Методы классификации. Решающие деревья. Случайный лес

Н.П. Алексеева,

СПбГУ, математико-механический факультет

2020 г.

Характеристики степени разнообразия

Информационное разнообразие

Пусть $x_1, \ldots, x_n$ выборка наблюдений дискретной величины с m градациями, каждая из которых встречается a_i раз, $a_1 + a_2 + \ldots + a_m = n$. Информационное разнообразие этой выборки измеряется формулой:

$$I(x_1, \ldots, x_n) = n \ln n - \sum_{i=1}^{m} a_i \ln a_i.$$

В группе с одинаковыми значениями признака информационное разнообразие равно нулю, так как $I = n \ln n - 0 \cdot \ln 0 - n \ln n = 0$. Очевидно, что разнообразие группы, состоящего из одного индивида, также равно нулю.

Энтропия

Энтропия случайного эксперимента

$$\xi : \begin{pmatrix} z_1 & \dots & z_m \\ p_1 & \dots & p_m \end{pmatrix},$$

где через $z_1, \dots, z_n$ обозначены возможные исходы эксперимента, которые осуществляются с вероятностями $p_1, \dots, p_m$, $\sum\limits_{k=1}^{m} p_k = 1$, вычисляется по формуле Шеннона:

$$H(\xi) = \sum_{k=1}^{m} p_k \log_2 \frac{1}{p_k}. \qquad (1)$$

Информационное разнообразие выборки $x_1, \ldots, x_n$ равно $I(x_1, \ldots, x_n) = nH(x_1, \ldots, x_n)/\ln 2$, где $H(x_1, \ldots, x_n)$ энтропия выборочного распределения.

Итак, $a_i = \#\{x_j : x_j = z_i\}$ частота элемента z_i,
$a_1 + \ldots + a_m = n$, энтропия выборочного распределения

$$\begin{pmatrix} z_1 & \ldots & z_m \\ \frac{a_1}{n} & \ldots & \frac{a_m}{n} \end{pmatrix}$$

имеет вид $H(x_1, \ldots, x_n) =$

$$= -\sum_{i=1}^{m} \frac{a_i}{n} \log_2 \frac{a_i}{n} = -\frac{1}{n} \sum_{i=1}^{m} a_i \log_2 a_i + \frac{1}{n} \sum_{i=1}^{m} a_i \log_2 n,$$

$$nH(x_1, \ldots, x_n) = n \log_2 n - \sum_{i=1}^{m} a_i \log_2 a_i$$

Индекс разнообразия Джини (Gini impurity)

$$\xi : \begin{pmatrix} z_1 & \dots & z_m \\ p_1 & \dots & p_m \end{pmatrix} \qquad J = 1 - \sum_{i=1}^{m} p_i^2$$

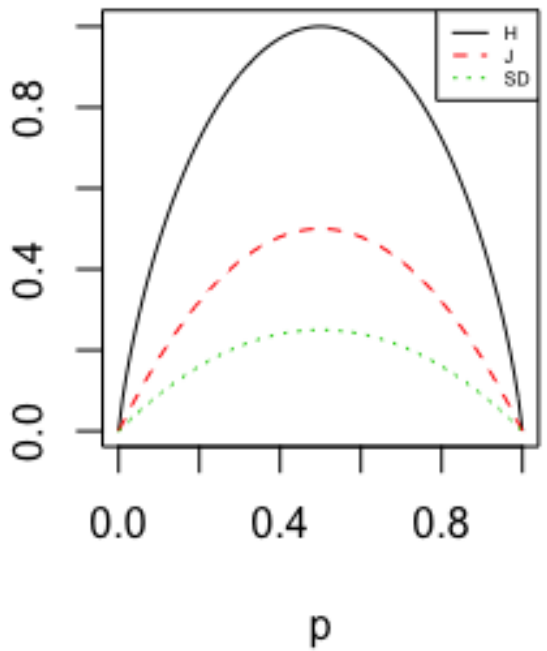


Figure: $m = 2$, $H(p) = -p \log_2 p - (1-p)\log_2(1-p)$, $J(p) = 1 - p^2 - (1-p)^2$, $SD(p) = p(1-p)$.

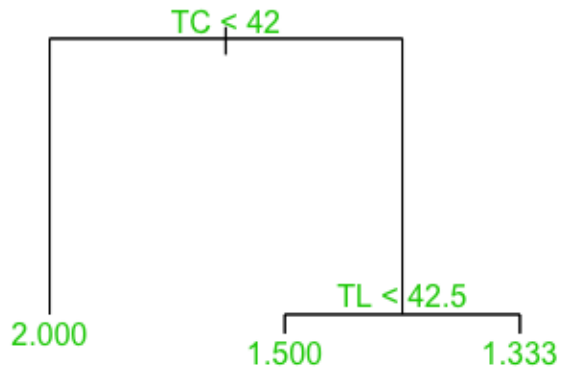
Построение дерева классификации: пример данных

	Moris	TC	TL	class
1	15	29	38	2
2	12	40	27	2
3	10	52	30	2
4	8	48	47	1
7	17	53	69	2
8	1	44	26	1
9	13	51	33	1
10	14	52	49	1
11	3	26	30	2
12	3	64	50	1
14	1	22	30	2
15	15	51	51	2
16	13	52	39	1
20	1	40	39	2
21	13	78	46	1
22	1	45	36	2
23	15	57	39	2

Функции в R для построения дерева

$$treeA < -tree(class \sim \ ., data)$$
$$plot(treeA); text(treeA, col = 3)$$

split	n	deviance	value
root	17	4.118	1.588
TC $<$ 42	5	0.000	2.000 *
TC $>$ 42	12	2.917	1.417
TL $<$ 42.5	6	1.500	1.500*
TL $>$ 42.5	6	1.333	1.333*



Критика алгоритма построения дерева в R

Информационный выигрыш от объединения

Пусть $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_m)$, тогда $\Delta = I(X, Y) - I(X) - I(Y)$.

- Чему равен Δ, если рядом с березкой и елочкой посадили березку и елочку?
- Чему равен Δ, если рядом с двумя березками посадили две елочки?

Применим Δ для поиска наилучшего разбиения множества на две части.

Код для вычисления Δ

```
I_inf <- function(X)
{nn <- table(X);  n <- sum(nn);  n * log(n) - sum(nn * log(nn))}
```

```
D_inf <- function(X, Y)I_inf(c(X, Y)) - I_inf(X) - I_inf(Y)
```

```
D_inf(X = c(1, 1), Y = c(2, 2))
```

```
2.7726
```

```
D_inf(X = c(1, 2), Y = c(1, 2)))
```

```
0
```

```
GinyDA. < −function(y, class, k1, k2)
{ #          y независимая переменная, class,
  #          k1,k2 минимальные объемы подвыборок
ss < −sort(unique(y)); ss
ss. < −(ss[−1] + ss[−length(ss)])/2;
#          определяем середины интервалов между наблюдениями
l < −lapply(ss., function(x)
{
   L1 < −class[y <= x]; L2 < −class[y > x];
   if(length(L1) > k1&length(L2) > k2)
   res < −c(D_inf(L1, L2), x)else res < −c(NA, NA)
})
# считаем информационные выигрыши при размежевании данных
data.frame(na.omit(do.call('rbind', l)))
}
```

Упорядочим значения признака TC

$22 < 26 < 29 < 40 < 44 < 45 < 48 < 51 < 52 < 53 < 57 < 64 < 78$

и рассмотрим середины интервалов между ними

$M = \{24, 27.5, 34.5, 42, 44.5, 46.5, 49.5, 51.5, 52.5, 55, 60.5, 71\}$

GinyDA.(df$Morris, class, 2, 2)
GinyDA.(df$TC, class, 2, 2)
GinyDA.(df$TL, class, 2, 2)

Morris	Δ	x
1	0.30	2.00
2	0.12	5.50
3	0.01	9.00
4	0.04	11.00
5	0.24	12.50
6	0.70	13.50
7	2.54	14.50

TC	Δ	x
1	1.81	34.50
2	3.37	42.00
3	1.23	44.50
4	1.92	46.50
5	0.84	49.50
6	0.63	51.50
7	0.08	52.50
8	0.48	55.00

TL	Δ	x
1	0.70	31.50
2	0.12	34.50
3	0.40	37.00
4	0.84	38.50
5	1.25	42.50
6	0.51	46.50
7	0.08	48.00
8	0.05	49.50

Разделение выборки на две части: A и B

A	Moris	TC	TL	class
1	15	29	38	2
2	12	40	27	2
11	3	26	30	2
14	1	22	30	2
20	1	40	39	2

B	Moris	TC	TL	class
3	10	52	30	2
4	8	48	47	1
7	17	53	69	2
8	1	44	26	1
9	13	51	33	1
10	14	52	49	1
12	3	64	50	1
15	15	51	51	2
16	13	52	39	1
21	13	78	46	1
22	1	45	36	2
23	15	57	39	2

Группа B

Morris	Δ	x
1	0.06	5.50
2	0.36	9.00
3	0.00	11.50
4	1.40	13.50
5	3.38	14.50

TC	Δ	x
1	0.06	49.50
2	0.00	51.50
3	0.09	52.50
4	0.06	55.00

TL	Δ	x
1	0.06	34.50
2	0.09	37.50
3	0.17	42.50
4	0.00	46.50
5	0.09	48.00
6	0.51	49.50

	Moris	TC	TL	class
3	10	52	30	2
4	8	48	47	1
8	1	44	26	1
9	13	51	33	1
10	14	52	49	1
12	3	64	50	1
16	13	52	39	1
21	13	78	46	1
22	1	45	36	2

	Moris	TC	TL	class
7	17	53	69	2
15	15	51	51	2
23	15	57	39	2

Если TC ≤ 42, то W_2 иначе если Morris > 14.5, то тоже W_2, и в случае, когда TC > 42 и Morris < 14.5 с вероятностью $\frac{7}{9}$ популяция W_1.

Сравнение со встроенными функциями

	Moris	TC	TL	class
3	10	52	30	2
8	1	44	26	1
9	13	51	33	1
16	13	52	39	1
22	1	45	36	2
23	15	57	39	2

	Moris	TC	TL	class
4	8	48	47	1
7	17	53	69	2
10	14	52	49	1
12	3	64	50	1
15	15	51	51	2
21	13	78	46	1

Вероятность правильной классификации меньше, чем при разделении по информационному выигрышу, но ветки дерева с большим числом листьев.

Случайный лес

```
library("randomForest")
data.rf <-randomForest(as.factor(class) ~ ., data = data)
table(data$class, predict(data.rf))
```

from \ to	1	2
1	5	2
2	5	5

Деревья по категориальным данным

Если группа индивидов характеризуется несколькими признаками то разнообразие складывается из информационных разнообразий по каждому признаку.

Процедура кластеризации индивидов

- глаза: 1 - большие, круглые, 0 - узкие, мелкие;
- нос: 1 - большой, толстый, 0 - тонкий, маленький;
- рот: 1 - большой, 0 - узкий;
- уши: 1 - большие, музыкальные, 0 - мелкие.

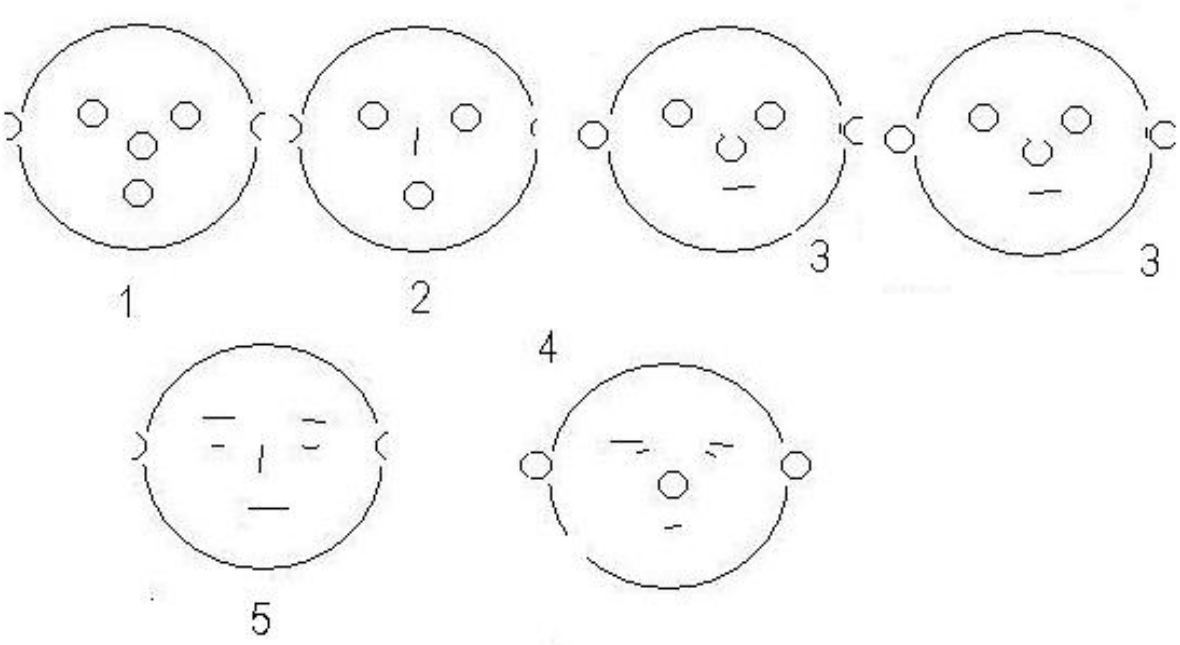


Figure: Данные для кластеризации.

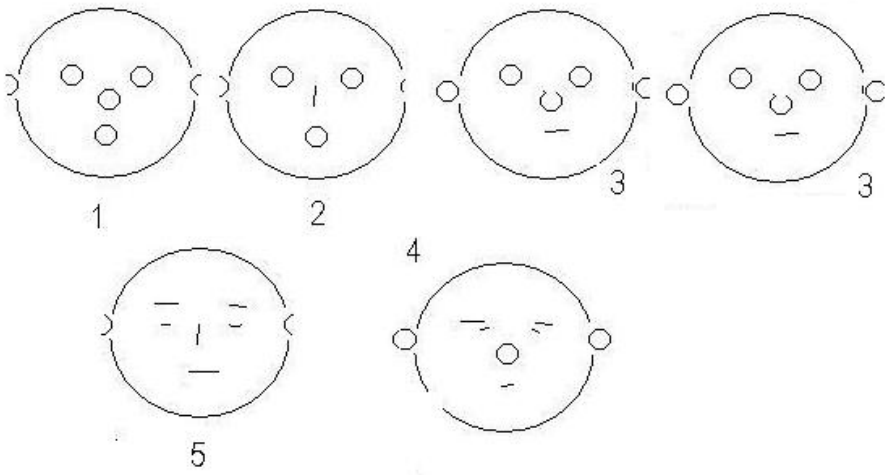
Количество признаков, по которым индивиды отличаются прпарно.

	1	2	3	4	5
1	–	1	2	3	3
2		–	3	4	2
3			–	1	3
4				–	2
5					–

$$I_{A(1,2)} = 2\ln 2 - 1\ln 1 - 1\ln 1 = 1.37,$$

$$I_{B(3,3,4)} = 3\ln 3 - 2\ln 2 - 1\ln 1 = 1.91,$$

$$I_C(5) = 0.$$



Вычисление информационного выигрыша

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{A+C}	1.91	1.91	1.91	0	5.73

$$\Delta I_{A+C} = I_{A+C} - I_A - I_C = 5.73 - 1.37 - 0 = 4.34\,.$$

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{B+C}	2.77	2.25	0	2.25	7.27

$$\Delta I_{B+C} = I_{B+C} - I_B - I_C = 7.27 - 1.91 - 0 = 5.36.$$

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
		I_{A+B}	2.5	2.5	3.37	3.37	11.73

$$\Delta I_{A+B} = I_{A+B} - I_A - I_B = 11.73 - 1.37 - 1.91 = 8.44,$$

Деревья по категориальным данным

Наименьший выигрыш достигается от объединения групп A(1, 2) и C(5). Поэтому получаем два кластера: (1, 2, 5) и (3,3,4), которые отличаются ушками.

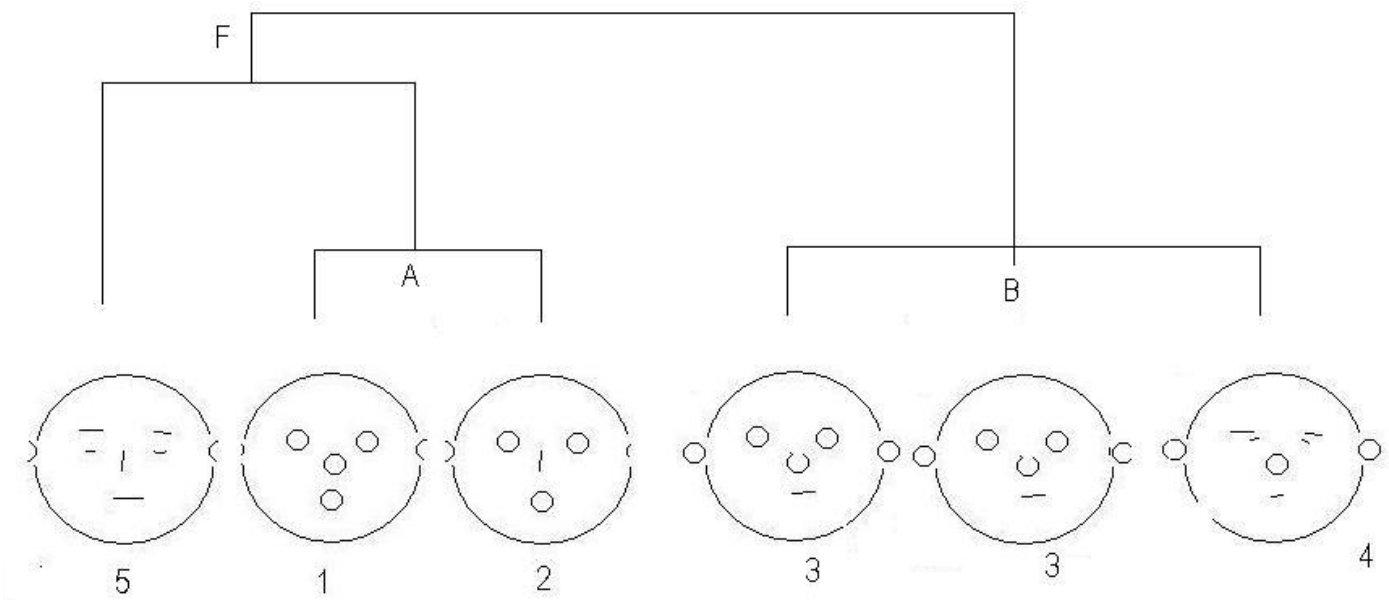


Figure: Дерево кластеризации.

Задание

Вычислить информационные выигрыши от объединения кластеров F(1, 2, 5) и B(3, 3, 4)

$$\Delta I_{F+D} = I_{F+D} - I_F - I_D, \qquad \Delta I_{B+D} = I_{B+D} - I_B - I_D,$$

и выяснить, к кому из них примкнет D(6) индивид (1001) шестого типа.

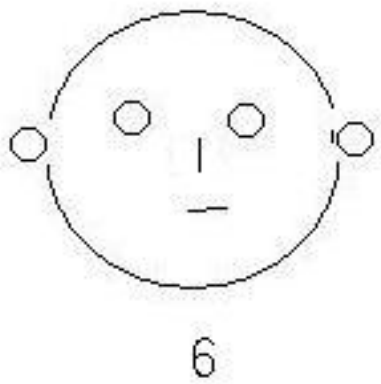


Figure: Индивид (1001) шестого типа.

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
	2		1	0	1	0	
F	5		0	0	0	0	
		I_F					
D	6		1	0	0	1	
		I_D					
		I_{F+D}					

Table: Форма для вычисления разнообразия объединений групп F(1, 2, 5)-D(6), B(3, 3, 4)-D(6).