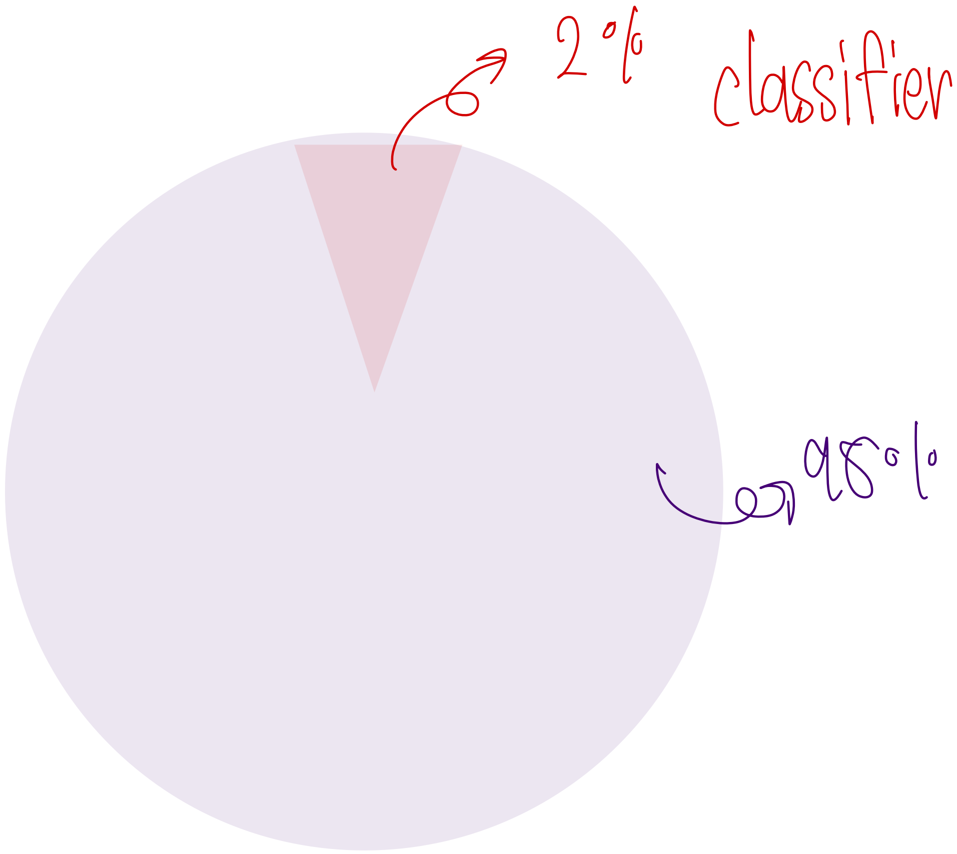
Chapter 8. Classification: Basic Concepts

- Classification: Basic Concepts
- Decision Tree Induction
- Bayes Classification Methods
- Linear Classifier
- Model Evaluation and Selection
- Techniques to Improve Classification Accuracy: Ensemble Methods
- Additional Concepts on Classification
- Summary

Model Evaluation and Selection

- Evaluation metrics
  - How can we measure accuracy?
  - Other metrics to consider?
- Use **validation test set** of class-labeled tuples instead of training set when assessing accuracy
- Methods for estimating a classifier’s accuracy
  - Holdout method
  - Cross-validation
  - Bootstrap
- Comparing classifiers:
  - ROC Curves

classifier → ไม่เก่ง (ความแม่นยำ 98%)



Classifier Evaluation Metrics: Confusion Matrix

- **Confusion Matrix:** คำตอบที่แท้จริง

Precision

เป็น yes ทาย yes

เป็น yes ทาย no

Actual class\Predicted class	C$_1$ ทายถูก	¬ C$_1$ ทายผิด
C$_1$	**True Positives (TP)**	**False Negatives (FN)**
¬ C$_1$	**False Positives (FP)**	**True Negatives (TN)**

Recall

เป็น no ทาย yes

เป็น no ทาย no

- In a confusion matrix w. *m* classes, *CM*$_{i,j}$ indicates # of tuples in class *i* that were labeled by the classifier as class *j*
  - May have extra rows/columns to provide totals
- **Example of Confusion Matrix:**

post

neg

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes positive	**6954**	**46**	7000
buy_computer = no negative	**412**	**2588**	3000
Total	7366	2634	10000

Classifier Evaluation Metrics: Accuracy, Error Rate, Sensitivity and Specificity

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P’	**N’**	**All**

- **Classifier accuracy,** or recognition rate
  - Percentage of test set tuples that are correctly classified

  Accuracy = (TP + TN)/All

- **Error rate:** *1 – accuracy*, or

  Error rate = (FP + FN)/All

- **Class imbalance problem**
  - One class may be *rare*
    - E.g., fraud, or HIV-positive
  - Significant *majority of the negative class* and minority of the positive class
  - Measures handle the class imbalance problem
    - **Sensitivity** (recall): True positive recognition rate
      - **Sensitivity = TP/P**
    - **Specificity:** True negative recognition rate
      - **Specificity = TN/N**

Classifier Evaluation Metrics: Precision and Recall, and F-measures

- **Precision**: Exactness: what % of tuples that the classifier labeled as positive are actually positive?

$$P = \text{Precision} = \frac{TP}{TP + FP}$$

ตัวที่ model ทายว่าเป็น Pos ถูกต้องมากน้อยแค่ไหน

- **Recall:** Completeness: what % of positive tuples did the classifier label as positive?

$$R = \text{Recall} = \frac{TP}{TP + FN}$$

ตัวที่เป็น Pos จริงๆ ทายถูกน้อยแค่ไหน
มากน้อยแค่ไหน

  - Range: [0, 1]
- The "inverse" relationship between precision & recall
- ***F* measure (or *F*-score):** harmonic mean of precision and recall
  - In general, it is the weighted measure of precision & recall

$$F_\beta = \frac{1}{\alpha \cdot \frac{1}{P} + (1 - \alpha) \cdot \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Assigning β times as much weight to recall as to precision)

  - ***F1-measure (balanced F-measure)***
    - That is, when β = 1,

$$F_1 = \frac{2PR}{P + R}$$

P สูง ยิ่งดี
R ต่ำ ไม่ดี } ต้องให้บาลานซ์กัน