DS-поток, 3 курс, осень 2022

Статистика

1. Введение

Введение

Теория вероятностей

Зная природу случайного явления,

посчитать характеристику этого явления.

Математическая статистика

По результатам экспериментальных данных

высказать суждение о том, какова была природа этого явления.

Классический пример

На третьем курсе N студентов; из них M выбирает DS-поток.

Задача в теории вероятностей

$$\mathsf{P}(\text{среди } n \text{ чел. ровно } m \text{ слушателей DS-потока}) - ?$$

Предполагается, что M *известно.*

Задача в математической статистике

Среди случайных n чел. есть m слушателей DS-потока.

Оценить M.

Предполагается, что M *не известно.*

Еще пример

$\xi \sim \mathcal{N}(a, \sigma^2)$ — случайная величина

Задача в теории вероятностей

Известно, что $a = 2.3, \sigma = 7.1$

$$\mathsf{P}(\xi \in [0, 1]) - ?$$

$$\mathsf{E}\xi - ?$$

Задача в математической статистике

$x_1, ..., x_n$ — независимые реализации случайной величины ξ.

Оценить a и σ.

Задача математической статистики

Пусть $x_1, ..., x_n$ — численные характеристики n-кратного повторения некоторого явления.

Будем их воспринимать как независимые реализации $\xi \sim \mathsf{P}$.

Задача: по значениям $x_1, ..., x_n$ высказать некоторое суждение о распределении P.

Решение: *статистический вывод* или *обучение*.

1. Введение

Однократный эксперимент

$\mathscr{X}$ — *выборочное пространство* = множество всех возможных значений эксперимента;

$\mathscr{B}_{\mathscr{X}}$ — некоторая σ-алгебра на $\mathscr{X}$;

P — некоторое неизвестное распределение на $(\mathscr{X}, \mathscr{B}_{\mathscr{X}})$;

Предполагается $\mathsf{P} \in \mathscr{P}$ — некоторое семейство распределений.

Вероятностно-статистическая модель

$$(\mathscr{X}, \mathscr{B}_{\mathscr{X}}, \mathscr{P}).$$

Для оперирования с результатами эксперимента как со случайными величинами, определим случайную величину $X : \mathscr{X} \to \mathscr{X}$ по правилу $X(x) = x \ \forall x \in \mathscr{X}$, которую будем называть *наблюдением*.

Многократный эксперимент

Вероятностно-статистическая модель

$$(\mathscr{X}^n, \mathscr{B}_{\mathscr{X}^n}, \mathscr{P}^n),$$

- $\mathscr{X}^n = \mathscr{X} \times ... \times \mathscr{X}$;
- $\mathscr{B}_{\mathscr{X}^n} = \sigma(B_1 \times ... \times B_n, B_i \in \mathscr{B}_{\mathscr{X}})$;
- $\mathscr{P}^n = \{\mathsf{P}^n, \mathsf{P} \in \mathscr{P}\}$, причем

  $\mathsf{P}^n(B_1 \times ... \times B_n) = \mathsf{P}(B_1)...\mathsf{P}(B_n)\ \forall B_i \in \mathscr{B}_{\mathscr{X}}$.

  Для любой $\mathsf{P} \in \mathscr{P}$ определенная таким образом P^n существует и единственна по теореме о продолжении вероятностной меры.

Многократный эксперимент

Наблюдение, соответствующее i-му эксперименту:

Сл. вел. $X_i : \mathscr{X}^n \to \mathscr{X}$, т.ч. $X_i(x) = x_i \ \forall x \in \mathscr{X}^n$.

Случайный вектор $X = (X_1, ..., X_n)$ — *выборка* размера n.

Выборка является вектором независимых

одинаково распределенных случайных величин,

каждая компонента которого имеет распределение P.

Бесконечный эксперимент

Вероятностно-статистическая модель

$$(\mathscr{X}^\infty, \mathscr{B}_{\mathscr{X}^\infty}, \mathscr{P}^\infty),$$

- $\mathscr{X}^\infty = \mathscr{X} \times \mathscr{X} \times \dots;$
- $\mathscr{B}_{\mathscr{X}^\infty} = \sigma(B_1 \times \dots \times B_n \times \mathscr{X}^\infty,\ B_i \in \mathscr{B}_{\mathscr{X}}, n \in \mathbb{N});$
- $\mathscr{P}^\infty = \{\mathsf{P}^\infty, \mathsf{P} \in \mathscr{P}\}$, причем

  $\mathsf{P}^\infty(B_1 \times \dots \times B_n \times \mathscr{X} \times \dots) = \mathsf{P}^n(B_1 \times \dots \times B_n)\ \forall B_i \in \mathscr{B}_{\mathscr{X}}.$

  Для любой $\mathsf{P} \in \mathscr{P}$ определенная таким образом P^n существует и единственна по теореме о продолжении вероятностной меры.

Бесконечный эксперимент

Наблюдение, соответствующее i-му эксперименту:

Сл. вел. $X_i : \mathscr{X}^\infty \to \mathscr{X}$, т.ч. $X_i(x) = x_i \; \forall x \in \mathscr{X}^\infty$.

Случайная последовательность $X = (X_1, X_2, ...)$ —

выборка неограниченного размера.

Выборка неограниченного размера является последовательностью независимых одинаково распределенных случайных величин, каждая компонента которого имеет распределение P.

Далее

- Для простоты будем опускать индексы n и ∞;
- Будем считать, что в качестве сигма-алгебры используется борелевская, если не сказано обратное;
- Да и вообще забудем про ее существование :)

1. Введение

Классификация по типу методов вывода

1. *Параметрический*

   Предполагается, что истинное распределение P принадлежит некоторому классу распределений $\mathscr{P}$,

   которое параметризовано параметром $\theta \in \Theta$.

   $\mathsf{P} \in \{\mathsf{P}_\theta \mid \theta \in \Theta\}$

2. *Непараметрический*

   Предполагается, что истинное распределение P принадлежит некоторому классу распределений $\mathscr{P}$,

   на котором не введен параметр.

Параметрический подход

1. $X_1, ..., X_n \sim Exp(\theta)$, где $\theta > 0$, подразумевает:

   $\mathscr{X} = (0, +\infty), \mathscr{P} = \{Exp(\theta) \mid \theta > 0\}, \Theta = (0, +\infty)$.

   Статистический вывод: указание числа из множества Θ.

2. Схема испытаний Бернулли $X_1, ..., X_n$ подразумевает:

   $\mathscr{X} = \{0, 1\}, \mathscr{P} = \{Bern(\theta) \mid 0 \leqslant \theta \leqslant 1\}, \Theta = [0, 1]$.

   Статистический вывод: указание числа из множества Θ.

3. $X_1, ..., X_n \sim \mathcal{N}(a, \sigma^2)$, где оба параметра неизвестны:

   $\mathscr{X} = \mathbb{R}, \mathscr{P} = \{\mathcal{N}(a, \sigma^2) \mid a \in \mathbb{R}, \sigma > 0\}$,

   $\theta = (a, \sigma), \Theta = \mathbb{R} \times (0, +\infty)$.

   Статистический вывод: указание пары чисел из множества Θ.

Пример

$\mathscr{X} = \mathbb{R}$;

$\mathscr{P} = \{U(0, \theta) \mid \theta > 0\}$;

Дана выборка $(X_1, X_2, X_3) = (1, 2, 3)$.

Может ли истинное значение θ быть равным 100 , 3 , 1.5 , -1?

- 100 и 3 — да;
- -1 — нет, поскольку $\theta > 0$;
- 1.5 — да, поскольку $\mathscr{X} = \mathbb{R}$

  $\Rightarrow$ возможны любые вещественные числа, правда вероятность получения хотя бы одного числа вне отрезка $[0, \theta]$ равна нулю.

Непараметрический подход

1. В отсутствии предположений:

   $\mathscr{X} = \mathbb{R}$;

   $\mathscr{P}$ — все распределения на $\mathbb{R}$.

   В качестве статистического вывода можно

   некоторым образом оценить функцию распределения;

2. Предполагается, что выборка взята из непрерывного распред.:

   $\mathscr{X} = \mathbb{R}$;

   $\mathscr{P}$ — все непрерывные распределения на $\mathbb{R}$.

   В качестве статистического вывода можно оценить плотность.

- Любое семейство распределений может рассматриваться как в параметрическом подходе, так и в непараметрическом;
- Смысл деления на два типа — принципиально разные методы к оценке неизвестного распределения:
  - Методы параметрического подхода как-либо оценивают параметр, соответствующий неизвестному распределению. На практике обычно $\Theta \subset \mathbb{R}^d$, размерность d фиксирована;
  - Непараметрические методы пытаются некоторым способом напрямую оценить неизвестное распределение. На практике обычно содержат нефиксированное количество параметров.

Классификация по способу вывода

Два доминирующих подхода:

1. *Частотный*

   Построение суждения о распределении P происходит только на основе выборки $X_1, \dots, X_n$.

   Все распределения из класса $\mathscr{P}$ равноправны.

2. *Байесовский*

   На распределениях из $\mathscr{P}$ задано некоторое распределение ("априорное знание"), которое учитывается при построении суждения, как правило, с помощью формулы Байеса.

   Пример: $\mathscr{P} = \{\mathcal{N}(\theta, 1) \mid \theta \in \mathbb{R}\}$,

   и предполагается, что истинное значение θ_0 выбрано из $\mathcal{N}(0, 1)$.

1. Введение

Пример

$X_1, \ldots, X_n \sim \mathcal{N}(a, \sigma^2)$, где оба параметра неизвестны.

Строго супер формально: $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathscr{P}^n)$,
где $\mathscr{P} = \{\mathcal{N}(a, \sigma^2) \mid a \in \mathbb{R}, \sigma > 0\}$, $\theta = (a, \sigma), \Theta = \mathbb{R} \times (0, +\infty)$.
Это мы упрощаем до $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathscr{P})$, работаем только с $\mathbb{R}$ и $\mathscr{P}$.

Найдем оценку по методу моментов. Решаем систему

$$\begin{cases} \mathsf{E}_\theta X_1 = \overline{X}; \\ \mathsf{E}_\theta X_1^2 = \overline{X^2}. \end{cases} \qquad \begin{cases} a = \overline{X}; \\ a^2 + \sigma^2 = \overline{X^2}. \end{cases} \qquad \begin{cases} a = \overline{X}; \\ \sigma = \sqrt{\overline{X^2} - a^2}. \end{cases}$$

Получаем оценки

Оценка $\widehat{a}$ сильно состоятельна: $\widehat{a} = \overline{X} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta X = a$.

Строго супер формально:

1. беск. выборка $X_1, X_2, \ldots \sim \mathcal{N}(a, \sigma^2)$ на пр-ве $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty), \mathscr{P}^\infty)$;
2. посл-ть оценок $\widehat{a}_n = \frac{1}{n}\sum_{i=1}^n X_i$ сильно сост., т.к. $\widehat{a}_n \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} a$.


ВСЁ!