Теория вероятности и математическая статистика

«Машинное обучение», 21.01.2018

Теория вероятности – это просто

- **Испытание**(опыт, эксперимент) – наблюдение какого-либо явления при соблюдении определенного комплекса условий, который должен каждый раз строго выполняться при повторении данного испытания.
- **Событие** – факт, появление которого регистрируется в результате испытания.
- **Пример 1**: испытание – подбрасывание монетки, событие – выпадение орла.
- **Пример 2**: испытание – выстрел в мишень, событие – попадание в десятку.

Какие бывают события

- **Достоверное** событие: если монету подбросить, то она всегда упадет на Землю
- **Невозможное** событие: если монету подбросить, она улетит вверх
- **Случайное** событие: выпадение орла при подбрасывании монетки – может выпасть, а может и нет

Равновероятные события

Игральный кубик

Грани одинаковые – вероятность выпадения каждой из граней одинаковая

Карточная колода

Тянем одну карту из колоды – вероятность вытащить конкретную карту одинаковая для всех карт

Определение вероятности

- $P(A) = \frac{m}{n}$

- A – событие
- m – количество благоприятствующих событий
- n – общее количество событий

Условия применения

- События должны быть элементарные
- События должны быть равновероятными

- Пример 1.
- Какая вероятность выпадения 1 или 2 на кубике?
- $P(A) = \frac{2}{6} = \frac{1}{3}$

Условия применения

- Пример 2.
- Какая вероятность выпадения 2 орлов при 3 подбрасываниях монетки
- $P(A) = \frac{2}{3}$ - неверно!
- Элементарные события: (О,О,О), (О,О,Р), (О,Р,О), (О,Р,Р), (Р,О,О), (Р,О,Р), (Р,Р,О), (Р,Р,Р)
- $P(A) = \frac{3}{8}$

Как записывают вероятности

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

Первая строка – события

Вторая строка – вероятности событий

Все вероятности в сумме должны давать 1

Задание: составить таблицу для выпадения орлов при 3 подбрасываниях монетки

Еще примеры

Игровой автомат

X	-100	200	400	600
P	3/4	1/6	1/24	1/24

Как считать вероятности сложных событий

- Независимые события – события, которые не влияют друг на друга (если произошло одно событие, то вероятность второго события не меняется)

- Вероятность того, что произойдет несколько независимых совместимых событий:
- $P(A, B) = P(A) * P(B)$

- Вероятность того, что произойдет хотя бы одно из двух независимых совместимых событий:
- $P(A + B) = P(A) + P(B) - P(A) * P(B)$

Примеры

- *Пусть A – пойдет дождь, B – я съем на завтрак творог*
- *P(A)=0.5,   P(B)=0.2*

- $P(A, B) = P(A) * P(B) = 0.5 * 0.2 = 0.1$
- $P(A, B) = P(A) + P(B) - P(A) * P(B) = 0.5 + 0.2 - 0.1 = 0.6$

Математическое ожидание

- Описывает некоторый закон распределения
- Как считать:

$$E(X) = (p_1 * x_1 + p_2 * x_2 + \ldots + p_n * x_n)$$

Пример с игральным кубиком

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = \left(\frac{1}{6} * 1 + \frac{1}{6} * 2 + \ldots + \frac{1}{6} * 6\right) = 3.5$$

Дисперсия

- Описывает некоторый закон распределения
- Как считать:

$$\mathrm{D}(X) = (p_1(x_1 - E(X))^2 + \ldots + p_n(x_n - E(X))^2)$$

Пример с игральным кубиком

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

$$D(X) = \left(\frac{1}{6}(1 - 3.5)^2 + \ldots + \frac{1}{6}(6 - 3.5)^2\right) \approx 2.91$$

Стандартное отклонение

- *Стандартное отклонение – это квадратный корень из дисперсии*
- $Std = \sqrt{D} \approx 1.71$

Статистика – зачем она нужна

- Статистика – важный инструмент **анализа** данных. Если мы посчитаем разные статистики
- Статистический анализ проводится для некоторой **выборки** – некоторый набор величин

- **Пример 1**: 10 испытаний игрального кубика

  1, 2, 6, 1, 4, 4, 5, 2, 3, 2, 5

- **Пример 2**: рост учеников класса

  160, 158, 172, 151, 157, 162, 160, 159, 167, 155

Выборочное среднее – аналог математического ожидания, но только для выборки

$$\bar{X} = (x_1 + x_2 + x_3 + \ldots + x_n)/n$$

- **Пример 1**: средняя значение выпавшего значения на грани кубика

  1, 2, 6, 1, 4, 4, 5, 2, 3, 2, 5

  $\bar{X}$ = (1+2+6+1+4+4+5+2+3+2)/10=3

Выборочная мода

- Выборочная **мода** – это значение, которое встречается в выборке чаще остальных

- **Пример 1**: мода значения на грани кубика =2

  1, 2, 6, 1, 4, 4, 5, 2, 3, 2

Выборочная медиана

- Выборочная **медиана** – одно из значение выборки, которая делит её пополам.

- Как можно посчитать: отсортировать выборку по возрастанию и взять значение из серединки выборки

- **Пример 1**: сортируем покупки за последнюю неделю

  1, 1, 2, 2, 2, 3, 4, 4, 5, 6

- Берем значение посередине – либо 2, либо 3

Выборочная дисперсия

- $\bar{X}$ – среднее
- $\bar{D} = ((x_1 - \bar{X})^2 + (x_1 - \bar{X})^2 + \ \dots + (x_1 - \bar{X})^2)/n$

- **Пример 1**: средняя стоимость покупок за последнюю неделю

  1, 2, 6, 1, 4, 4, 5, 2, 3, 2

  $\bar{D} = \left((1-3)^2 + (2-3)^2 + \ \dots + (3-2)^2\right)$/10=2.5

Выборочное стандартное отклонение

- Выборочное стандартно отклонение, так же как и обычное стандартное отклонение – это квадратный корень из выборочной дисперсии
- $\overline{Std} = \sqrt{\overline{D}} \approx 1.58$