Введение в анализ данных и статистику

Омелюсик Владимир Степанович

Национальный Исследовательский Университет
«Высшая школа экономики»
–
Факультатив «Введение в анализ данных и машинное обучение на Python»

19 октября 2019 г.

Направления математики

- Линейная алгебра
- Математический анализ
- Дифференциальные уравнения
- Теория вероятностей
- ...

Теория вероятностей и математическая статистика

- Теория вероятностей:
  - Знаем модель некоторого явления.
  - Можем посчитать вероятность наступления какого-то события.
  - И другие параметры модели.
- Математическая статистика:
  - Модель явления нам неизвестна.
  - Но есть наблюдения над этим явлением – данные.
  - Идея: оценить параметры модели по данным.

Пример: подбрасывание монетки

- Пусть мы знаем вероятность выпадения орла при подбрасывании монетки равна $\frac{1}{2}$. Запишем это так:

$$\mathbb{P}\{\text{Выпадет орёл}\} = \frac{1}{2}.$$

  Сумма вероятностей должна равняться 1. Поэтому вероятность выпадения решки тоже равна $\frac{1}{2}$.

- Это наша модель, в которой мы знаем, чему равны вероятности выпадения орла и решки. В реальной жизни мы эти вероятности не знаем и никогда не узнаем.

Пример: подбрасывание монетки

- Что делать, если мы хотим узнать эти вероятности? Попробуем получить их **оценки**.
- Проведём эксперимент: попросим человека подбрасывать монетку 100 раз. Запишем результаты подбрасывания:

$$\text{О, О, Р, О, Р, Р, Р, О} \ldots \text{О, Р.}$$

- Оценку вероятности выпадения орла можно рассчитать, например, таким образом:

$$\hat{\mathbb{P}}\{\text{Выпадет орёл}\} = \frac{\text{Число раз, когда выпал орёл}}{\text{Общее число подбрасываний}}.$$

- Оценка вероятности равна доле. Можно показать, что при некоторых условиях $\hat{\mathbb{P}}$ является «хорошей» оценкой $\mathbb{P}$ (будем понимать «хорошей» пока на интуитивном смысле).

Пример: среднее число посетителей

- Случайная величина – величина, значение которой зависит от результата случайного происшествия (не подчиняющегося какому-либо шаблону).
- Математическое ожидание – среднее значение случайной величины при проведении эксперимента много раз. Очень много раз.
- Пусть случайная величина X – число посетителей ресторана за один день. Пусть математическое ожидание X:

$$\mathbb{E}(X) = 78.$$

Пример: среднее число посетителей

- Чтобы оценить математическое ожидание, будем записывать число посетителей в течение 100 дней:

$$X_1, X_2, X_3 \dots X_{100}.$$

- В качестве оценки математического ожидания рассчитаем среднее арифметическое по всем наблюдениям:

$$\hat{\mathbb{E}}(X) = \frac{X_1 + X_2 + \dots X_{100}}{100}.$$

- Оценка математического ожидания равна среднему. Можно показать, что при некоторых условиях $\hat{\mathbb{E}}$ является «хорошей» оценкой $\mathbb{E}$.

Зачем пытаться узнать параметры модели?

Для решения практических задач:

- Банку, выдающему кредиты, необходимо знать, с какой вероятностью кредит могут не вернуть.
- Владельцу торговой сети необходимо знать, где открыть новый магазин. Среднее число людей, проходящих в конкретном месте за день, может быть полезной характеристикой.
- Рекомендательная система должна определить, с какой вероятностью контент понравится пользователю, и предложить наиболее подходящий вариант.

Данные

- Для получения оценок нам необходимы наблюдения (данные). **Качество оценки** сильно зависит от **качества данных**.
- Откуда берутся данные? Вообще говоря, из генеральной совокупности.

Генеральная совокупность

Совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реализованном комплексе условий.

Данные

Выборка

Часть генеральной совокупности, используемая для проведения эксперимента.

Пример:

- Оценить вероятность того, что случайно выбранный ученик Лицея НИУ ВШЭ изучает английский язык.
- Генеральная совокупность: все ученики Лицея НИУ ВШЭ.
- Выборка: 10-е классы Лицея НИУ ВШЭ.

Репрезентативность выборки

Репрезентативность выборки

Способность выборки описывать свойства генеральной совокупности.

- Исследуем выборку, получаем среднее, оценки вероятностей и проч.
- Можем ли сказать, что генеральная совокупность имеет то же среднее, оценки вероятностей и проч.?
- Да, если выборка репрезентативна.

Репрезентативность выборки

Чтобы быть репрезентативной, выборка должна быть:

- Несмещённой в том смысле, что в ней должны присутствовать те же классы, что и в генеральной совокупности, в тех же пропорциях.
  Продолжение примера: если мы знаем, что в Лицее НИУ ВШЭ большинство изучает английский, но некоторые также изучают французский и немецкий, то в несмещённой выборке большинство лицеистов должны изучать английский, но также должны быть изучающие французский и немецкий.

- Достаточной по числу наблюдений (чем больше, тем лучше). Это позволяет учесть больше информации и получить более точные оценки.

Случайность выборки

Чтобы быть репрезентативной, выборка должна быть случайной:

- То есть данные должны быть собраны случайным образом.
- Участие экспериментатора должно быть минимальным.

Пример: исследуем зависимость числа людей, посмотревших фильм, от рейтинга фильма на «Кинопоиске».

- Как сформировать случайную выборку?

Как выглядит выборка?

- А если мы хотим изучить зависимость от нескольких факторов?
- Нужно включить в выборку несколько переменных!

Общий вид выборки – таблица «объекты-признаки» (N – число наблюдений, k – число признаков):

	X_1	X_2	X_3	$\ldots$	X_k
1					
2					
3					
$\vdots$					
N					

Про терминологию

- Зависимая переменная (target) – переменная, значение которой хотим предсказать (например, прибыль кафе):

$$Y$$

- Признаки или Объясняющие переменные (features) – переменные, при помощи которых хотим предсказать значение зависимой переменной (например, расстояние в метрах до станции метро, время года, координаты расположения кафе):

$$X_1, X_2, \ldots X_k$$

- Наблюдения или объекты (observations) – конкретная реализация зависимой и объясняемой переменной. Наблюдений N штук.

Пример: кафе

- Исследуем зависимость прибыли кафе от расстояния в метрах до станции метро, времени года и координат расположения кафе. Формализуем:
  - Y – прибыль кафе за месяц в тысячах рублей.
  - X_1 – расстояния в метрах до станции метро.
  - X_2 – время года.
  - X_3 – координаты расположения кафе.
- Пусть нам удалось собрать всего три наблюдения. Тогда таблица «объекты-признаки» может выглядеть следующим образом (числа придуманы):

	X_1	X_2	X_3
1	100	Зима	(30, 76)
2	40	Зима	(21, 12)
3	150	Лето	(4, 49)

Типы признаков

- Вещественные признаки (абсолютные или относительные): $X_k \in \mathbb{R}$.
  - Возраст, площадь класса.
- Категориальные номинальные признаки: $X_k \in \{\text{неупорядоченное множество}\}$.
  - Цвет, название города.
- Категориальные ранговые признаки: $X_k \in \{\text{упорядоченное множество}\}$.
  - Воинское звание, ступень образования.
- Бинарные признаки: $X_k \in \{0, 1\}$.
  - Материал, из которого сделан стол, – дерево?

Пример: типы признаков

Определить типы признаков следующей таблицы «объекты-признаки»:

	X_1	X_2	X_3	X_4	X_5	X_6
1	110	Россия	13%	Small	0	13.5
2	90	Россия	10%	Medium	0	10
3	105	Германия	30%	Medium	1	20
4	92	США	11%	Small	1	18.9
5	114	Россия	11%	Large	0	10
6	90	Франция	5%	Small	1	19

$N = ?$, $k = ?$.

Визуализация выборки

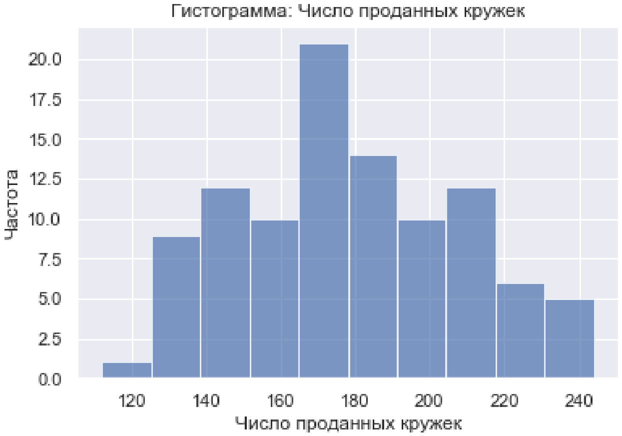
- Мы можем с лёгкостью визуализировать количественные переменные, чтобы понять примерный вид зависимостей в данных.
- Это важно для дальнейшей интерпретации адекватности моделей:
  - Если на «хороших» данных точно видно, что зависимость положительная, а модель показывает отрицательную модель, то вероятно, что-то не так с моделью.

Диаграмма рассеяния (scatter plot)

По осям отложены интересующие переменные (данные сгенерированы).

Гистограмма (histogram)

По нижней оси отложены значения переменных, а по левой оси – частота встречаемости (данные сгенерированы).



Типы данных

- Кроссекционные данные: время фиксировано, зависимая переменная изменяется по объектам.

N	Y
1	10
2	40
3	100

- Временные ряды: объект фиксирован, зависимая переменная изменяется во времени.

t	Y
1	100
2	40
3	150

Типы данных

- Панельные данные: зависимая переменная изменяется как по времени, так и по объектам.

N	t	Y
1	1	100
1	2	120
2	1	150
2	2	152