นางสาวสุภาภรณ์ บุตุธรรม
623020543-8

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

Classifier Evaluation Metrics: Confusion Matrix

- **Confusion Matrix:**

Class ที่แท้จริง

Precision

Actual class\Predicted class	C_1	$\neg C_1$	
C_1	**True Positives (TP)**	**False Negatives (FN)**	Recall
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**	

- 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:**

Actual class\Predicted class	buy_computer = yes (Positive)	buy_computer = no (Negative)	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

    ความแม่นยำ

    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?

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

ตัวที่ model ทายว่าเป็น Positive ถูกต้องมากแค่ไหน

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

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

Model เราหาตัวที่เป็น Positive จริงๆ เจอมากน้อยแค่ไหน

  - 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

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

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

- ***F1-measure*** (*balanced F-measure)*

  P = Precision

  R = Recall

  - That is, when β = 1, ยิ่งสูงยิ่งดี ต่ำไม่ค่อยดี

$$\mathrm{F}_1 = \frac{2\mathrm{PR}}{\mathrm{P}+\mathrm{R}}$$

P สูง R ต่ำ = F ต่ำ ต้องให้สมดุลกัน

นางสาวสุภาภรณ์ บุตุธรรม
623020543-8

Chapter 9. Classification: Advanced Methods

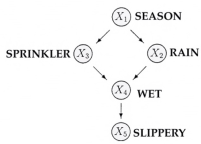
หาวิธีเพื่อสอนคอมพิวเตอร์ให้สามารถตัดสินใจแทนตัวเราได้

- ❑ Bayesian Belief Networks
- ❑ Support Vector Machines
- ❑ เรียนแบบเซลล์ประสาทของคน Neural Networks and Deep Learning
- ❑ Pattern-Based Classification
- ❑ Lazy Learners and K-Nearest Neighbors
- ❑ Other Classification Methods
- ❑ Summary

ของเทียมArtification + เลียนแบบเซลล์ประสาทของคน

From Naïve Bayes to Bayesian Networks

- ❑ Naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable
  - ❑ This assumption is often too simple to model the real world well
- ❑ Bayesian network (or Bayes network, belief network, Bayesian model or probabilistic directed acyclic graphical model) is a probabilistic **graphical model**
  - ❑ Represented by a set of *random variables* and *their conditional dependencies* via a *directed acyclic graph* (DAG)
  - ❑ Ex. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases



Discussion on the *k*-NN Algorithm

มี 2 ตัวสำคัญ คือ
1. K
2. หาระยะห่าง

- ❑ *k*-NN for real-valued prediction for a given unknown tuple
  - ❑ Returns the mean values of the *k* nearest neighbors ถามคนที่อยู่ใกล้เคียง เป็นจำนวนคี่เพื่อเอกฉันท์
- ❑ Distance-weighted nearest neighbor algorithm
  - ❑ Weight the contribution of each of the *k* neighbors according to their distance to the query x_q
    - ❑ เชื่อคนที่อยู่ใกล้เรามากที่สุด Give greater weight to closer neighbors

$$w \equiv \frac{1}{d(x_q, x_i)^2}$$

= 1 / ระยะทาง (ยิ่งใกล้ยิ่งมาก & ยิ่งไกลยิ่งน้อย)

- ❑ Robust to noisy data by averaging ถามคนที่ใกล้ก่อน *k*-nearest neighbors จำนวนเพื่อนบ้านที่จะถาม
- ❑ Curse of dimensionality: distance between neighbors could be dominated by irrelevant attributes
  - ❑ To overcome it, axes stretch or elimination of the least relevant attributes