xgboost для dummies

Винни-Пух

9 февраля 2017 г.

Предисловие

План:

1. Энтропия/Джини

Распределения максимизирующие энтропию? что-то про ROC кривые до кучи?

2. Одиноко стоящий дуб

Типичное заданичие: Вырастить дерево согласно такому-то критерию. Сюда борьбу с NA. Сюда же регуляризацию? Или отдельно?

3. Логит-модель

Логистическое распределение? Перевод y=0/1 в y=-1/1. Максимум правдоподобия в минимум штрафа? Предельные эффекты?

4. Мини-мими-лес

Типичное: Два-три дерева. По ним построить прогноз/оценить важность переменных. Что еще?

5. Регуляризация.

Общая идея. Парадокс James-Stein. Для среднего, для регрессии, для дерева. L1 и L2.

6. Про кросс-валидацию?

Как это делать руками? Какие тут теоретические задачи?

Упр: Дано одно-два-три дерева. И 5 наблюдений. Посчитать кросс-валидационную ошибку.

Упр: На наборе данных в 5 наблюдений подобрать параметр жесткости с помощью кросс-валидации.

7. Несколько практических упражнений.

Упр: сделайте с дефолтными параметрами и ответьте на все подробности про алгоритм
тут решения в python/R.

Упр: Нарисуйте дерево номер 5.

Из теории:

- определения
- табличка с параметрами xgboost, rforest
- несколько практик подбора параметров

```
library("knitr") # грамотное программирование
library("tikzDevice") # сохранение графиков в формате tikz

library("tidyverse") # Хэдли на нашей стороне
library("xtable")

theme_set(theme_bw()) # чёрно-белая тема для графиков
```

Глава 1

Неразобранные :)

1.1 Для случайных величин X и Y найдите индекс Джини и энтропию

X	0	1
$\mathbb{P}()$	0.2	0.8

,

Y	0	1	5
$\mathbb{P}()$	0.2	0.3	0.5

1.2 Случайная величина X принимает значение 1 с вероятностью p и значение 0 с вероятностью $1 - p$.

1. Постройте график зависимости индекса Джини и энтропии от p.
2. Являются ли функции монотонными? выпуклыми?
3. При каком p энтропия и индекс Джини будут максимальны?

1.3 Кот Леопольд анкетировал 20 мышей по трём вопросам: x — «Одобряете ли Вы непримиримую к котам позицию Белого и Серого?», y — «Известно ли Вам куда пропала моя любимая кошка Мурка?» и z — «Известны ли Вам настоящие имена Белого и Серого?» Результаты опроса в таблице:

```
set.seed(1975)
x <- sample(c("yes", "no"), size = 20, rep = TRUE)
y <- sample(c("yes", "no"), size = 20, rep = TRUE)
z <- sample(c("yes", "no"), size = 20, rep = TRUE)
xtable(data.frame(x, y, z))
```

1. Какой фактор нужно использовать при прогнозировании y, чтобы минимизировать энтропию?
2. Какой фактор нужно использовать при прогнозировании y, чтобы минимизировать индекс Джини?

1.4 Постройте регрессионное дерево для набора данных:

y_i	x_i
5	0
6	1
4	2
100	3

Критерий деления узла на два — минимизация RSS. Дерево строится до трёх терминальных узлов.

	x	y	z
1	no	no	yes
2	no	yes	yes
3	yes	yes	yes
4	yes	yes	no
5	no	no	no
6	no	yes	yes
7	no	no	yes
8	no	no	no
9	yes	no	yes
10	yes	no	yes
11	no	no	no
12	yes	yes	yes
13	no	yes	yes
14	no	yes	no
15	yes	no	no
16	yes	no	yes
17	no	no	no
18	no	yes	no
19	no	yes	no
20	yes	no	no

1.5 Постройте регрессионное дерево для набора данных:

y_i	x_i
100	1
102	2
103	3
50	4
55	5
61	6
70	7

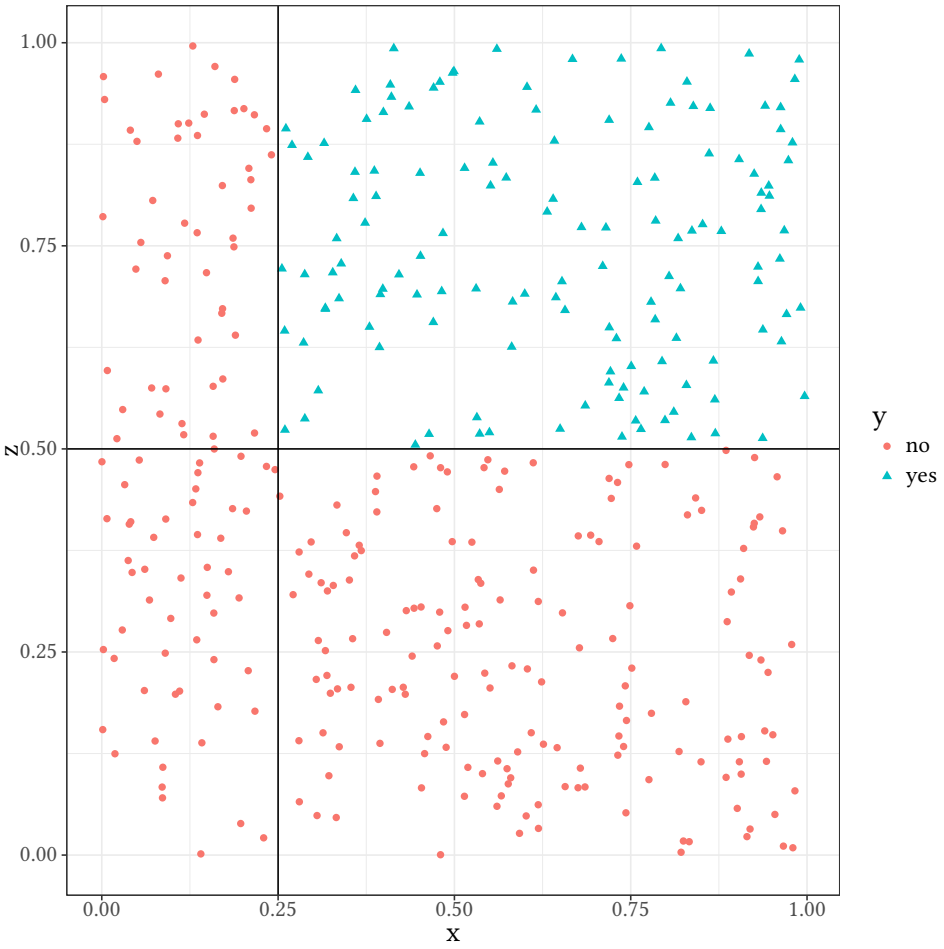
Критерий деления узла на два — минимизация RSS. Узлы делятся до тех пор, пока в узле остаётся больше двух наблюдений.

1.6 Дон-Жуан предпочитает брюнеток. Перед Новым Годом он посчитал, что в записной книжке у него 20 блондинок, 40 брюнеток, две рыжих и восемь шатенок. С Нового Года Дон-Жуан решил перенести все сведения в две записные книжки, в одну — брюнеток, во вторую — остальных.

Как изменились индекс Джини и энтропия в результате такого разбиения?

1.7 Машка пять дней подряд гадала на ромашке, а затем выкладывала очередную фотку «Машка с ромашкой» в инстаграмчик. Результат гадания — переменная y_i, количество лайков у фотки — переменная x_i. Постройте классификационное дерево.

y_i	x_i
плюнет	10
поцелует	11
поцелует	12
к сердцу прижмёт	13
к сердцу прижмёт	14



Дерево строится до идеальной классификации. Критерий деления узла на два — максимальное падение индекса Джини.

1.8 У Винни-Пуха есть 100 песенок (кричалок, вопелок, пыхтелок и сопелок). Каждый день он выбирает и поёт одну из них равновероятно наугад. Одну и ту же песенку он может петь несколько раз. Сколько в среднем песенок оказываются неспетыми за 100 дней?

1.9 По данной диаграмме рассеяния постройте классификационное дерево для зависимой переменной y:

```
set.seed(42)
df <- data.frame(x = runif(400), z = runif(400))
df$y <- factor(ifelse( df$x > 0.25 & df$z > 0.5, "yes", "no"))
qplot(data = df, x = x, y = z, col = y, shape = y) +
  geom_vline(xintercept = 0.25) + geom_hline(yintercept = 0.5)
```

Дерево необходимо построить до идеальной классификации, в качестве критерия деления узла на два используйте минимизацию индекса Джини.

1.10 Рассмотрим табличку:

y_i	x_i	z_i
y_1	1	2
y_2	1	2
y_3	2	2
y_4	2	1
y_5	2	1
y_6	2	1
y_7	2	1

Сколько существует принципиально разных классификационных деревьев для данного набора данных?

1.11 Исследовательница Мишель строит классификационное дерево для бинарной переменной y_i. Может ли при разбиении узла на два расти индекс Джини? Энтропия?

1.12 Приведите примеры наборов данных, для которых индекс Джини равен 0, 0.5 и 0.999.

1.13 Рассмотрим задачу построения классификационного дерева для бинарной переменной y_i. Приведите пример такого набора данных, что никакое разбиения стартового узла на два не снижает индекс Джини, однако двух разбиений достаточно, чтобы снизить индекс Джини до нуля.

Глава 2

Решения и ответы к избранным задачам

1.1.

1.2. $I = 2p(1-p)$, энтропия и индекс Джини максимальны при $p = 0.5$.

1.3.

1.4.

1.5.

1.6.

1.7.

1.8. $100 \cdot \left(\frac{99}{100}\right)^{100} \approx 100/e \approx 37$

1.9. Сначала делим по z, потом по x, так как индекс Джини в таком порядке падает сильнее.

1.10.

1.11. Нет, в силу выпуклости функций.

1.12. Все y_i одинаковые; поровну y_i двух типов; 1000 разных типов y_i, по одному наблюдению каждого типа.

1.13.

y_i	x_i	z_i
1	1	1
1	2	2
0	1	2
0	2	1

Список обозначений

Оглавление