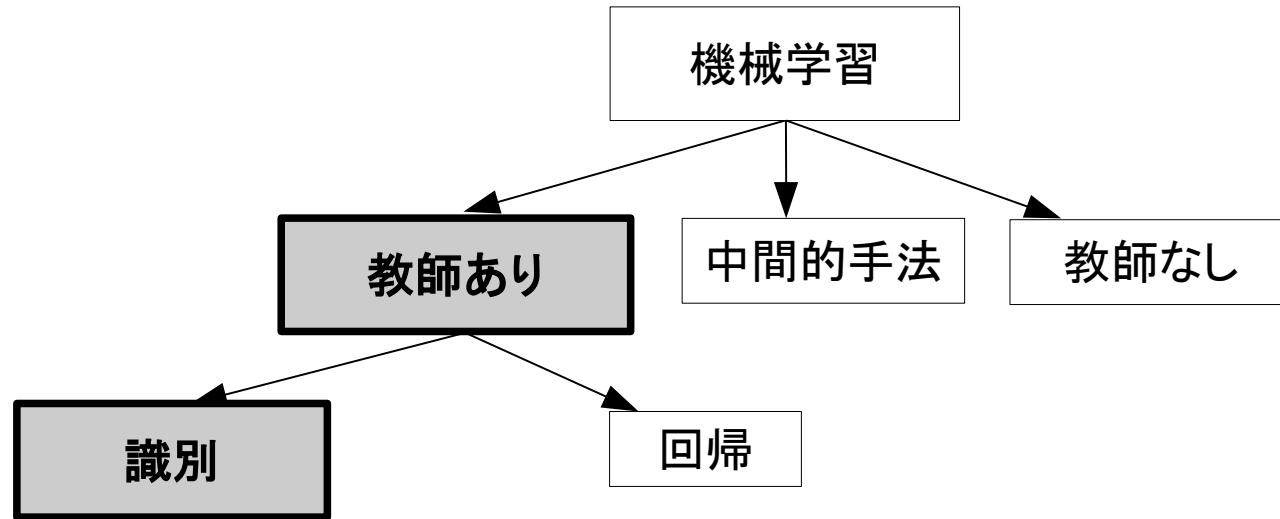


4. 識別 —統計的手法—



- ラベル特徴



- 数値特徴

4.1 統計的識別とは

- 最大事後確率則による識別

\mathbf{x} : 特徴ベクトル

$$C_{MAP} = \arg \max_i P(\omega_i | \mathbf{x}) \quad \omega_i \ (1 \leq i \leq c) : \text{クラス}$$

- データから直接的にこの確率を求めるのは難しい

- ベイズの定理 $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

$$\begin{aligned} C_{MAP} &= \arg \max_i P(\omega_i | \mathbf{x}) \\ &= \arg \max_i \frac{P(\mathbf{x} | \omega_i) P(\omega_i)}{P(\mathbf{x})} \\ &= \arg \max_i P(\mathbf{x} | \omega_i) P(\omega_i) \end{aligned}$$

4.1 統計的識別とは

- 事前確率

$$P(\omega_i)$$

- 特徴ベクトルを観測する前の、各クラスの起こりやすさ

- 事前確率の最尤推定

$$P(\omega_i) = \frac{n_i}{N}$$

N: 全データ数、 n_i : クラス i のデータ数

4.1 統計的識別とは

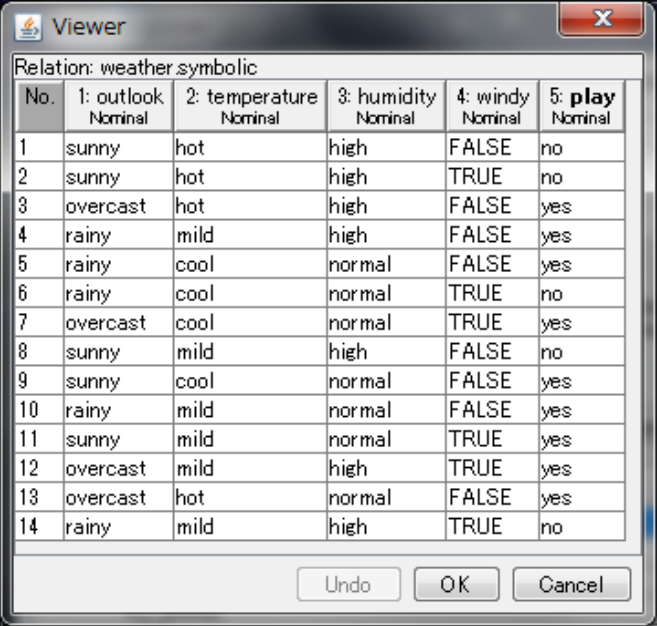
- 尤度

$$P(\boldsymbol{x}|\omega_i)$$

- 特定のクラスから、ある特徴ベクトルが出現する尤もらしさ

- d次元ベクトルの場合の最尤推定

- 値の組合せがデータ中に出現しないものの多数



| No. | 1: outlook Nominal | 2: temperature Nominal | 3: humidity Nominal | 4: windy Nominal | 5: play Nominal |
|-----|-----------------------|---------------------------|------------------------|---------------------|--------------------|
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

Weka の
weather.nominal データ
3×3×2×2=36 種類の組合せ

4.2.2 ナイーブベイス識別

- ナイーブベイズの近似
 - 全ての特徴が独立であると仮定

$$P(\boldsymbol{x}|\omega_i) = P(x_1, \dots, x_d|\omega_i)$$

$$= \prod_{j=1}^d P(x_j|\omega_i)$$

$$C_{NB} = \arg \max_i P(\omega_i) \prod_{j=1}^d P(x_j|\omega_i)$$

4.2.2 ナイーブベイス識別

- 尤度の最尤推定

$$P(x_j|\omega_i) = \frac{n_{ij}}{n_i}$$

n_{ij} : クラス i のデータのうち、
 j 次元目の値が x_j の個数

ゼロ頻度問題

- 確率の m 推定

$$P(x_j|\omega_i) = \frac{n_{ij} + mp}{n_i + m}$$

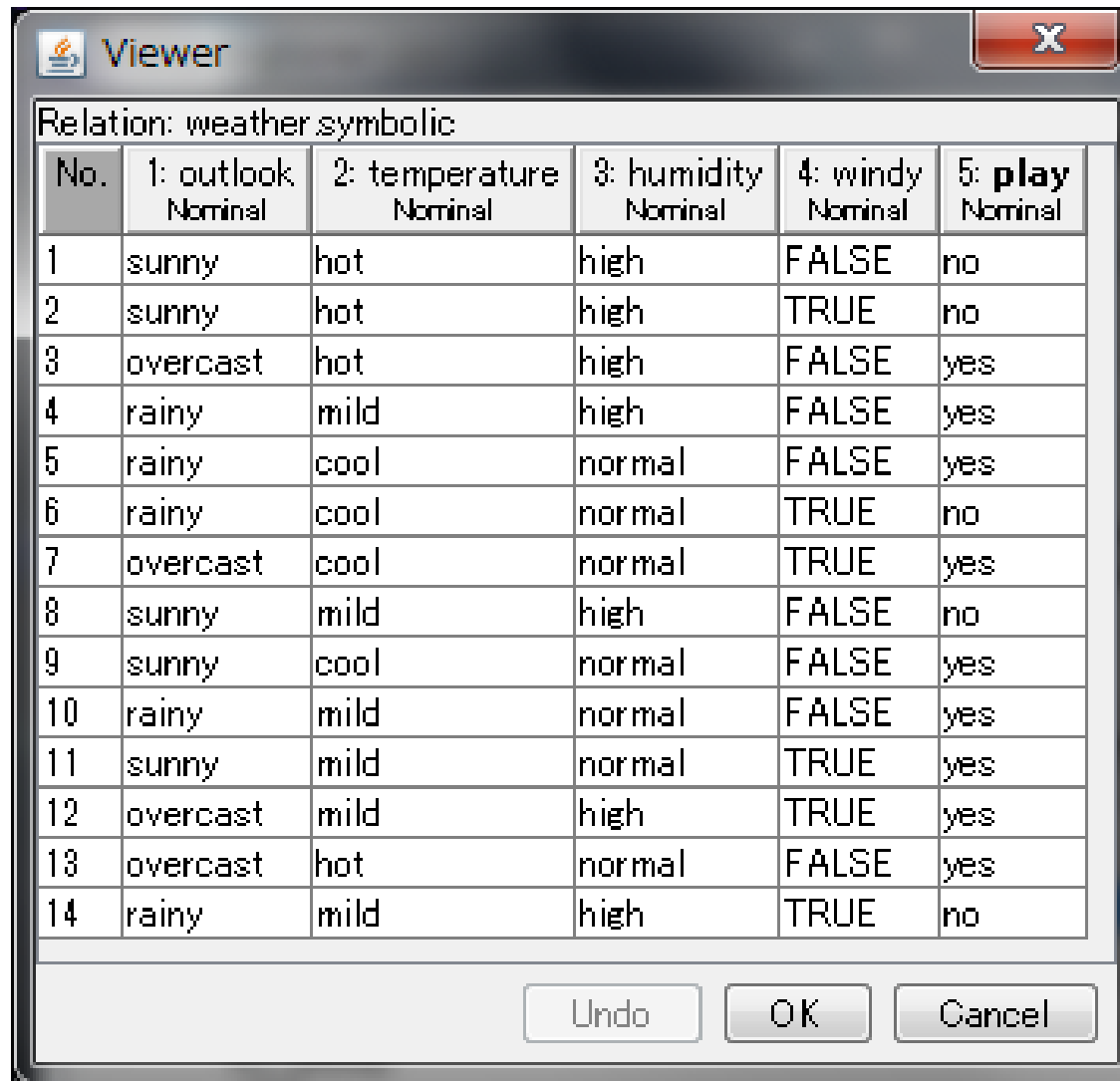
p : 事前に見積もった各特徴値の割合
 m : 事前に用意する標本数

- ラプラス推定

– m : 特徴値の種類数、 p : 等確率 とすると、 $mp=1$

4.2.2 ナイーブベイス識別

weather.nominal データ



Relation: weather.symbolic

| No. | 1: outlook Nominal | 2: temperature Nominal | 3: humidity Nominal | 4: windy Nominal | 5: play Nominal |
|-----|-----------------------|---------------------------|------------------------|---------------------|---------------------------|
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

Undo OK Cancel

4.2.2 ナイーブベイス識別

Classifier
Choose NaiveBayes

Test options
☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) play
Start Stop

Result list (right-click for options)
12:44:23 - trees.J48
13:10:03 - bayes.NaiveBayes

Classifier output

| Attribute | Class | |
|-------------|---------------|--------------|
| | yes (0.63) | no (0.38) |
| ===== | | |
| outlook | | |
| sunny | 3.0 | 4.0 |
| overcast | 5.0 | 1.0 |
| rainy | 4.0 | 3.0 |
| [total] | 12.0 | 8.0 |
| temperature | | |
| hot | 3.0 | 3.0 |
| mild | 5.0 | 3.0 |
| cool | 4.0 | 2.0 |
| [total] | 12.0 | 8.0 |
| humidity | | |
| high | 4.0 | 5.0 |
| normal | 7.0 | 2.0 |
| [total] | 11.0 | 7.0 |
| windy | | |
| TRUE | 4.0 | 4.0 |
| FALSE | 7.0 | 3.0 |
| [total] | 11.0 | 7.0 |

Time taken to build model: 0 seconds
=== Evaluation on training set ===

Class

| Attribute | yes (0.63) | no (0.38) |
|-----------|---------------|--------------|
| ===== | | |
| outlook | | |
| sunny | 3.0 | 4.0 |
| overcast | 5.0 | 1.0 |
| rainy | 4.0 | 3.0 |
| [total] | 12.0 | 8.0 |

ラプラス推定
各カウントの初期値が 1
例) no は 5 事例で、
overcast の事例は
データ中にない

Status OK Log x 0

実行例

- 入力

天気 気温 湿度 風
 $\mathbf{x}=(\text{sunny}, \text{hot}, \text{high}, \text{false})$

$$P(\text{yes}) = 0.63$$

$$P(\text{no}) = 0.38$$

$$P(\mathbf{x}|\text{yes}) = 3/12 \times 3/12 \times 4/11 \times 7/11 = 0.0144$$

$$P(\mathbf{x}|\text{no}) = 4/8 \times 3/8 \times 5/7 \times 3/7 = 0.0574$$

$$P(\mathbf{x}|\text{yes}) \cdot P(\text{yes}) = 0.0091 < P(\mathbf{x}|\text{no}) \cdot P(\text{no}) = 0.0218$$

- 出力

no

4.3 ベイジアンネットワーク

- ベイジアンネットワークの仮定
 - 変数の部分集合が、ある分類値のもとで独立である

$$P(x_1, \dots, x_d) = \prod_{i=1}^d P(x_i | Parents(X_i))$$

