CPE 232 Data Models

Mock Final Exam (ข้อสอบจำลองปลายภาค)

เวลาสอบ: 3 ชั่วโมง | คะแนนเต็ม: 100 คะแนน | สอบ Onsite

ประกาศและขอบเขตข้อสอบ

ข้อสอบครอบคลุมเนื้อหา Lecture 6–9 ได้แก่ Machine Learning Classification, Regression, Parameter Tuning และ Clustering ข้อสอบ**เน้นข้อเขียนและคำนวณ (ประมาณ 90%)** MCQ มีเพียงเล็กน้อย

สิ่งที่อนุญาต: เครื่องเขียน, บัตรนักศึกษา และ **Formula Sheet** ที่อาจารย์แจกให้ (ไม่ต้องท่องสูตร)

โครงสร้างคะแนน: Part 1 (10) | Part 2 (10) | Part 3 (15) | Part 4 (30) | Part 5 (35)

Part 1: Multiple Choice / Multiple Selection / True-False (10 คะแนน)

1.1 Multiple Choice — เลือกคำตอบที่ถูกต้องที่สุด 1 ข้อ (5 ข้อ ข้อละ 1 คะแนน)

1. ข้อใดอธิบาย Naive Bayes Classifier ได้ถูกต้องที่สุด?

a. สร้างต้นไม้ตัดสินใจโดยใช้ Information Gain เลือก attribute ที่ดีที่สุด

b. ใช้ทฤษฎีของเบย์และสมมติว่า features ทุกตัวมีความเป็นอิสระต่อกัน

c. หาจุดที่ใกล้ที่สุด k จุดในพื้นที่ feature แล้ว vote ผลลัพธ์

d. สร้าง Hyperplane เพื่อแบ่งกลุ่มข้อมูลให้มี margin กว้างที่สุด

2. เหตุใด MSE จึงลงโทษ error ขนาดใหญ่หนักกว่า MAE?

a. MSE ยก error เป็นกำลังสองทำให้ค่า error ที่มากถูก amplify มากกว่า

b. MSE คำนวณค่า median แทนค่าเฉลี่ยจึงไม่ไว้วางใจ outlier

c. MSE หารด้วย $\sqrt{n}$ ส่วน MAE หารด้วย n

d. MSE ใช้ log scale แทน linear scale ในการวัด error

3. ข้อใดเป็นข้อดีหลักของ k-Fold Cross-Validation เมื่อเทียบกับ Single Train/Test Split?

a. ใช้เวลา training น้อยกว่าเพราะ dataset เล็กลง

b. ลด variance ของการประเมิน model เพราะใช้ข้อมูลทุกส่วนเป็นทั้ง train และ validation

c. ไม่จำเป็นต้องสุ่มแบ่งข้อมูลก่อน

d. รับประกันว่า model จะไม่ overfit กับ training data

4. DBSCAN มีข้อได้เปรียบเหนือ K-Means อย่างไร?

a. ทำงานเร็วกว่าบน dataset ขนาดใหญ่เสมอ

b. ไม่ต้องระบุจำนวน cluster ล่วงหน้า และจัดการ noise/outlier ได้

c. Cluster ที่ได้มี centroid ชัดเจนทุกครั้ง

d. ต้องการ memory น้อยกว่าเพราะไม่เก็บ proximity matrix

5. Model มี Training Accuracy = 97%, Test Accuracy = 62% สถานการณ์นี้คืออะไร?

a. Overfitting

b. Underfitting

c. Good Generalization

d. High Bias, High Variance พร้อมกัน

1.2 Multiple Selection — เลือกทุกข้อที่ถูกต้อง (3 ข้อ ข้อละ 1 คะแนน)

หมายเหตุ: ต้องเลือกครบถ้วนทุกข้อที่ถูกจึงจะได้คะแนน

6. ข้อใดบ้างเป็น **Hyperparameter** (กำหนดก่อน training)?

a. `max_depth` ของ Decision Tree

b. ค่า weight ที่ model เรียนรู้ระหว่าง Gradient Descent

c. Learning Rate (α) ใน Gradient Descent

d. จำนวนต้นไม้ใน Random Forest (`n_estimators`)

e. ค่า bias term ที่ optimize ผ่าน loss function

7. ข้อใดบ้างช่วย**ลด Overfitting** ได้?

a. เพิ่มข้อมูล training

b. ลด `max_depth` ใน Decision Tree

c. เพิ่มจำนวน features ใน dataset

d. ใช้ L1/L2 Regularization

e. เพิ่ม learning rate เป็น 2 เท่า

8. ข้อใดบ้างเป็นข้อความที่**ถูกต้อง**เกี่ยวกับ Silhouette Score?

a. ค่าอยู่ในช่วง $[-1, +1]$

b. ค่าใกล้ +1 บ่งชี้ว่า cluster มีความแน่นและแยกจากกันชัดเจน

c. ค่าลบหมายความว่าจุดนั้นถูก assign ผิด cluster

d. ค่า 0 บ่งชี้ว่า cluster สมบูรณ์แบบ

e. ใช้ได้เฉพาะกับ K-Means เท่านั้น

1.3 True/False — ระบุ จริง (T) หรือ เท็จ (F) (4 ข้อ ข้อละ 0.5 คะแนน)

เขียน T หรือ F ในช่องว่าง

9. [ ] Entropy = 0 หมายความว่า node นั้น**ไม่บริสุทธิ์** (impure) อย่างสมบูรณ์

10. ☐ ค่า R^2 (Coefficient of Determination) อาจมีค่าติดลบได้ในบางกรณี

11. ☐ K-Means รับประกันว่า converge ไปที่ global optimum เสมอ

12. ☐ Stratified k-Fold Cross-Validation รักษาสัดส่วน class ในแต่ละ fold ให้ใกล้เคียงกับ dataset ต้นฉบับ

Part 2: Definition — เขียนนิยามความหมายของคำศัพท์ต่อไปนี้ (10 คะแนน, ข้อละ 1 คะแนน)

คำชี้แจง: ตอบเป็นภาษาไทยหรือภาษาอังกฤษก็ได้ ให้นิยามที่ครอบคลุมทั้งความหมายและบริบทการใช้งาน

1. **Descriptive Analytics** [Intro]

2. **Diagnostic Analytics** [Intro]

3. **Predictive Analytics** [Intro]

4. **Prescriptive Analytics** [Intro]

5. **Confusion Matrix** [Lecture 6]

6. **Overfitting** [Lecture 8]

7. **Hyperparameter** [Lecture 8]

8. **k-Fold Cross-Validation** [Lecture 8]

9. **Silhouette Score** [Lecture 9]

10. **DBSCAN (Density-Based Spatial Clustering)** [Lecture 9]

Part 3: Short Answer & Data Interpretation (15 คะแนน, ข้อละ 3 คะแนน)

คำชี้แจง: ตอบสั้น 2–4 บรรทัด เน้นการ “ตีความ” และ “อธิบายเหตุผล” ไม่ใช่เพียงแค่นิยาม

1. **[Lecture 6]** เพราะเหตุใดจึงควรเลือกใช้ **Random Forest** แทน **Decision Tree** เดี่ยว ในกรณีที่ต้องการ model ที่ generalize ดี? อธิบายกลไกที่ทำให้ Random Forest ดีกว่า

2. **[Lecture 8]** ระบบ AI ผ่าน training ด้วย Training Accuracy = 98% แต่เมื่อทดสอบกับข้อมูลใหม่ Test Accuracy = 62% ปัญหาคืออะไร? ควรแก้ไขอย่างไรบ้าง (ระบุอย่างน้อย 2 วิธี)

3. **[Lecture 6]** จาก Confusion Matrix ด้านล่าง จงตีความผลลัพธ์และอธิบายว่า classifier นี้มีจุดแข็งและจุดอ่อนอะไร ควรปรับปรุงอย่างไร

	ทำนาย: บวก (+)	**ทำนาย: ลบ (-)**
จริง: บวก (+)	20 (TP)	80 (FN)
จริง: ลบ (-)	5 (FP)	95 (TN)

4. **[Lecture 9]** ในการทำ Clustering เพราะเหตุใดจึงอาจพิจารณาเลือกใช้ **K-Medoids** แทน **K-Means**? อธิบายความแตกต่างของทั้ง 2 วิธีนี้มาพอสังเขป

5. **[Business Application]** สมมติคุณสร้างโมเดลทำนายการซื้อสินค้าและพบผลลัพธ์ว่า ‘กลุ่มลูกค้าที่น่าจะซื้อสินค้าใหม่มากที่สุดคือ กลุ่มนิสิตนักศึกษา ในช่วงเวลา 20.00-24.00 น.” ในเชิงธุรกิจ คุณจะออกนโยบายหรือกลยุทธ์ (Take action) อย่างไรเพื่อเพิ่มยอดขายหรือสร้างประสบการณ์ที่ดีให้กับลูกค้ากลุ่มนี้? (ยกตัวอย่าง 2 กลยุทธ์)

Part 4: Long Answer / Essay (30 คะแนน)

คำชี้แจง: ตอบอย่างครบถ้วน ใช้ตารางหรือเรียงความก็ได้ ประเมินจากความครบถ้วน ความถูกต้อง และความชัดเจน

ข้อ 1 — เปรียบเทียบ 4 Classification Models (10 คะแนน)

ระบบบริษัท FinTech ต้องการสร้าง model จำแนกว่าลูกค้ารายใดมีความเสี่ยงสูง (High Risk) หรือต่ำ (Low Risk) ในการผิดนัดชำระสินเชื่อ Dataset มี 50,000 samples, 30 features, class imbalance (High Risk 15%, Low Risk 85%)

จงเปรียบเทียบ 4 Classification Algorithms ต่อไปนี้:

- Decision Tree
- k-Nearest Neighbors (KNN)
- Naive Bayes
- Random Forest

การเปรียบเทียบต้องครอบคลุม:

1. หลักการทำงานโดยย่อ (1–2 ประโยค/algorithm)
2. ข้อดีและข้อจำกัดหลัก
3. ความเหมาะสมกับโจทย์นี้ โดยเฉพาะเรื่อง class imbalance และ dataset ขนาดใหญ่
4. สรุปว่าควรเลือก algorithm ใด พร้อมเหตุผล

ข้อ 2 — Types of Cross-Validation (10 คะแนน)

ส่วนที่ 2.1 อธิบาย Cross-Validation 4 แบบต่อไปนี้ โดยระบุ concept, ข้อดี, ข้อจำกัด, และเหมาะกับ dataset แบบใด:

- k-Fold Cross-Validation
- Stratified k-Fold Cross-Validation
- Leave-One-Out Cross-Validation (LOO-CV)
- Time Series Cross-Validation

ส่วนที่ 2.2 สถานการณ์ต่อไปนี้ควรใช้ Cross-Validation แบบใด พร้อมให้เหตุผล (1 คะแนน/ข้อ)

a. Dataset ราคาหุ้น SET50 ย้อนหลัง 5 ปี ต้องการ predict ราคาวันถัดไป

ตอบ:

b. Dataset ตรวจมะเร็ง 200 samples (Positive 10%, Negative 90%)

ตอบ:

ข้อ 3 — การประยุกต์ใช้ 4 Types of Analytics (10 คะแนน)

สมมติว่าคุณเป็นนักวิเคราะห์ข้อมูลของร้านค้าปลีก (Retail Store) บริษัทพบว่ายอดขายสาขาในห้างตกต่ำลง 15% จงประยุกต์ใช้ 4 Types of Analytics เพื่อนำเสนอโซลูชัน:

1. **Descriptive Analytics** (2.5 คะแนน) - จะดูข้อมูลส่วนใดหรือหา Insight อะไรเพื่อดูสิ่งที่เกิดขึ้น?

2. **Diagnostic Analytics** (2.5 คะแนน) - จะตั้งสมมติฐานหรือวิเคราะห์หาสาเหตุอย่างไร?

3. **Predictive Analytics** (2.5 คะแนน) - จะสร้างโมเดลคาดการณ์เรื่องใดในอนาคตเพื่อช่วยร้านค้า?

4. **Prescriptive Analytics** (2.5 คะแนน) - จะแนะนำ Automation หรือ Policy optimization อะไรให้ระบบทำงานได้เอง?

Part 5: Calculation Structure (35 คะแนน)

คำชี้แจง: แสดงวิธีทำให้ครบทุกขั้นตอน คำตอบที่ไม่มีวิธีทำจะไม่ได้คะแนน

ข้อ 5.1 — Confusion Matrix Metrics (10 คะแนน)

ระบบ AI วินิจฉัยโรคเบาหวาน ทดสอบกับ 200 patients ได้ผลดังนี้:

	ทำนาย: ป่วย (+)	**ทำนาย: ไม่ป่วย (-)**	**รวม**
จริง: ป่วย (+)	60	20	80
จริง: ไม่ป่วย (-)	40	80	120
รวม	100	100	200

5.1.1 (5 คะแนน) คำนวณค่าต่อไปนี้พร้อมแสดงวิธีคิด:

- Accuracy
- Precision
- Recall (Sensitivity)
- Specificity
- F1-Score

5.1.2 (5 คะแนน) จากผลลัพธ์ข้างต้น จงตีความ: ระบบนี้เหมาะสมสำหรับการวินิจฉัยโรคหรือไม่? metric ใดสำคัญที่สุดในบริบทนี้? และควรปรับปรุงอย่างไร?

ข้อ 5.2 — Regression: MSE, RMSE, MAE และ R^2 (10 คะแนน)

จาก dataset ต่อไปนี้ model Linear Regression ให้ค่า predict ดังนี้:

#	y_{actual}	$y_{\text{predicted}}$
1	2	2
2	4	4
3	5	6
4	8	8
5	11	10

5.2.1 (10 คะแนน) คำนวณค่าต่อไปนี้พร้อมแสดงวิธีคิดทุกขั้น:

- MSE (Mean Squared Error)
- RMSE (Root MSE)
- MAE (Mean Absolute Error)
- SSE และ SS_{total}
- R^2 (Coefficient of Determination)

ข้อ 5.3 — Hierarchical Agglomerative Clustering + Dendrogram + Nested Clusters (15 คะแนน)

ใช้ Complete Linkage: $d(\{C_i\}, \{C_j\}) = \max_{x \in C_i,\, y \in C_j} d(x, y)$

กำหนด Proximity Matrix ของจุดข้อมูล 5 จุด (A, B, C, D, E):

	A	B	C	D	E
A	0	1	5	6	7
B	1	0	4	6	8
C	5	4	0	2	3
D	6	6	2	0	3
E	7	8	3	3	0

5.3.1 (6 คะแนน) แสดงขั้นตอน Agglomerative Clustering ทีละ step จนเหลือ 1 cluster:
ระบุ: (1) คู่ที่ merge, (2) ระยะทางที่ merge, (3) Proximity Matrix ที่อัปเดตแล้ว ทุก iteration

5.3.2 (5 คะแนน) วาด **Dendrogram** โดยระบุแกน Y เป็นระยะทาง (distance) และแสดงระดับที่ merge แต่ละครั้งให้ชัดเจน

5.3.3 (2 คะแนน) วาด **Nested Cluster Diagram** แสดงการ group จุดข้อมูลในแต่ละ level

5.3.4 (2 คะแนน) ถ้าตัด Dendrogram ที่ระดับ $d = 4$ จะได้กี่ cluster และประกอบด้วยจุดใดบ้าง?

เฉลยและแนวคิด (Answer Key)

ส่วนนี้สำหรับตรวจหลังทำข้อสอบ แนะนำให้ทำข้อสอบเต็มเวลาก่อนเปิดเฉลย

Part 1: คำตอบ

1.1 MCQ

1. **c.** 2. **d.** 3. **b.** 4. **c.** 5. **d.**

1.2 Multiple Selection

- **ข้อ 6: a, c, d** (max_depth, Learning Rate, n_estimators คือ Hyperparameters; b,e คือ learned parameters)
- **ข้อ 7: a, b, d** (เพิ่มข้อมูล, ลด max_depth, L1/L2 Reg ลด overfit; c เพิ่ม features อาจทำ overfit มากขึ้น; e เพิ่ม LR อาจทำให้ diverge)
- **ข้อ 8: a, b, c** (ช่วง -1 ถึง +1, ค่าใกล้ +1 ดี, ค่าลบ = misclustered; d ผิด ค่า 0 = on boundary; e ผิด Silhouette ใช้กับ clustering ทั่วไป)

1.3 True/False

9. **F** (Entropy=0 คือ pure)
10. **T** ($R^2 < 0$ ได้)
11. **F** (แค่ local optima)
12. **T** (Stratified รักษา class ratio)

Part 2: แนวคำตอบ (Definitions)

1. **Entropy:** วัดความไม่แน่นอน (impurity) ของ node ใน Decision Tree $H(S) = -\sum_i p_i \log_2 p_i$ ค่า 0 = pure (class เดียว), ค่า 1 = equally mixed สองคลาส
2. **Information Gain:** การลดลงของ Entropy หลังแบ่งข้อมูลด้วย attribute นั้น $IG(S, A) = H(S) - \sum_v \frac{|S_v|}{|S|} H(S_v)$ ยิ่งสูงยิ่งเป็น attribute ที่ดีสำหรับ split
3. **Confusion Matrix:** ตาราง 2×2 (หรือ n×n) แสดงผล classification ประกอบด้วย TP (ทำนายถูก-บวก), TN (ทำนายถูก-ลบ), FP (ทำนายผิด-positive), FN (ทำนายผิด-negative)
4. **Overfitting:** สภาวะที่ model เรียนรู้ noise ใน training data จน Training acc สูงมาก แต่ Test acc ต่ำ — model ไม่สามารถ generalize กับข้อมูลใหม่ได้
5. **Gradient Descent:** Algorithm ปรับ parameter โดย update ในทิศทาง**ตรงข้าม** gradient ของ loss $\theta := \theta - \alpha \cdot \nabla_\theta L(\theta)$ เพื่อลด loss function ทีละขั้น
6. **Hyperparameter:** Parameter ที่**กำหนดก่อน training** และไม่ถูก optimize โดย model เช่น max_depth, learning_rate, k ใน KNN ต่างจาก learnable parameters ที่ model เรียนรู้ระหว่าง training
7. **k-Fold CV:** แบ่ง dataset เป็น k ส่วนเท่า ๆ กัน วนซ้ำ k รอบ โดยแต่ละรอบใช้ส่วนต่างกันเป็น validation set ที่เหลือเป็น training ค่า metric เฉลี่ยจาก k รอบ = ประสิทธิภาพของ model
8. **Centroid:** จุดศูนย์กลางของ cluster ใน K-Means คำนวณจากค่าเฉลี่ยของ feature ทุกมิติของจุดทั้งหมดใน cluster $\mu_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$

9. **Silhouette Score:** วัดคุณภาพ clustering ช่วง $[-1, +1]$ โดย $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$ โดย a=ระยะเฉลี่ยใน cluster, b=ระยะเฉลี่ยไปยัง cluster ใกล้ที่สุด ค่าสูงหมายถึง cluster ดี

10. **DBSCAN:** Algorithm clustering แบบ density-based กำหนด core points ด้วย ε (รัศมี) และ MinPts (จำนวนจุดขั้นต่ำ) ไม่ต้องระบุ k ล่วงหน้า จัดการ noise ได้ และ cluster มีรูปร่างได้อิสระ

Part 3: แนวคำตอบ

1. **Random Forest vs Decision Tree:** Random Forest ใช้ Bagging สร้าง Decision Trees หลายต้นจาก bootstrap samples ต่างกัน แต่ละต้น vote ผลลัพธ์ Ensemble หลายต้นลด variance ทำให้ model stable กว่า DT เดี่ยวที่ overfit ง่าย นอกจากนี้ Random Forest ยังสุ่มเลือก features แต่ละ split (feature subsampling) ลด correlation ระหว่างต้นไม้

2. **Overfitting:** Training 98%, Test 62% คือ Overfitting ชัดเจน Model จดจำ training data มากเกินไป วิธีแก้: (1) ลด max_depth / เพิ่ม min_samples_split ใน DT (2) เพิ่ม L1/L2 Regularization (3) ใช้ k-Fold CV เพื่อ estimate performance จริง (4) เพิ่มข้อมูล training (5) Early Stopping

3. **CM Interpretation:** TP=20, FN=80 หมายความว่า Recall = 20/100 = 0.20 ต่ำมาก — model "พลาด" ผู้ป่วยจริง 80% ถือว่า**อันตรายมาก** ในบริบทการแพทย์ ควรปรับ decision threshold ให้ต่ำลง หรือ oversample class บวก (SMOTE) เพื่อให้ model ตรวจจับ positive case ได้ดีขึ้น

4. **Silhouette 0.12:** ค่า 0.12 ใกล้ 0 บ่งชี้ว่า cluster ทับซ้อนกัน หรือจุดข้อมูลอยู่บน boundary ระหว่าง cluster — คุณภาพการ cluster**ต่ำมาก** วิธี diagnose: ใช้ Elbow Method (plot WCSS vs k) หรือ Silhouette Score vs k เพื่อหาค่า k ที่เหมาะสม

5. **Local Minimum:** เกิดเมื่อ gradient ลด loss ลงมาจนติดอยู่ที่จุดต่ำสุดเฉพาะที่ที่ไม่ใช่ global minimum gradient = 0 แต่ไม่ใช่คำตอบที่ดีที่สุด Adam optimizer แก้ปัญหาด้วยการรวม Momentum (สะสม gradient) และ Adaptive Learning Rate (ปรับ LR ตาม parameter) ทำให้ออกจาก local minimum และ saddle point ได้ดีกว่า SGD ธรรมดา

Part 4: แนวคำตอบ

ข้อ 1: เปรียบเทียบ 4 Models

Model	หลักการ	ข้อดี / ข้อจำกัด	FinTech Use Case
Decision Tree	แบ่ง dataset ด้วย attribute ที่ให้ IG สูงสุด	ตีความง่าย / overfit ง่าย, ไม่ stable	ใช้ได้แต่ต้อง prune; อ่านผลง่าย
KNN	วัดระยะทางหา k เพื่อนบ้านแล้ว vote	ง่าย / ช้ามากกับ 50K samples; ไม่ดีกับ imbalanced	**ไม่ เหมาะ** (dataset ใหญ่, class imbalance)
Naive Bayes	P(class|features) ด้วย Bayes + independence assumption	เร็ว / independence มักไม่จริง	ใช้ได้ เบื้องต้น แต่ features อาจ correlate กัน
Random Forest	Ensemble DT หลายต้น + Bagging + Feature sampling	Robust, ลด overfit / ตีความยากกว่า	**เหมาะ ที่สุด**: จัดการ imbalance ได้ดี, ไม่ overfit, scale ได้

สรุป: ควรเลือก **Random Forest** เนื่องจาก: (1) ทน imbalanced class ได้ดีกว่าเมื่อใช้ class_weight (2) generalize ดีกว่า DT เดี่ยว (3) scale ได้กับ 50K samples (4) feature importance ช่วยตีความได้

ข้อ 2: Cross-Validation Types

ประเภท	Concept	ข้อดี / ข้อจำกัด	เหมาะกับ
k-Fold	แบ่ง k ส่วน วน k รอบ	ประเมิน variance ดี / ใช้เวลา k เท่า	Dataset ทั่วไป balanced
Stratified k-Fold	k-Fold แต่รักษา class ratio ทุก fold	ดีกับ imbalanced / ใช้เวลาเท่า k-Fold	Class imbalanced
LOO-CV	k = n, test ทีละ 1 sample	ใช้ข้อมูลได้สูงสุด / ช้ามาก O(n)	Dataset เล็ก ($n < 100$)
Time Series CV	ขยาย training window ทีละ step	ไม่ data leakage จากอนาคต / ต้องเรียง	Time series, sequential data

ส่วน 2.2: (a) Stock price $\longrightarrow$ **Time Series CV** เพราะข้อมูลมี temporal dependency ห้ามสุ่ม
(b) Cancer dataset imbalanced $\longrightarrow$ **Stratified k-Fold** เพื่อรักษา class ratio ใน fold

Part 5: เฉลยการคำนวณ

5.1 Entropy & Information Gain

Dataset: 8 samples, 4 ใช่ (+), 4 ไม่ใช่ (-) $\Rightarrow H(S) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = -\frac{1}{2}(-1) - \frac{1}{2}(-1) =$ **1.0** bit

รายได้ (Income):

- สูง (High): rows 1,2,3 $\longrightarrow$ 3 ใช่, 0 ไม่ใช่ $\Rightarrow H = 0$
- ปานกลาง (Medium): rows 4,5 $\longrightarrow$ 1 ใช่, 1 ไม่ใช่ $\Rightarrow H = 1.0$
- ต่ำ (Low): rows 6,7,8 $\longrightarrow$ 0 ใช่, 3 ไม่ใช่ $\Rightarrow H = 0$
- $IG(\text{Income}) = 1.0 - \frac{3}{8}(0) - \frac{2}{8}(1.0) - \frac{3}{8}(0) = 1.0 - 0.25 =$ **0.75**

นักศึกษา (Student):

- ใช่: rows 1,2,4,7 $\longrightarrow$ 3 ใช่, 1 ไม่ใช่ $\Rightarrow H = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = -\frac{3}{4}(1.585 - 2) - \frac{1}{4}(-2) = \frac{3}{4}(0.415) + 0.5 \approx 0.811$
- ไม่ใช่: rows 3,5,6,8 $\longrightarrow$ 1 ใช่, 3 ไม่ใช่ $\Rightarrow H = 0.811$ (symmetric)
- $IG(\text{Student}) = 1.0 - \frac{4}{8}(0.811) - \frac{4}{8}(0.811) = 1.0 - 0.811 =$ **0.189**

อายุ (Age):

- น้อยกว่า 30: rows 1,2,5,6 $\longrightarrow$ 2 ใช่, 2 ไม่ใช่ $\Rightarrow H = 1.0$
- 30–50: rows 3,4,7,8 $\longrightarrow$ 2 ใช่, 2 ไม่ใช่ $\Rightarrow H = 1.0$
- $IG(\text{Age}) = 1.0 - \frac{4}{8}(1.0) - \frac{4}{8}(1.0) =$ **0.0**

สรุป IG: Income = 0.75 $>$ Student = 0.189 $>$ Age = 0.0

Root Node = รายได้ (Income) เพราะให้ Information Gain สูงสุด

5.2 Confusion Matrix

TP=60, FP=40, FN=20, TN=80, Total=200

$$\text{Accuracy} = \frac{60+80}{200} = \frac{140}{200} = \mathbf{0.70}$$

$$\text{Precision} = \frac{60}{60+40} = \frac{60}{100} = \mathbf{0.60}$$

$$\text{Recall} = \frac{60}{60+20} = \frac{60}{80} = \mathbf{0.75}$$

$$\text{Specificity} = \frac{80}{80+40} = \frac{80}{120} \approx \mathbf{0.667}$$

$$\text{F1} = \frac{2 \times 0.60 \times 0.75}{0.60 + 0.75} = \frac{0.90}{1.35} = \frac{2}{3} \approx \mathbf{0.667}$$

ตีความ: Recall = 0.75 หมายถึงผู้ป่วยจริง 25% ถูกพลาด (FN=20) ในบริบทการแพทย์ FN อันตรายกว่า FP เพราะผู้ป่วยไม่ได้รับการรักษา ควรลด threshold เพื่อเพิ่ม Recall แม้ Precision จะลดลงบ้าง

5.3 Regression

errors $(e_i = y_{actual} - y_{pred})$: 0, 0, -1, 0, 1

$$\text{MSE} = \frac{0^2 + 0^2 + (-1)^2 + 0^2 + 1^2}{5} = \frac{2}{5} = \mathbf{0.40}, \quad \text{RMSE} = \sqrt{0.40} \approx \mathbf{0.632}$$

$$\text{MAE} = \frac{0 + 0 + 1 + 0 + 1}{5} = \mathbf{0.40}$$

$$\bar{y} = \frac{2 + 4 + 5 + 8 + 11}{5} = 6.0 \quad \text{SSE} = \sum (y_i - \hat{y}_i)^2 = 0 + 0 + 1 + 0 + 1 = \mathbf{2}$$

$$SS_{total} = (2-6)^2 + (4-6)^2 + (5-6)^2 + (8-6)^2 + (11-6)^2 = 16 + 4 + 1 + 4 + 25 = \mathbf{50}$$

$$R^2 = 1 - \frac{SSE}{SS_{total}} = 1 - \frac{2}{50} = \mathbf{0.96}$$

Gradient Descent: $\theta_0^{new} = 1.0 - 0.1 \times 2.0 = \mathbf{0.80}, \quad \theta_1^{new} = 0.5 - 0.1 \times (-1.5) = 0.5 + 0.15 = \mathbf{0.65}$

5.4 K-Means

Iteration 1 ($\mu_1 = (1,1), \mu_2 = (8,4)$):

จุด	$d(\cdot, \mu_1)$	$d(\cdot, \mu_2)$	Cluster
A(1,1)	0	$\sqrt{58} \approx 7.62$	1
B(2,3)	$\sqrt{5} \approx 2.24$	$\sqrt{37} \approx 6.08$	1
C(3,2)	$\sqrt{5} \approx 2.24$	$\sqrt{29} \approx 5.39$	1
D(6,5)	$\sqrt{41} \approx 6.40$	$\sqrt{5} \approx 2.24$	2
E(7,6)	$\sqrt{61} \approx 7.81$	$\sqrt{5} \approx 2.24$	2
F(8,4)	$\sqrt{58} \approx 7.62$	0	2

$$\mu_1^{new} = \left(\tfrac{1+2+3}{3}, \tfrac{1+3+2}{3}\right) = (2,2), \quad \mu_2^{new} = \left(\tfrac{6+7+8}{3}, \tfrac{5+6+4}{3}\right) = (7,5)$$

Iteration 2 ($\mu_1 = (2,2), \mu_2 = (7,5)$): ทุกจุดยังอยู่ใน cluster เดิม (A,B,C $\longrightarrow$ Cluster 1; D,E,F $\longrightarrow$ Cluster 2)
Centroids ไม่เปลี่ยน $\Rightarrow$ **Converged!** ที่ Cluster 1 = {A,B,C}, Cluster 2 = {D,E,F}

5.5 Hierarchical Agglomerative Clustering

Step 1: min = $d(A,B) = 1 \longrightarrow$ Merge {A,B}
Complete Linkage: $d(\{AB\}, C) = \max(5,4) = 5$, $d(\{AB\}, D) = \max(6,6) = 6$, $d(\{AB\}, E) = \max(7,8) = 8$

	{AB}	C	D	E
{AB}	0	5	6	8
C	5	0	2	3
D	6	2	0	3
E	8	3	3	0

Step 2: min = $d(C,D) = 2 \longrightarrow$ Merge {C,D}
$d(\{AB\}, \{CD\}) = \max(5,6) = 6$, $d(\{CD\}, E) = \max(3,3) = 3$

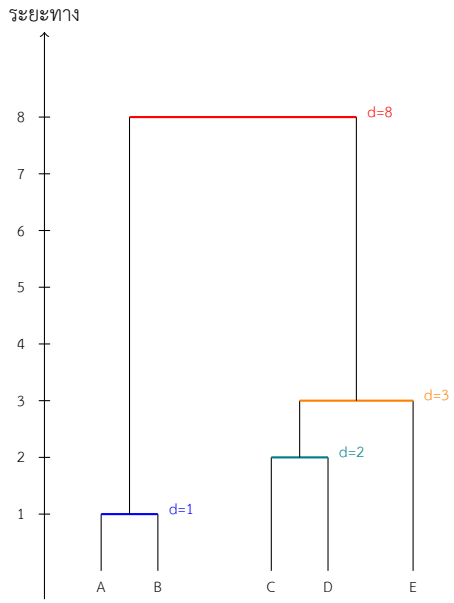
	{AB}	{CD}	E
{AB}	0	6	8
{CD}	6	0	3
E	8	3	0

Step 3: min = $d(\{CD\}, E) = 3 \rightarrow$ Merge {C,D,E}

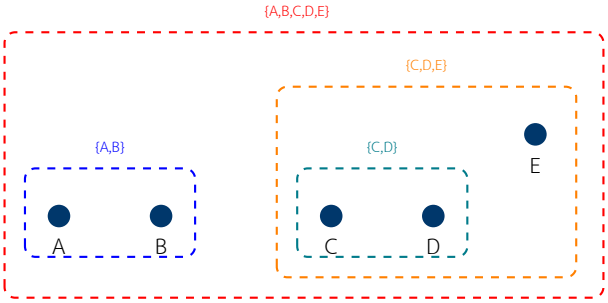
$d(\{AB\}, \{CDE\}) = \max(6, 8) = 8$

Step 4: Merge {AB} + {CDE} ที่ $d = 8$

Dendrogram (เฉลย):



Nested Cluster Diagram (เฉลย):



5.5.4: ตัดที่ $d = 4$ จะได้ **2 clusters**: {A,B} และ {C,D,E}

(เพราะ {A,B} merge กันที่ $d = 1 < 4$ และ {C,D,E} merge กันครบที่ $d = 3 < 4$ แต่ merge กัน 2 กลุ่มใหญ่ที่ $d = 8 > 4$)

Formula Sheet (สูตรสำหรับใช้ในห้องสอบ)

อาจารย์จะแจก Formula Sheet แผ่นนี้ในห้องสอบ — ไม่ต้องท่องสูตร แต่ต้องเข้าใจวิธีใช้

Lecture 6: Classification

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

$$IG(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

$$\text{Gini}(S) = 1 - \sum_{i=1}^{c} p_i^2$$

$$\text{Sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Lecture 7: Regression

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{RMSE} = \sqrt{\text{MSE}}, \quad \text{MAE} = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{SSE}{SS_{total}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$$\theta_j := \theta_j - \alpha \cdot \frac{\partial L}{\partial \theta_j}$$

$$L_{\text{Ridge}} = L + \lambda \sum \theta_j^2$$

$$L_{\text{Lasso}} = L + \lambda \sum |\theta_j|$$

Lecture 9: Clustering

$$d_{\text{Euclidean}} = \sqrt{\sum_k (x_k - y_k)^2}$$

$$d_{\text{Manhattan}} = \sum_k |x_k - y_k|$$

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2$$

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$$d_{\text{Complete}}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

$$d_{\text{Single}}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

RECAP: สรุปเนื้อหาจำสอบ (อ่านก่อนสอบ)

RECAP 1: Classification (Lecture 6)

- **Decision Tree:** เลือก root ด้วย IG สูงสุด, Entropy=0 คือ pure node, Gini ใช้แทน Entropy ได้
- **KNN:** หา k เพื่อนบ้านใกล้สุด, ช้าตอน predict, ไม่มี training phase, ต้อง scale features
- **Naive Bayes:** สมมติ features อิสระกัน ทำงานเร็วมาก เหมาะกับ text classification
- **Random Forest = Bagging + Feature Subsampling** ลด variance ดีกว่า DT เดี่ยว feature importance ช่วยตีความ
- **Confusion Matrix จำ:** TP=ทำนายถูก-บวก, FP=ทำนายผิด-บวก (False Alarm), FN=พลาด positive (Missed)
- ในทางการแพทย์ FN อันตรายกว่า FP $\Rightarrow$ เน้น Recall; ในการกรอง spam เน้น Precision

RECAP 2: Regression (Lecture 7)

- **MSE ลงโทษ outlier มากกว่า MAE** เพราะยกกำลังสอง; RMSE หน่วยเดียวกับ y
- $R^2 = 1$ = perfect fit; $R^2 = 0$ = ไม่ดีกว่าการใช้ค่าเฉลี่ย; $R^2 < 0$ = แย่กว่า
- **Gradient Descent:** $\theta := \theta - \alpha \cdot \nabla L$ อัปเดต parameter ลดทาง gradient
- **Batch GD:** ใช้ทุก sample/step (stable แต่ช้า) vs **SGD:** 1 sample/step (เร็วแต่ noisy) vs **Mini-batch:** กลาง ๆ
- **Ridge (L2):** ลด magnitude weights, **Lasso (L1):** ทำให้บาง weight = 0 (feature selection)

RECAP 3: Parameter Tuning (Lecture 8)

- **Hyperparameter vs Learnable:** max_depth, LR = Hyperparameter (กำหนดก่อน); weights, bias = Learnable (model เรียนรู้)
- **Overfitting:** Train high, Test low $\Rightarrow$ แก้ด้วย: regularization, ลด complexity, เพิ่มข้อมูล, k-Fold CV
- **k-Fold:** ทั่วไป; **Stratified k-Fold:** imbalanced class; **LOO-CV:** dataset เล็ก; **Time Series CV:** sequential
- **Grid Search:** ลอง combination ทั้งหมด (exhaustive แต่ช้า); **Random Search:** สุ่ม (เร็วกว่า); **Bayesian:** ใช้ prior ชี้นำ
- **Adam:** รวม Momentum + Adaptive LR ใช้ได้ดีกับส่วนใหญ่ ออกจาก local minimum ได้ดี
- **Learning Rate Scheduling:** ลด LR ตามเวลา (decay) ช่วย converge ได้ละเอียดขึ้น

RECAP 4: Clustering (Lecture 9)

- **K-Means:** (1) init centroids, (2) assign points, (3) update centroids, (4) repeat จน converge WCSS วัดความแน่น
- **Elbow Method:** plot WCSS vs k หา “ข้อศอก” ที่ WCSS ลดช้าลงกะทันหัน
- **Silhouette:** $[-1, +1]$; $s \approx 1$ = ดี; $s \approx 0$ = on boundary; $s < 0$ = misclustered

- **Hierarchical Agglomerative:** Bottom-up, Single/Complete/Average Linkage, ผล = Dendrogram ตัด Dendrogram ที่ระดับต่าง ๆ ได้จำนวน cluster ต่าง ๆ
- **Complete Linkage = MAX distance** ระหว่างคู่จุดของ 2 clusters; Single = MIN
- **DBSCAN:** ไม่ต้องกำหนด k, จัดการ noise ได้, cluster ไม่เป็น convex ได้ ε = รัศมี, MinPts = จุดขั้นต่ำ
- **K-Means ดีเมื่อ:** cluster กลม ๆ, รู้จำนวน k; **DBSCAN ดีเมื่อ:** cluster รูปร่างซับซ้อน, มี noise/outlier