

การเขียนโปรแกรมเพื่อใช้โครงข่ายประสาทเทียมสำหรับ

จำแนกชุดข้อมูล **Breast Cancer Wisconsin (Original)**

โดย

ชื่อ พิชยุทธ หันชัยเนาว์

รหัสนักศึกษา 640610653

Computer Engineering

Chiang Mai University

1. **ให้แสดงรายละเอียดของงานที่ส่งได้แก่ชื่อ และไฟล์**

a รายงาน ให้บันทึกชื่อไฟล์เป็นรหัสนักศึกษา แจ้งรายละเอียดและได้ส่งไฟล์ขึ้นระบบ (60 คะแนน)

b โปรแกรม แจ้งรายละเอียดและได้ส่งไฟล์ขึ้นระบบและ run ได้ (25 คะแนน)

c function ที่ใช้ แจ้งรายละเอียดและได้ส่งไฟล์ขึ้นระบบ (5 คะแนน)

d Training dataset แจ้งรายละเอียดและได้ส่งไฟล์ขึ้นระบบ (5 คะแนน)

e Testing dataset แจ้งรายละเอียดและได้ส่งไฟล์ขึ้นระบบ (5 คะแนน)

กำหนดให้ นักศึกษาสามารถใช้ dataset จากฐานข้อมูล UCI <https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original>

โดยให้เลือกชุดข้อมูล

Breast Cancer Wisconsin (Original)

Donated on 7/14/1992

Breast Cancer Wisconsin มี **2 class** ในcloumnสุดท้ายได้แก่ 2 คือclass benign และ 4 คือclass malignant, และมี **9 Attribute** ได้แก่ 1.Clump Thickness, 2.Uniformity of Cell Size, 3.Uniformity of Cell Shape, 4.Marginal Adhesion, 5.Single Epithelial Cell Size, 6.Bare Nuclei, 7.Bland Chromatin, 8.Normal Nucleoli, 9.Mitoses จำนวน699 instances ประเภทข้อมูลเป็นจำนวนจริง และมี missing value จำเป็นต้องมีการ Preprocess โดยในการทดลองนี้ จะมีการ fill missing value โดยใช้ mode

เผยแพร่ตั้งแต่ 15 July 1992

**นำรายละเอียดข้อมูลมากจาก**

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

(เข้าถึงข้อมูลเมื่อ 20 Feb. 2024)

1. **รายละเอียดเกี่ยวกับ Dataset ที่ใช้ในการทดลองตามที่ นศ เข้าใจ (5 คะแนน)**

Dataset นี้เป็นข้อมูลเกี่ยวกับมะเร็งเต้านม ที่ได้มาจาก มหาวิทยาลัย Wisconsin โรงพยาบาล Madison จาก Dr. William H. Wolberg. United States of America(USA) ในDataset มีทั้งหมด 11 column 1. **ID** (หมายเลขประจำตัว) 2. **Clump Thickness** (ความหนาของก้อนมะเร็ง) 3. **Uniformity of Cell Size** (ความสม่ำเสมอของขนาดเซลล์) 4. **Uniformity of Cell Shape** (ความสม่ำเสมอของรูปร่างของเซลล์) 5. **Marginal Adhesion** (การยึดเกาะขอบ) 6. **Single Epithelial Cell Size** (เซลล์เยื่อบุผิวเดี่ยวขนาด) 7. **Bare Nuclei** (นิวเคลียสเปลือย) **8. Bland Chromatin** (เบลนด์โครมาติน) 9. **Normal Nucleoli** (นิวคลีโอลีปกติ) 10. **Mitoses** (ไมโทซิส) 11. **Class** (หมวดหมู่ 2คือไม่เป็นมะเร็ง 4คือเป็นมะเร็ง) โดยในDataset มี Missing Value ใน column **Bare Nuclei** จำเป็นต้องมีการ preprocessก่อนโดยใช้การ fill missing value โดยใช้ mode (ฐานนิยม) และข้อมูลที่จะนำไป processได้ คือตั้งแต่column 2-11 โดย Attribute ID จะไม่ถูกนำมาprocess

1. **ขั้นตอนการทำ Data pre-processing (เช่น noise, missing value, transform etc) อธิบายที่มาและความเหมาะสมที่เลือกใช้วิธีการดังกล่าว (10 คะแนน)**

ในOriginal Dataset มี missing value ทั้งหมด 16 ตัว

































จึงได้มีการเลือกการทำpreprocess โดยการ fill missing value โดยใช้ mode หรือ ฐานนิยม เพราะAttribute **Bland Chromatin** (เบลนด์โครมาติน) Attribute เดียวที่มี missing value และ value ของ Attribute นี้มีการกระจากตัวที่ value=1 มากที่สุด และ จากการทดลองใช้การ fill missing value จาก mean เกิดการกระจายตัวที่เกินความเป็นจริงและเป็น outlier และ รูปแบบเป็น Decimal

1. **ออกแบบการทดลอง สมติฐานการทดลอง เช่น การทดสอบโครงสร้าง(อธิบายว่ามีกี่ Layer และรายละเอียดของแต่ละ Layer) ให้ใช้ activation function เป็น sigmoid การทดสอบจำนวนโหนด(จำนวนโหนดในแต่ละชั้น) เป็นต้น กำหนดให้ทำ นศ ทำการแบ่งสัดส่วน Train และ Test เอง**
   1. แนวคิดการแบ่งสัดส่วนเพื่อกำหนด Training set และ Testing set (5 คะแนน)



* 1. แนวคิดในการทดลองเพื่อกำหนดโครงสร้าง NN ที่เหมาะสม เช่น มี output layer, มี 2 hidden layer (5 คะแนน)
  2. แนวคิดในการทดลองเพื่อกำหนด จำนวน node ที่เหมาะสม (5 คะแนน)
  3. ความสัมพันธ์ระหว่างการกำหนดค่า learning rate และ จำนวน epoch ที่ได้ทดลอง (5 คะแนน)

1. **สรุป โครงสร้างที่เลือกใช้ (5 คะแนน)**

**ผลการทดลอง การรัน ความแม่นยำในการทำนาย Training data และ Testing data (10 คะแนน)**

**table**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Graph**

1. **วิเคราะห์และสรุปผลการทดลอง , ข้อเสนอแนะในการพัฒนาต่อไป (10 คะแนน)**