CS第2　テーマ2　演習ガイド

これからの予定

1. レポート課題2の説明
2. グループごとに課題に挑戦

1. 準備

0. グループごとに分かれる
1. ログインする.
2. Terminal を動かす（TSUBAME と直接対話する窓口）.
   - 2.1. mkdir cs2kadai2　← レポート課題2用の部屋を作る
   - 2.2. cd cs2kadai2　← レポート課題 2用の部屋へ行く.
   - 2.3. 必要なファイルを共通のお部屋から cs2kadai2 へコピーする.

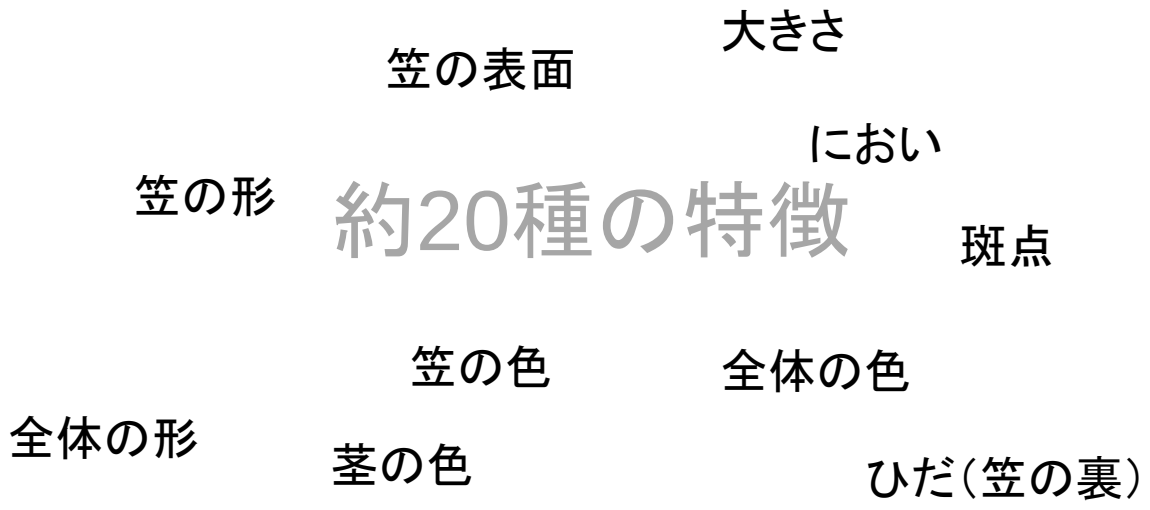
共通ファイルの置き場所：Desktop/shared/CS/cs5a/data1

2. レポート課題2：導入

データマイニングとは
法則を発見することない

毒キノコの法則を発見せよ

キノコの特徴（属性 attribute）から，与えられたキノコが毒キノコか否かを判定するルール（判別規則 binary decision rule）を見つけよう



2. レポート課題2：導入

データマイニングとは
法則を発見することない

毒・無毒

p: poisonous
e: edible

毒キノコの法則を発見せよ

笠の色

y: yellow
p: pink
b: brown
・・・

19 個の属性

4000 種類のキノコ

p	f	y	n	f	y	f	c	n	b	t	?	s	s	w	w	p	w	o	e	w	v	p
p	x	s	b	t	f	f	c	b	h	t	b	s	s	w	w	p	w	o	p	h	s	u
e	x	y	n	t	a	f	c	b	w	e	r	s	y	w	w	p	w	o	p	k	y	p
e	x	y	n	t	n	f	c	b	w	t	b	s	s	g	g	p	w	o	p	k	v	d
e	k	s	n	f	n	a	c	b	y	e	?	s	s	o	o	p	o	o	p	n	c	l
p	k	y	e	f	s	f	c	n	b	t	?	k	s	w	p	p	w	o	e	w	v	p
e	b	s	w	f	n	f	w	b	g	e	?	k	k	w	w	p	w	t	p	w	s	g
p	f	s	w	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	v	g
p	f	f	y	f	f	f	c	b	h	e	b	k	k	b	n	p	w	o	l	h	y	d
e	f	f	n	t	n	f	c	b	n	t	b	s	s	w	g	p	w	o	p	n	y	d
e	f	f	g	f	n	f	w	b	n	t	e	f	f	w	w	p	w	o	e	k	a	g

データのみ
を使って！

UC Irvine ML Repository, 1982

2. レポート課題2：導入

データマイニングとは
法則を発見することない

少し簡単にしました

毒・無毒

+1: poisonous
−1: edible

毒キノコの法則を発見せよ

19 個の属性

0, 1, 2 の
三種類

4000 種類のキノコ

−1	0	0	1	2	2	2	2	1	0	2	2	2	0	2	1	1	1	0	1
+1	0	2	0	1	1	2	1	1	0	2	2	0	0	2	1	2	0	1	0
+1	0	0	2	2	1	2	1	1	0	1	1	1	0	0	1	0	0	2	2
−1	0	0	1	1	2	2	1	1	0	2	2	2	0	2	1	2	0	1	2
+1	2	2	0	2	0	2	1	2	2	2	2	2	0	2	1	1	0	1	2
+1	1	0	0	2	1	2	1	1	0	1	1	1	0	0	1	0	0	2	2
−1	1	0	0	1	0	2	2	2	1	2	2	2	0	2	1	2	0	1	2
−1	2	2	1	2	0	2	1	2	2	2	1	2	1	1	1	1	0	1	0
+1	2	2	1	2	1	2	1	2	2	2	2	1	0	2	1	1	0	1	0
−1	0	1	0	1	0	2	1	1	0	1	2	2	0	2	1	2	0	0	1
+1	1	2	2	1	1	2	1	1	0	2	2	0	0	2	1	2	0	0	1
⋮																			

UC Irvine ML Repository, 1982

2. レポート課題2：導入

データマイニングとは
法則を発見することない

何を作る？
what?

毒キノコの法則を発見せよ

毒・無毒
+1: poisonous
−1: edible

19 個の属性

0, 1, 2 の三種類

4000 種類のキノコ

−1	0	0	1	2	2	2	2	1	0	2	2	2	0	2	1	1	1	0	1
+1	0	2	0	1	1	2	1	1	0	2	2	0	0	2	1	2	0	1	0
+1	0	0	2	2	1	2	1	1	0	1	1	1	0	0	1	0	0	2	2
−1	0	0	1	1	2	2	1	1	0	2	2	2	0	2	1	2	0	1	2
+1	2	2	0	2	0	2	1	2	2	2	2	2	0	2	1	1	0	1	2
+1	1	0	0	2	1	2	1	1	0	1	1	1	0	0	1	0	0	2	2
−1	1	0	0	1	0	2	2	2	1	2	2	2	0	2	1	2	0	1	2
−1	2	2	1	2	0	2	1	2	2	2	1	2	1	1	1	1	0	1	0
+1	2	2	1	2	1	2	1	2	2	2	2	1	0	2	1	1	0	1	0
−1	0	1	0	1	0	2	1	1	0	1	2	2	0	2	1	2	0	0	1
+1	1	2	2	1	1	2	1	1	0	2	2	0	0	2	1	2	0	0	1

UC Irvine ML Repository, 1982

法則って何？

属性値を使って毒性の
有無±1を判定する条件式

属性 5 == 2

属性5 の値が 2　⇒ 毒（つまり，+1）
そうでない　⇒ 無毒（つまり，-1）

a_0 a_4 ↓ 一般の属性 a_j

毒性	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
−1	0	0	1	2	2	2	2	1	0	2	2	2	0	2	1	1	1	0	1
+1	0	2	0	1	1	2	1	1	0	2	2	0	0	2	1	2	0	1	0
+1	0	0	2	2	1	2	1	1	0	1	1	1	0	0	1	0	0	2	2
−1	0	0	1	1	2	2	1	1	0	2	2	2	0	2	1	2	0	1	2

注）簡単のため毒性を 0 番目の属性 a_0 とする.

一般に条件式とは

a_j == k もしくは a_j != k を
AND, OR, NOT でつなげた論理式

⇒ 宿題

結局欲しいのは
このように
使える条件式 ⟶

if 条件式（が成立）
　　a_0 は +1 と判定
else
　　a_0 は -1 と判定

2. レポート課題2：導入

データマイニングとは 法則を発見することない

何を作る？
what?

毒キノコの法則を発見せよ

毒性判定の条件式

どうやって作る？
how?

毒・無毒
+1: poisonous
−1: edible

19 個の属性

0, 1, 2 の三種類

4000 種類のキノコ

−1	0	0	1	2	2	2	2	1	0	2	2	2	0	2	1	1	1	0	1
+1	0	2	0	1	1	2	1	1	0	2	2	0	0	2	1	2	0	1	0
+1	0	0	2	2	1	2	1	1	0	1	1	1	0	0	1	0	0	2	2
−1	0	0	1	1	2	2	1	1	0	2	2	2	0	2	1	2	0	1	2
+1	2	2	0	2	0	2	1	2	2	2	2	2	0	2	1	1	0	1	2
+1	1	0	0	2	1	2	1	1	0	1	1	1	0	0	1	0	0	2	2
−1	1	0	0	1	0	2	2	2	1	2	2	2	0	2	1	2	0	1	2
−1	2	2	1	2	0	2	1	2	2	2	1	2	1	1	1	1	0	1	0
+1	2	2	1	2	1	2	1	2	2	2	2	1	0	2	1	1	0	1	0
−1	0	1	0	1	0	2	1	1	0	1	2	2	0	2	1	2	0	0	1
+1	1	2	2	1	1	2	1	1	0	2	2	0	0	2	1	2	0	0	1

UC Irvine ML Repository, 1982

1. 専門家と考える

専門家とデータをながめて整理する

← データ解析の鉄則

専門家がいない
専門家の見落としを発見したい

いろいろな道具がある
でも. . .

2. ランダムサンプリングして傾向を見る

基本は数えること!!

2. レポート課題2：導入

頻度を調べる

$a_4 == 2 \;\&\&\; a_0 == +1$
$a_4 == 2 \;\&\&\; a_0 == -1$

前にもやったぞ！

毒性	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	1
−1	0	0	1	2	2	2	2	1	0	2	2	2	0	2	1	1	1	(
+1	0	2	0	1	1	2	1	1	0	2	2	0	0	2	1	2	0	1	0
+1	0	0	2	2	1	2										0	0	2	2
−1	0	0	1	1	2	2										2	0	1	2
+1	2	2	0	2	0	2										1	0	1	2
+1	1	0	0	2	1	2										0	0	2	2
−1	1	0	0	1	0	2										2	0	1	2
−1	2	2	1	2	0	2										1	0	1	0
+1	2	2	1	2	1	2	1	2	2	2	2	1	0	2	1	1	0	1	0
−1	0	1	0	1	0	2	1	1	0	1	2	2	0	2	1	2	0	0	1
+1	1	2	2	1	1	2	1	1	0	2	2	0	0	2	1	2	0	0	1
−1	0	2	1	2	2	1	1	1	1	1	2	2	0	0	1	2	2	0	0
−1	0	1	1	1	2	2	1	1	0	2	2	2	0	0	1	2	0	2	2
+1	0	1	0	2	1	2	1	1	0	1	1	1	0	1	1	0	0	2	0
+1	0	0	0	2	1	2	1	1	0	1	1	1	0	1	1	0	0	2	1
+1	0	1	1	2	0	2	1	2	2	2	1	1	1	1	1	1	0	1	0

$a_4 = 2$ の時，ほとんどが+1（毒性有り）ならば，それをルールに取り入れるのがよいのでは！？

⋮

3. レポート課題2

やるべきこと

1. 毒キノコを判定する条件式を求める

   - ランダムにサンプルした**サンプルデータ**を解析
     ⇒ 条件式を導き出す
   - その条件式の精度を**テストデータ**で確認する

提出物と採点基準（満点 15）

1. 条件式　※グループ共通でよい
2. その条件式の精度　※これもグループ共通でよい

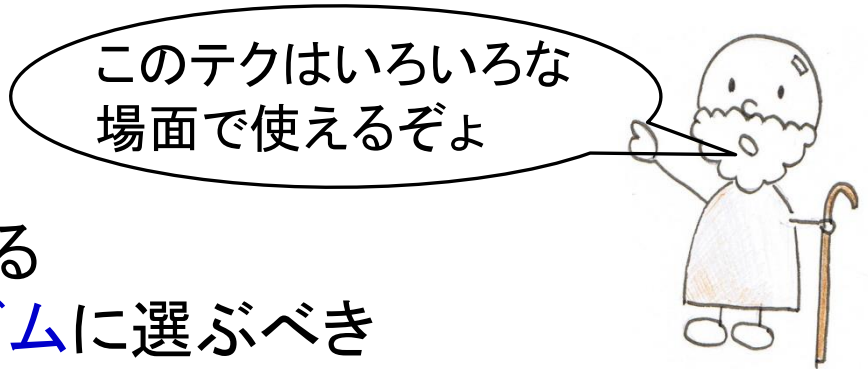
} 合わせて 5 点

3. その条件式を導いた方法についての説明 (10)
   ※もちろんグループの構成員が同じ方法を使って求めるのは当然．でも，**説明は自分の言葉で書くこと**

以下はオプショナル（加点 ≦ 5）

4. 自分なりの解析 ⇒ 自分なりの条件式

3. レポート課題2：アドバイス



(1) ランダムサンプルの作り方

- 元データは偏っている場合もある
- 条件式作成用のデータはランダムに選ぶべき
- 残りはテスト用に取っておこう

1つの有効な方法（1つのデータが1行になっている場合）

1. データの各行の先頭に (0,1) 区間の乱数を付ける
   ruby addrnd.rb < data4000org.txt > tmp.txt
2. tmp.txt をエクセルで見て先頭の行でソート
   - エクセルでセミコロン ; 区切りで開く
   - 最初の行（A 欄）でソートし，最初の行を削除
   - テキスト（タブ区切り）で保存（data4000.txt などの名前で）
3. 最初の n 行をランダムサンプルとして取り出す
   head -500 < data4000.txt > data500.txt ← 行数は適宜
4. 最後の行に 0 を付けておく ← 今回は行数がわからないので最後の行の印が必要
   echo 0 >> data500.txt

3. レポート課題2：アドバイス

(2) データ解析のためのツール

・要するに数勘定の道具
・このくらいは，エクセル上でも可能
・もちろん，本職はもっと高級な道具を使います．でも，この程度でも結構できる

1. 条件式の精度テスト用プログラム

プログラム名：test.rb
使い方：・hantei = の右辺に条件式を書く
・ruby test.rb < data500.txt

↑ファイルの最後に 0 のみの行が必要
※データの行数が不定なので

2. 属性 $a_j == k$ と毒性との関係を調べるプログラム

プログラム名：count.rb
使い方：・調べたい属性と属性値でプログラムを修正
・ruby count.rb < data500.txt

まとめ Terminal 上のコマンド

命令	使用例	意味
mkdir	mkdir kadai2	kadai2 というフォルダ（部屋）を作る
cd	cd kadai2	kadai2 というお部屋に入る
	cd ..	上の（大きな）部屋に戻る
ls	ls	その部屋にあるファイルを表示する
cat	cat xxx.txt	ファイル xxx.txt の中身を見る（全部）
less	less xxx.txt	ファイル xxx.txt の中身を見る（部分的） 注）スペースキーで先，やめるのは q
head	head $-n$ xxx.txt	最初の n 行だけ表示
tail	tail $-n$ xxx.txt	最後の n 行だけ表示
rm	rm foo.rb	foo.rb を消す（戻らないので注意）
リダイレクト <	ruby xx.rb < aa	xx.rb を実行．入力は aa から取り込む
リダイレクト >	ruby xx.rb > bb	xx.rb を実行．結果は bb へ出す
echo + >>	echo 文字列 >> aa	文字列を aa の最後の行に付ける