

การบ้านครั้งที่ 3 CS358 ภาคเรียนที่ 2/2560

กำหนดส่ง วันศุกร์ที่ 9 มีนาคม 2561 เวลา 23.55น.

วัตถุประสงค์ เพื่อให้ น.ศ. ได้ฝึกฝนทักษะการเขียนโปรแกรมวิเคราะห์ข้อมูลด้วยภาษา R และการสร้าง Tree-based model จากความรู้ที่ได้เรียนในชม.บรรยาย

1. ให้ น.ศ. สร้าง Decision Tree Model สำหรับชุดข้อมูล Mushroom จาก UCI Machine Learning repository (จากแหล่งอ้างอิง 1) เพื่อแก้ปัญหาในการจำแนกเห็ดประเภทรับประทานได้และเห็ดมีพิษ
2. ให้ดำเนินการและเขียนรายงานผลการดำเนินการ ประกอบด้วยหัวข้อดังนี้
 - 2.1. การเตรียมชุดข้อมูล (Data acquisition)
อธิบายว่า น.ศ. อ่านข้อมูลเข้ามาอย่างไร ทำอะไรบ้างเพื่อศึกษาให้เข้าใจข้อมูล และเข้าใจข้อมูลว่าอย่างไรบ้าง แนบ code ประกอบ
 - 2.2. การแบ่งข้อมูลเพื่อ Train และ Test แบบจำลอง (Data partitioning)
อธิบายหลักในการแบ่งข้อมูลเป็นชุดสำหรับ train และ test แบบจำลอง แนบ code ประกอบ
 - 2.3. การเลือก Attribute เพื่อสร้างแบบจำลอง (Attribute selection)
อธิบายกระบวนการในการเลือก attribute ที่มีความสำคัญในการสร้างแบบจำลอง ใช้หลักการอะไร ดำเนินการอย่างไร แนบ code ประกอบ ได้ผลอย่างไร
 - 2.4. การแสดงภาพเกี่ยวกับ Attribute ที่เลือก (Attribute visualization)
อธิบายสิ่งที่ทำให้เข้าใจความสัมพันธ์ของ attribute ที่ได้จากข้อ 2.3 กับค่าที่ต้องการพยากรณ์ให้มากขึ้น แนบ code ประกอบ และระบุผลที่ได้
 - 2.5. Classification ด้วย Decision Tree (Classification with Decision Tree) แบบที่ได้เรียนมาในชม.บรรยาย
 - 2.5.1. ใช้วิธีการเขียนฟังก์ชันต่างๆ เพื่อสร้าง decision tree เอง อาศัย data.tree package (อ้างอิงจาก Lab)
อธิบายขั้นตอนการสร้างแบบจำลองด้วยวิธีการแบบที่ทำใน lab แนบ code ประกอบ และแสดง decision tree ที่ได้ รวมทั้งผลการทดสอบด้วย
 - 2.5.2. ใช้วิธีการเรียกไลบรารีสำเร็จรูปในการสร้าง decision tree เช่น rpart package
อธิบายขั้นตอนการสร้างแบบจำลองด้วยการเรียกใช้ฟังก์ชันในไลบรารีสำเร็จรูป แนบ code ประกอบ อธิบายที่มาของค่าพารามิเตอร์ต่างๆ ที่ใช้ และแสดง decision tree ที่ได้ รวมทั้งผลการทดสอบด้วย
 - 2.6. สรุปองค์ความรู้ที่ได้จากการใช้แบบจำลองในการแก้ปัญหา และสิ่งที่ได้เรียนรู้เกี่ยวกับกระบวนการในการใช้ข้อมูลแก้ปัญหาจากการบ้านนี้

เกณฑ์การให้คะแนน

- | | |
|--|-----|
| 1) มี source code ที่เกี่ยวข้องทั้งหมด และสามารถรันเพื่อแสดงกระบวนการทั้งหมดได้ (น.ศ. อาจจะต้องแนบ readme เพื่ออธิบายวิธีการรันด้วย) | 30% |
| 2) ประสิทธิภาพของแบบจำลองที่ได้ | 10% |
| 3) คุณภาพรายงาน (pdf) | 60% |
| 4) Extra credit | 20% |

หาก น.ศ. อัป source code และรายงานบน GitHub หรือแหล่งออนไลน์สาธารณะอื่นๆ เช่น Jupiter notebook, Kaggle notebook, Shiny, Medium, Facebook, Youtube เป็นต้น

แหล่งอ้างอิงสำคัญที่ไม่ควรมองข้าม

- 1) <https://www.kaggle.com/uciml/mushroom-classification>
- 2) <https://www.kaggle.com/mokosan/mushroom-classification>
- 3) <https://fderyckel.github.io/machinelearningwithr/case-study-mushrooms-classification.html>