสมาชิกทุกคนชัดเจน (เขียนเป็น Markdown ไว้) หากไม่มีจะถูกหัก 3 คะแนน

Problem 1

Problem Statement

การวิเคราะห์ข้อมูลลูกค้าช่วยให้ธุรกิจปรับแต่งสินค้าและบริการให้เหมาะกับลักษณะและความต้องการของลูกค้าแต่ละกลุ่มได้ แทนที่ธุรกิจจะทุ่มเงินไปทำการตลาดผลิตภัณฑ์ใหม่ตัวหนึ่งกับลูกค้าทุกราย บริษัทสามารถวิเคราะห์ลูกค้ากลุ่ม (segment) ที่น่าจะซื้อผลิตภัณฑ์ใหม่ตัวนั้นมากที่สุด และจึงทำการตลาดไปยังลูกค้ากลุ่มนั้น

Attributes (File marketing.csv)

People

1. ID: Customer's unique identifier
2. BirthYear: Customer's birth year
3. Education: Customer's education level
4. Marital_Status: Customer's marital status
5. Income: Customer's yearly household income
6. NKids: Number of children in customer's household
7. EnrollDate: Date of customer's enrollment with the company
8. Recency: Number of days since customer's last purchase
9. Complain: 1 if the customer complained in the last year, 0 otherwise

Products

10. AmtFruits: Amount spent on fruits in last 2 years
11. AmtMeats: Amount spent on meat in last 2 years
12. AmtFish: Amount spent on fish in last 2 years
13. AmtSweets: Amount spent on sweets in last 2 years
14. AmtGold: Amount spent on gold in last 2 years

Promotions

15. NumDiscPurchases: Number of purchases made with a discount
16. AccCampaign: "First" if customer accepted the offer in the first campaign,
"Second" if customer accepted the offer in the first campaign,
"Both" if customer accepted the offers in both campaigns,
0 otherwise
17. Response: "Yes" if customer accepted the offer in the last campaign, "No" otherwise

Channels

18. NumWebPurchases: Number of purchases made through the company's website
19. NumCatalogPurchases: Number of purchases made using a catalogue
20. NumStorePurchases: Number of purchases made directly in stores

Your Tasks

- ✓ ให้แบ่งลูกค้าออกเป็นกลุ่ม (segment) อย่างเหมาะสม แล้ววิเคราะห์ (เปรียบเทียบ) ลักษณะเฉพาะของแต่ละกลุ่มนั้น
- ✓ ใช้ข้อมูลเพียงสองกลุ่มใหม่ที่สุ่มที่แบ่งได้จากข้อมูลที่ย่อยที่แล้ว มาสร้างโมเดลคาดการณ์ (หรือพยากรณ์) Response ของลูกค้า. วิเคราะห์ผลด้วย

(10 + 10 = 20 คะแนน)

Problem 2

จากข้อมูลรถยนต์ที่กำหนดให้ (car-mpg.csv) 428 รายการ (ดังตัวอย่างบางส่วนที่แสดงตามมา) หากมีรถคันหนึ่งที่มีระบบขับเคลื่อน (drivetrain) ล้อหน้า เครื่องยนต์ความจุ 2.8 ลิตร 4 สูบ 150 แรงม้า น้ำหนัก 2500 กก. ความยาวฐานล้อ 108" ความยาวตัวรถ 172" ถามว่า... 1) รถคันนี้น่าจะกินน้ำมันเท่าใด (กิโลเมตร/แกลลอน (MPG))

2) วิเคราะห์ผลที่ได้รวมถึงน้ำหนักของแต่ละปัจจัยที่ใช้ด้วย (5 คะแนน)

Make	DriveTrain	EngineSize	Cylinders	Horsepower	MPG	Weight	Wheelbase	Length
Acura	All	3.5	6	265	17	4451	106	189
Acura	Front	2	4	200	24	2778	101	172
Acura	Front	2.4	4	200	22	3230	105	183
Acura	Front	3.2	6	270	20	3575	108	186
Acura	Front	3.5	6	225	18	3880	115	197
Acura	Front	3.5	6	225	18	3893	115	197
Acura	Rear	3.2	6	290	17	3153	100	174
Audi	Front	1.8	4	170	22	3252	104	179
Audi	Front	1.8	4	170	23	3638	105	180
Audi	Front	3	6	220	20	3462	104	179

Problem 3

ไฟล์ data3.csv แสดงข้อมูลของ x และ y โดยมีทั้งหมด 120 รายการ (ดังแสดงตัวอย่างบางส่วนตามมา) ให้หาความสัมพันธ์ทางคณิตศาสตร์ที่ดีที่สุดระหว่างตัวแปรต้น x กับตัวแปรอิสระ y พร้อมวิเคราะห์ผลที่ได้.

นอกจากนี้ให้ประมาณค่าของ y หาก x มีค่าเป็น 170 (5 คะแนน)

x	y
-70	15766.63
-68	15547.53
-66	14575.77
-64	11426.51
-62	11376.55
-60	9810.183
-58	9095.248
-56	9688.487
-54	7055.753
-52	6239.81