**Lab 5** **204453: Pattern Recognition**

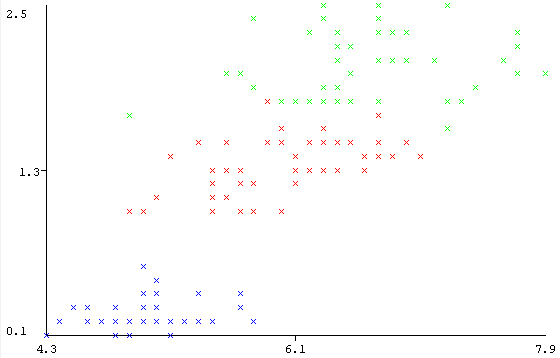
**คำสั่ง:** ให้นักศึกษาดาวน์โหลดไฟล์ชื่อ Iris จาก UC Irvine Machine Learning Repository จากนั้นให้ทำการทดลอง

ขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเคตัว (IBk) โดยให้ k มีค่าเท่ากับ 1, 3, 5, 7, 9 โดยรายงานมีรายละเอียด ดังนี้

1. ไฟล์ (ศึกษาจากเอกสารหรือในเว็บไซต์)

* รายละเอียดการรู้จำแบบ (ไฟล์ด้านไหน ต้องการทำนายอะไร)  
  **เป็นไฟล์ชุดข้อมูลของสปีชีส์ดอกไม้ Iris มี 3 สายพันธ์ คือ Setosa, Versicolor, และ Virginica โดยมีจำนวน Sample ดอกไม้ Iris สายพันธ์ละ 50 ดอก ต้องการทำนายว่าดอกไม้ Iris ที่มีความยาวและความกว้างของกลีบเลี้ยงและกลีบดอก จะถูกจัดว่าดอกไม้ดอกนั้นเป็นสายพันธ์ใด**
* ชั้นข้อมูล (แต่ละชั้นคือประเภทอะไร มีความแตกต่างกันอย่างไร)  
  **มี 3 ชั้นข้อมูล (Class) โดยมีชั้นข้อมูลดังนี้  
  1. Iris-setosa  
  มีความยาวและความกว้างของกลีบเลี้ยงและกลีบดอกที่น้อยที่สุดในทั้ง 3 สายพันธ์  
  2. Iris-versicolor  
  มีความยาวและความกว้างของกลีบเลี้ยงและกลีบดอกที่อยู่กึ่งกลางในทั้ง 3 สายพันธ์  
  3. Iris-virginica  
  มีความยาวและความกว้างของกลีบเลี้ยงและกลีบดอกที่มากที่สุดในทั้ง 3 สายพันธ์**
* ค่าทางสถิติ (จำนวนระเบียน Row, จำนวนสดมภ์ Column, จำนวนชั้นข้อมูล Class, จำนวนข้อมูลในแต่ละชั้นข้อมูล)  
  **จำนวนระเบียน (Rows) = 150 ระเบียน  
  จำนวนสดมภ์ (Columns) = 5 สดมภ์ (ไม่นับคอลัมน์ Id)  
  จำนวนชั้นข้อมูล (Classes) = 3 ชั้นข้อมูล  
  จำนวนข้อมูลในแต่ละชั้นข้อมูล = 50, 50, 50**

1. การกระจายตัวของข้อมูล

* เข้าเมนู Visualize แล้วให้เลือกคู่ของแกนที่เห็นการกระจายตัวชัดเจนที่สุด  
  **คู่ของแกนที่เลือก คือ SepalLengthCm, PetalWidthCm**  
  

1. ผลการทดลองแบบ Cross-validation (Folds = 10) ให้ทดลองแยกแต่ละค่า K

* Accuracy รวมทุกคลาสคำนวณจาก Correctly Classified Instances  
  **K = 1 มี Accuracy = 95.3333 %  
  K = 3 มี Accuracy = 95.3333 %  
  K = 5 มี Accuracy = 95.3333 %  
  K = 7 มี Accuracy = 96.6667 %  
  K = 9 มี Accuracy = 96 %  
  K = 11 มี Accuracy = 95.3333 %  
  K = 13 มี Accuracy = 97.3333 %**
* Accuracy แต่ละคลาสคำนวณจาก Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **Accuracy ของ**  **Setosa** | **Accuracy ของ Versicolor** | **Accuracy ของ Virginica** |
| 1 | **100 %** | **94 %** | **92 %** |
| 3 | **100 %** | **94 %** | **92 %** |
| 5 | **100 %** | **94 %** | **92 %** |
| 7 | **100 %** | **96 %** | **94 %** |
| 9 | **100 %** | **94 %** | **94 %** |
| 11 | **100 %** | **94 %** | **92 %** |
| 13 | **100 %** | **98 %** | **94 %** |

1. สรุปผลการทดลอง

* ชั้นข้อมูลใดถูกทำนายแม่นยำมากที่สุด ในแต่ละค่า K  
  **ชั้นข้อมูล Iris-setosa ซึ่งมี Accuracy = 100 % ในแต่ละค่า K**
* ชั้นข้อมูลใดถูกทำนายผิดพลาดมากที่สุด ในแต่ละค่า K  
  **ชั้นข้อมูล Iris-virginica เนื่องจากชั้นข้อมูลนี้ถูกทำนายแม่นยำได้น้อยที่สุดใน 3 ชั้นข้อมูล**

1. ทดลองปรับค่าพารามิเตอร์ distanceWeighting และ distanceFunction (อยู่ใน LinearNNSearch) อภิปรายผลการทดลอง ให้นักศึกษาตอบคำถามด้านล่างนี้พร้อมเหตุผลประกอบ

**“พารามิเตอร์ชุดใดให้ค่า Accuracy สูงสุด”   
จากค่าที่พบว่ามีค่าความแม่นยำสูงสุด จึงเลือกใช้ค่า K = 13  
distanceFunction แบบ Euclidean Distance  
ได้ Accuracy = 97.3333 %**

**ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1/distance**

**ได้ Accuracy = 96.6667 %  
ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1-distance**

**ได้ Accuracy = 97.3333 %  
พบว่าพารามิเตอร์ distanceFunction แบบ Euclidean Distance และใช้ distanceWeighting ที่เป็น Weight by 1-distance หรือ ไม่มี distance weighting ให้ค่า Accuracy สูงสุด คือ 97.3333 % ซึ่งสามารถทำนายและจำแนกพันธ์ดอกไม้ได้อย่างถูกต้องและแม่นยำสูง  
  
ทดลองเปลี่ยน distanceFunction เป็น Chebyshev Distance**

**ได้ Accuracy = 96.6667 %**

**ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1/distance**

**ได้ Accuracy = 96.6667 %  
ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1-distance**

**ได้ Accuracy = 96.6667 %  
พบว่าพารามิเตอร์ distanceFunction แบบ Chebyshev Distance และการเปลี่ยนค่า distanceWeighting ซึ่งไม่ทำให้ความแม่นยำในการทำนายเปลี่ยนแปลงเลย ให้ค่า Accuracy สูงสุด คือ 96.6667 % ซึ่งในกรณีนี้ สามารถทำนายและจำแนกพันธ์ดอกไม้ได้แม่นยำน้อยกว่า Euclidean Distance**

**ทดลองเปลี่ยน distanceFunction เป็น Manhattan Distance**

**ได้ Accuracy = 94.6667 %**

**ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1/distance**

**ได้ Accuracy = 95.3333 %  
ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1-distance**

**ได้ Accuracy = 94.6667 %  
พบว่าพารามิเตอร์ distanceFunction แบบ Manhattan Distance และใช้ distanceWeighting ที่เป็น Weight by 1/distance ให้ค่า Accuracy สูงสุด คือ 95.3333 % ซึ่งในกรณีนี้ สามารถทำนายและจำแนกพันธ์ดอกไม้ได้แม่นยำน้อยกว่า Euclidean Distance  
  
ทดลองเปลี่ยน distanceFunction เป็น Filter Distance**

**ได้ Accuracy = 95.3333 %**

**ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1/distance**

**ได้ Accuracy = 96 %  
ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1-distance**

**ได้ Accuracy = 95.3333 %  
พบว่าพารามิเตอร์ distanceFunction แบบ Filter Distance และใช้ distanceWeighting ที่เป็น Weight by 1/distance ให้ค่า Accuracy สูงสุด คือ 96 % ซึ่งในกรณีนี้ สามารถทำนายและจำแนกพันธ์ดอกไม้ได้แม่นยำน้อยกว่า Euclidean Distance  
  
ทดลองเปลี่ยน distanceFunction เป็น Minkowski Distance**

**ได้ Accuracy = 97.3333 %**

**ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1/distance**

**ได้ Accuracy = 96.6667 %  
ทดลองปรับค่าพารามิเตอร์ distanceWeighting เป็น Weight by 1-distance**

**ได้ Accuracy = 97.3333 %  
พบว่าพารามิเตอร์ distanceFunction แบบ Minkowski Distance และใช้ distanceWeighting ที่เป็น Weight by 1-distance หรือ ไม่มี distance weighting ให้ค่า Accuracy สูงสุด คือ 97.3333 % ซึ่งสามารถทำนายและจำแนกพันธ์ดอกไม้ได้อย่างถูกต้องและแม่นยำสูงเทียบเท่ากับฟังก์ชัน Euclidean Distance**

**หมายเหตุ:** ออกแบบรายงานตามความเหมาะสม ควรให้เรียบร้อย สวยงาม และ อ่านง่าย