Эконометрика. Лекция 10. Три сюжета напоследок

Три сюжета

- Квантильная регрессия
- Алгоритм случайного леса
- Байесовский подход

Квантильная регрессии

Моделировать можно не только среднее, но и медиану или другой определённый квантиль.

Классическая регрессия — модель для среднего

Предпосылки классической модели:

- $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
- экзогенность, $E(\varepsilon_i | x_i) = 0$
- другие предпосылки

Следствие:

$$E(y_i | x_i) = \beta_1 + \beta_2 x_i$$

Минимизация суммы квадратов

Модель: $E(y_i|x_i) = \beta_1 + \beta_2 x_i$

- Сумма квадратов остатков, $Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_i (y_i - \hat{y}_i)^2$
- При минизации $Q(\hat{\beta}_1, \hat{\beta}_2)$ получаем состоятельные оценки $\hat{\beta}_1$, $\hat{\beta}_2$

Медианная регрессия

Модель: $Med(y_i|x_i) = \beta_1 + \beta_2 x_i$

Алгоритм получения оценок

- Сумма модулей остатков, $M(\hat{\beta}_1, \hat{\beta}_2) = \sum_i |y_i - \hat{y}_i|$
- Минимизируя $M(\hat{\beta}_1, \hat{\beta}_2)$ получаем состоятельные оценки $\hat{\beta}_1$, $\hat{\beta}_2$

Пример у неоновой доски

Найдите оценку $\hat{\beta}$ медианной регрессии:

$$Med(y_i|x_i) = \beta x_i$$

Набор данных:

y	x
1	1
2	5
6	5

Медианная и классическая регрессия

- Классическая: от каких факторов зависит $E(y_i|x_i)$?
- Медианная: от каких факторов зависит $Med(y_i|x_i)$?
- Если распределение ε_i симметрично, то оба подхода дают асимптотически одинаковые оценки

Медианная регрессия: минусы

- Нет явных формул для оценок коэффициентов и стандартных ошибок
- Только асимптотические свойства оценок коэффициентов

Медианная регрессия: плюсы

- Взгляд на данные с другой стороны
- Более устойчивые оценки в случае “выбросов” в ε_i

Произвольная квантиль

- Медиана, $Med(y_i)$, — квантиль 50%

$$P(y_i \leq Med(y_i)) = 0.5$$

- Квантиль порядка τ, q_τ:

$$P(y_i \leq q_\tau) = \tau$$

Квантильная регрессия

Модель: $q_\tau(y_i|x_i) = \beta_1^\tau + \beta_2^\tau x_i$

- Зависимость для разных квантилей может быть разная!

Асимметричная сумма модулей остатков:

$$M(\hat{\beta}_1, \hat{\beta}_2) = \sum_i \rho_\tau(y_i - \hat{y}_i)$$

где

$$\rho_\tau(y_i - \hat{y}_i) = \begin{cases} (1-\tau) \cdot |y_i - \hat{y}_i|, & y_i < \hat{y}_i \\ \tau \cdot |y_i - \hat{y}_i|, & y_i \geq \hat{y}_i \end{cases}$$

- Минимизируя $M(\hat{\beta}_1, \hat{\beta}_2)$ получаем состоятельные оценки $\hat{\beta}_1$, $\hat{\beta}_2$

Квантильная регрессия стоимости квартир

totsp 1.31148

totsp 2.09259

totsp 3.64286

(здесь вставить график)

Алгоритм случайного леса

- Очень хорошо прогнозирует
- Не объясняет, как устроены данные

Две версии алгоритма

- Для непрерывной y_i
- Для качественной y_i

Каждый мужчина должен посадить дерево

(здесь картинка дерева)

Работа с деревом

- Построение дерева по имеющимся данным
- Прогнозирование с помощью дерева

Как посадить дерево?

- Из имеющихся k переменных случайно отбираем $k' = \lceil k/3 \rceil$ переменных
- Из отобранных k' переменных выбираем ту, которая даёт наилучшее деление ветви дерева на две
- Повторяем до тех пор, пока в каждом терминальном узле остаётся больше $nodesize = 5$ наблюдений

Алгоритм случайный

Повторное применение алгоритма к тому же набору данных даст слегка другие оценки

Неоновая доска

Пример построения классификационного дерева

y	x
1	1
2	2
9	3
10	4
10	5

Мужчина, владеющий R, может посадить целый лес!

- Случайным образом отбираем (с повторениями) n наблюдений из исходных n наблюдений
- Сажаем дерево по случайной подвыборке
- Повторяем до получения $n_{tree} = 500$ деревьев

Прогноз случайного леса:

- Каждое из $n_{tree} = 500$ деревьев даёт свой прогноз $\hat{y}_i$
- Усредняем и получаем финальный прогноз

Байейсовский подход

Опишем наше незнание параметра θ в виде априорного закона распределения!

Пример. Неизвестная вероятность

- $p \in [0; 1]$

Априорная плотность:

$$f(p) = \begin{cases} 1, & p \in [0; 1] \\ 0, & \text{иначе} \end{cases}$$

(здесь картинка)

Пример. Неизвестный положительный коэффициент

- $\beta \in [0; +\infty)$

Априорная плотность:

$$f(\beta) = \begin{cases} exp(-\beta), \ \beta \in [0; \infty) \\ 0, \quad \text{иначе} \end{cases}$$

(здесь картинка)

Модель

Модель задаёт закон распределения наблюдений, y_i, при фиксированном значении параметров

Например,

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$$

Кристально-чистая логика байесовского подхода

Определяем:

- Априорное распределение, $f(\theta)$
- Модель для данных, $f(y|\theta)$

Получаем:

- Апостериорное распределение, $f(\theta|y)$

Формула условной вероятности

$$f(\theta|y) = \frac{f(y|\theta) \cdot f(\theta)}{f(y)} \sim f(y|\theta) \cdot f(\theta)$$

Пример у неоновой доски

Караси, щуки

- нет информации
- Бабушка: караси встречаются чаще щук!

Как описать сложную функцию плотности?

(картинка)

- Большая выборка независимых значений случайной величины r
- Можно оценить всё: $E(r)$, $E(r^2)$, $P(r > 0)$

Монте-Карло по схеме Марковской цепи

MCMC (Markov Chain Monte Carlo)

На входе:

- Априорное распределение, $f(\theta)$
- Модель для данных, $f(y|\theta)$

На выходе:

- Большая выборка из апостериорного распределения, $f(\theta|y)$

(картинка)

Регрессия пик-плато

Пример

Большое спасибо

Большое спасибо тем, кто прошел вместе с нами этот курс до конца