**Lab 6** **204453: Pattern Recognition**

**คำสั่ง:** ให้นักศึกษาดาวน์โหลดไฟล์ชื่อ Adult จาก UC Irvine Machine Learning Repository ให้ทำการทดลอง

ตัวจำแนกนาอีฟเบส์ (NaiveBayes) และ ~~เครือข่ายความเชื่อเบร์เซี่ยน (Bayesian Belief Network)~~ ตัวจำแนกเพื่อนบ้านใกล้ที่สุด (*k*-NN, k=5)

1. ไฟล์ Adult – UCI Machine Learning

* รายละเอียดการรู้จำแบบ – ลักษณะไฟล์ที่ใช้ คือ adult.data ซึ่งเป็น Text File เราจึงต้องทำการแปลงไฟล์นี้ให้อยู่ในรูปตาราง CSV ด้วยโปรแกรม Microsoft Excel ก่อน จึงจะได้ adult.csv เพื่อที่จะนำเข้าไปทดลองตัวจำแนก Naïve Bayes และ K-Nearest Neighbours ที่มีค่า k=5
* ค่าตัวแปร (Feature)  
  – age เป็นตัวแปรประเภท Integer แทนตัวเลขอายุของผู้ใหญ่ โดยมีอายุตั้งแต่ 17-90 ปี  
  – workclass เป็นตัวแปรประเภท Categorical แทนชนิดของงานที่ผู้ใหญ่ทำ โดยมีค่าที่เป็นไปได้ คือ Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.  
  – fnlwgt เป็นตัวแปรประเภท Integer แทนจำนวนผู้ใหญ่ที่มีข้อมูลคล้ายกับผู้ใหญ่ที่อยู่ในระเบียนนั้น โดยชื่อของตัวแปรนี้ย่อมาจากคำว่า final weight  
  – education เป็นตัวแปรประเภท Categorical แทนระดับการศึกษาสูงสุดของผู้ใหญ่ โดยมีค่าที่เป็นไปได้ คือ Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.  
  – education-num เป็นตัวแปรประเภท Integer แทนระดับการศึกษาสูงสุดของผู้ใหญ่ โดยเป็นตัวเลขซึ่งสัมพันธ์กับตัวแปร education ตั้งแต่ 1 ถึง 16 เมื่อ 16 เป็นระดับการศึกษาที่สูงที่สุด  
  – marital-status เป็นตัวแปรประเภท Categorical แทนสถานภาพการสมรสของผู้ใหญ่ โดยมีค่าที่เป็นไปได้ คือ Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.  
  – occupation เป็นตัวแปรประเภท Categorical แทนอาชีพที่ผู้ใหญ่ทำ โดยมีค่าที่เป็นไปได้ คือ Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.  
  – relationship เป็นตัวแปรประเภท Categorical แทนลักษณะความสัมพันธ์ของผู้ใหญ่ โดยมีค่าที่เป็นไปได้ คือ Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.  
  – race เป็นตัวแปรประเภท Categorical แทนเชื้อชาติของผู้ใหญ่ โดยมีค่าที่เป็นไปได้ คือ White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.  
  – sex เป็นตัวแปรประเภท Categorical แทนลักษณะเพศของผู้ใหญ่ โดยมีค่าที่เป็นไปได้ คือ Female, Male.  
  – capital-gain เป็นตัวแปรประเภท Integer แทนจำนวนรายได้ของผู้ใหญ่จากการค้าขาย  
  – capital-loss เป็นตัวแปรประเภท Integer แทนจำนวนขาดทุนของผู้ใหญ่จากการค้าขาย  
  – hours-per-week เป็นตัวแปรประเภท Integer แทนจำนวนชั่วโมงที่ผู้ใหญ่ทำงานต่อสัปดาห์  
  – native-country เป็นตัวแปรประเภท Categorical แทนประเทศเกิดของผู้ใหญ่ โดยมีค่าที่เป็นไปได้ คือ United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.  
  – income เป็นตัวแปรประเภท Binary แทนช่วงรายได้ของผู้ใหญ่ 2 ช่วง ได้แก่ >50K: รายได้มากกว่า $50,000 และ <=50K: รายได้น้อยกว่าหรือเท่ากับ $50,000
* ชั้นข้อมูล – ชุดข้อมูล Adult มีชั้นข้อมูล (Class) 2 ชั้นเป็นข้อมูลจำนวนรายได้ (Income) ของผู้ใหญ่ โดยชั้นที่ 1 คือ <= 50K (รายได้น้อยกว่าหรือเท่ากับ $50,000) และชั้นที่ 2 คือ > 50K (รายได้มากกว่า $50,000)
* ค่าทางสถิติ – ชุดข้อมูล Adult มี 32,561 ระเบียน, มี 15 สดมภ์, มี 2 ชั้นข้อมูล, มี 24,720 ข้อมูลในชั้นข้อมูลที่ 1 และมี 7,841 ข้อมูลในชั้นข้อมูลที่ 2

2. การกระจายตัวของข้อมูล

– เลือกคู่แกน age กับ hours-per-week มาเพื่อให้สามารถเห็นการกระจายตัวของข้อมูลได้ชัดเจนที่สุด  


3. ผลการทดลองแบบ Cross-validation (Folds = 10)

|  |  |  |
| --- | --- | --- |
| **Accuracy** | **Naïve Bayes** | ***k*-NN (*k* = 5)** |
| Dataset | 83.4772 % | 82.5374 % |
| <= 50K | 93.4264 % | 90.4045 % |
| > 50K | 52.1107 % | 57.7350 % |

4. สรุปผลการทดลอง

– จากการทดลองใช้ตัวจำแนก Naïve Bayes และตัวจำแนก *k*-NN พบว่าตัวจำแนก Naïve Bayes สามารถทำการจำแนกกลุ่มชุดข้อมูล Adult ในภาพรวมได้ดีกว่า *k*-NN แต่เมื่อเราได้ทำการตรวจสอบ Confusion matrix แล้ว พบว่าตัวจำแนก Naïve Bayes สามารถจำแนกกลุ่มข้อมูลสำหรับกลุ่ม <= 50K ได้แม่นยำกว่า *k*-NN แต่ *k*-NN นั้นสามารถจำแนกกลุ่มของ > 50K ได้แม่นยำกว่าตัวแบบ Naïve Bayes

– เนื่องจาก Class: <= 50K มีจำนวนข้อมูล 24,720 ข้อมูล แต่ Class: > 50K มีจำนวนข้อมูล 7,841 ข้อมูล พบว่า Class: <= 50K มีขนาดใหญ่กว่ามาก ทำให้เมื่อมีการจำแนกกลุ่มได้ถูกต้องเป็นจำนวนมากใน Class นี้ จะทำให้ Class: > 50K ซึ่งอาจมีปัญหาในการจำแนกกลุ่มได้อย่างถูกต้อง ถูกลดความสำคัญลงไปด้วย ดังนั้น Accuracy รวมที่ได้ในชุดข้อมูล Adult จึงยังไม่มีความน่าเชื่อถือพอ

**หมายเหตุ:** รายงาน ให้ใช้รูปแบบของ Lab 5