

МСМС для dummies

Винни-Пух

18 февраля 2017 г.

Предисловие

План:

1. Энтропия/Джини

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

2. Одиноко стоящий дуб

Типичное задание: 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. Марковские цепи

- 1.1 Шахматный конь начинает в клетке A1. Каждый свой ход он выбирает равновероятно из возможных. Какова вероятность того, что через много-много ходов он окажется в клетке H8? Сколько в среднем длится путь от клетки A1 до клетки A1?

1.2. Ручные задачи

- 1.2 Случайные величины X_i независимы и одинаково распределены с табличкой

X	1	2	6
$\mathbb{P}()$	β	2β	$1 - 3\beta$

Известно, что $X_1 = 1, X_2 = 2, X_3 = 2, X_4 = 4$.

1. Найдите оценку $\hat{\beta}$ методом моментов
2. Найдите оценку $\hat{\beta}$ методом максимального правдоподобия
3. Предположим, что β равномерно на отрезке $[0; 1/3]$. Найдите апостериорную условную функцию плотности β с учётом полученных наблюдений. С какой функцией она совпадает?
4. Предположим, что β имеет функцию плотности $f(t) = 18t$ на отрезке $[0; 1/3]$. Найдите апостериорную функцию плотности β .
5. Решите данную задачу с помощью STAN.

1.3. Гиббс

1.4. Метрополис-Гастингс

- 1.3 Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для биномиального распределения $\text{Bin}(n, p)$ из равновероятного на множестве $\{0, 1, 2, \dots, n\}$
- 1.4 Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для биномиального распределения $\text{Bin}(n, p)$ из симметричного случайного блуждания на \mathbb{Z}
- 1.5 Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для геометрического распределения $\text{Geom}(p)$ из симметричного случайного блуждания на \mathbb{Z}
- 1.6 Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для пуассоновского распределения $\text{Pois}(\lambda)$ из симметричного случайного блуждания на \mathbb{Z}

- 1.7** Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для функции плотности $\pi(x) \sim \exp(-x^2)(3 + x^2 + \cos x)$ из нормального $N(0, 1)$. Из нормального $N(0, \sigma^2)$
- 1.8** Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для функции плотности $\pi(x) \sim \exp(-x^2)(3 + x^2 + \cos x)$ из случайного блуждания $X_{t+1} = X_t + \varepsilon_t$, где $\varepsilon_t \sim N(0, 1)$. Вариант с $N(0, \sigma^2)$
- 1.9** Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для стандартного нормального распределения $N(0, 1)$ из случайного блуждания $X_{t+1} = X_t + \varepsilon_t$, где $\varepsilon_t \sim U[-1, 1]$
- 1.10** Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для двумерного нормального распределения $N(0, A)$, $A = \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}$ из случайного блуждания $X_{t+1,i} = X_{t,i} + \varepsilon_{t,i}$, где $\varepsilon_{t,i} \sim U[-1, 1]$.
- 1.11** Используя алгоритм Метрополиса-Хастингса сгенерируйте выборку для двумерного распределения с функцией плотности $p(x, y) = \exp(-4x^2 - 6y^2 + 2x - y + xy)$, $x > 0, y > 0$ из случайного блуждания $X_{t+1,i} = X_{t,i} + \varepsilon_{t,i}$, где $\varepsilon_{t,i} \sim U[-1, 1]$.

1.5. Регуляризация

- 1.12** Рассмотрим модель

$$y = X\beta + u,$$

где u_i независимы и $\mathcal{N}(0; \sigma^2)$.

Метод гребневой регрессии предполагает минимизацию функции

$$Q(\beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^k \beta_j^2.$$

Рассмотрим байесовский подход к регрессии. Предположим, что априорное распределение имеет вид $\sigma^2 \sim InvGamma(a, b)$, $\beta_j | \sigma^2 \sim \mathcal{N}(0; c)$.

При каких a, b и c апостериорная мода $\hat{\beta}_{MAP}$ совпадёт с $\hat{\beta}_{Ridge}$?

- 1.13** Рассмотрим модель

$$y = X\beta + u,$$

где u_i независимы и $\mathcal{N}(0; \sigma^2)$.

Метод LASSO предполагает минимизацию функции

$$Q(\beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^k |\beta_j|.$$

Рассмотрим байесовский подход к регрессии. Предположим, что априорное распределение имеет вид $\sigma^2 \sim InvGamma(a, b)$, $\beta_j | \sigma^2 \sim DoubleExp(c)$.

При каких a, b и c апостериорная мода $\hat{\beta}_{MAP}$ совпадёт с $\hat{\beta}_{LASSO}$?

- 1.14** Храбрый Охотник ловит Покемонов в случайном порядке. Вес i -го пойманного Покемона, y_i , имеет нормальное распределение $\mathcal{N}(\mu; \sigma^2)$. Параметры μ и σ неизвестны.

Храбрый охотник хочет оценить μ по формуле $\hat{\mu} = c \sum_{i=1}^n y_i$.

1. При каком c величина $\mathbb{E}((\hat{\mu} - \mu)^2)$ будет минимальна?
2. Возможно ли использовать на практике данное c ?

Глава 2

Решения и ответы к избранным задачам

1.1.

1.2.

1.3.

1.4.

1.5.

1.6.

1.7.

1.8.

1.9.

1.10.

1.11.

1.12.

1.13.

1.14. $c = \frac{1}{n+\mu^2/\sigma^2}$, нет, так как μ и σ^2 неизвестны.

Список обозначений

Оглавление

1	Неразобранные :)	5
1.1	Марковские цепи	5
1.2	Ручные задачи	5
1.3	Гиббс	5
1.4	Метрополис-Гастингс	5
1.5	Регуляризация	6
2	Решения и ответы к избранным задачам	7