монета

МАТБАЗА

Курс по основам теории вероятностей, математической статистики и теории информации

Игорь Нехаев

Специалист по анализу данных
и машинному обучению

Модель. Параметры. Параметрическая Модель

Тезис 1. Чтобы управлять процессом (т.е. вести процесс к нужному результату) нужна модель процесса и информация.

Модель процесса – это его описание, способствующее решению Задачи управления процессом.

Самая простая Модель процесса – это **Параметрическая Модель** (основные параметры, описывающие поведение системы, и множества возможных значений данных параметров, получаемых после измерения).

M = {(v[i], V[i]), i=1..nv}, где
v[i], V[i] - это обозначение i-го параметра (var) и множества его возможных значений (Values)

Параметрическая модель

Пример «Процесс знакомства с девушкой в парке»:

параметр v1="Ответ" принимает значения:
“презрительный взгляд”, “я не знакомлюсь на улице”, “а меня - Света, а вот мой парень идет”, “а меня - Лена, спасибо за мороженое, сегодня действительно жарковато”;

параметр “Потраченное Время”: 1 мин. - 5 часов.

параметр “Результат”: “Получил отлуп”,
“Съела мороженое и убежала”,
“Познакомились, оставила телефончик”;

Параметрическая модель

Пример «Процесс подключения клиента к системе интернет-платежей»:

Клиент: client_id; 20-значный хеш-код;
Сайт: адрес подключаемого интернет-магазина; допустимый интернет-адрес;
Статус процедуры проверки MRM: “в работе MRM”, “отказано”, “согласовано”;
Статус процедуры подключения CRM: не заполнен ЛК, подтверждено (требуется открыть счет), счет подключен, транзачит,
время, затраченное на подключение клиента к счету;
Другие параметры: дата регистрации заявки, ФИО клиента, название юр.лица, …

Виды данных и параметров

Параметры, в зависимости от принимаемых значений делятся на:

- категориальные или номинальные параметры – те, которые после измерения принимают текстовые значения; например, переменная “вид_клиента” может принимать значения СЗ, ИП, ЧП, ГБ,...

- порядковые или упорядоченные (ординальные) параметры; упорядоченными называются категориальные данные, в которых для категорий может быть установлено отношение порядка, а расстояния между категориями не определены; например, переменная “уровень подключения клиента” может принимать значения, выстроенные в порядке “проверка клиента”, “заполнение ЛК”, “подключение счета”, “транзачит”.

- числовые параметры – параметры с числовыми, интервальными данными; для этих данных определено как отношение порядка, так и расстояние между ними; соответствующие параметры, при измерении которых фиксируются данные этого вида, называются числовыми.

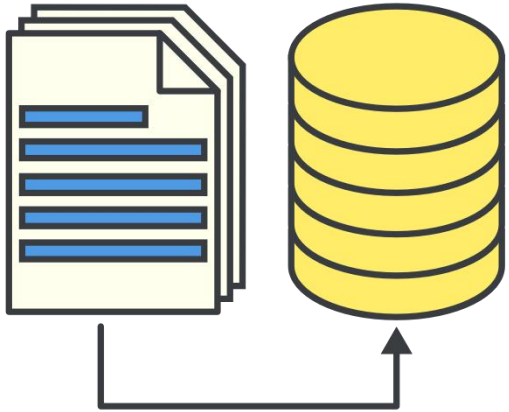
Модель Данных

Следующий уровень модели - это Модель Данных.

Это когда вы к параметрической модели добавили фактические каналы измерения параметров и добавили наблюдаемые данные.

$$o[i]: process \rightarrow V[i], i=1..nv$$

В зависимости от канала наблюдения один и тот же параметр может быть и числовым и, например, порядковым. (Рост - высокий, выше среднего, средний, ниже среднего, …).



Мы будем начинать работу в курсе с этого уровня моделей, т.е. когда определены параметры модели и данные.

Данные и информация. Порождающая Модель

Тезис 2. Чтобы данные стали информацией (чтобы извлечь из данных информацию) необходима Модель наблюдаемого процесса.

Вы получили данные. Что они означают?
Какую информацию несут в себе?
Например, вы узнали, что оборот клиента за месяц составил 100 тыс.руб. Это много или это мало? Это падение или это прирост? Это аномалия или норма?

А вот для этого нужно построить следующий уровень Модели – порождающие Модели, которые содержат знания о некоторых инвариантных характеристиках параметров и отношений между параметрами. Например, средний рост человека, нормальное отклонение веса для данного роста и т.п. Тогда мы и сможем найти ответы на поставленные выше вопросы.

Неопределенность в данных. Виды неопределенностей

Тезис 3. Данные о процессе мы получаем с помощью неидеальных каналов наблюдения.
Наблюдаемое значение параметра отличается от истинного измеряемого свойства объекта.

Пример канала измерения роста. При измерении роста, мы округляем его значение с точностью до 1 см. Кроме того, результат измерения может быть разным утром и вечером.

Поэтому очень важно при определении Модели данных указать не только параметры и множества возможных значений, но и точно описать канал измерения/наблюдения для каждого параметра/признака Модели. Как он измеряется? С помощью какого инструмента? В какие моменты времени? В каких условиях?

Пример канала измерения

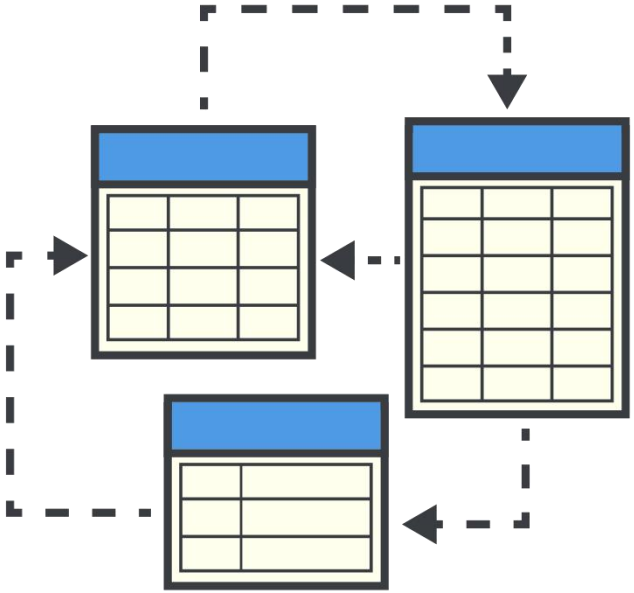
Измерение объема ежедневного/ежемесячного дохода клиента, полученного с использованием подключенного интернет-магазина.

Казалось бы, что здесь все прозрачно, мы имеем все данные поступлений. Но оказывается все не так просто. Поступающие суммы дробятся, разделяются между несколькими акторами, возможны какие-то отложенные платежи… И здесь также будет важно четко описать алгоритм, по которому мы будем оценивать доход клиента.

Несовершенство используемых Моделей

Тезис 4. Модель далеко не всегда хорошо описывает процесс - учитываемые и наблюдаемые параметры не могут полностью объяснить/описать ход/свойства/результаты процесса.

Это может происходить и из-за того, что ряд важных факторов процесса мы не контролируем и не можем наблюдать и из-за того, что мы не знаем какие параметры надо учитывать и измерять, для того, чтобы модель была бы точнее.



Неопределенность в объяснении полученных Данных

Пример “Процесс знакомства с девушкой в парке”.
куча неконтролируемых факторов: настроение, состояние, есть ли парень, отношения с ним в данный момент, мотивация для знакомств;

Пример “Процесс подключения клиента к системе интернет-платежей”.
неконтролируемые факторы: мотивация клиента, внешние обстоятельства;

Управление таким процессом затруднено из-за того, что мы
не понимаем или не можем построить хороших инвариантов, моделей, порождающих этих данных.
На помощь нам приходит теория вероятности и вероятностные модели, которые могут помочь нам описать нашу степень незнания/непонимания/ошибки в описании процесса.

Описание неопределенности с помощью вероятностей

Пример “Процесс знакомства с девушкой в парке”.
Мы завели блокнотик, в который вносим наши результаты попыток знакомств по воскресеньям в парке: отлуп, отлуп, убежала, убежала, телефончик, убежала, телефончик, отлуп, отлуп, убежала, отлуп, телефончик, отлуп, телефончик, …;

Через полгода мы подвели итоги:

Отлуп : 55 раз	Убежала: 21 раз	Телефончик оставила: 24 раз.

Через год мы подвели итоги:

Отлуп : 100 раз	Убежала: 50 раз	Телефончик оставила: 50 раз.

Можно сделать вывод, что вероятность знакомства с Вами составляет 0,24 по итогам полугода и 0,25 по итогам года. Вероятность исхода можно считать некоторой мерой нашей уверенности в данном исходе.

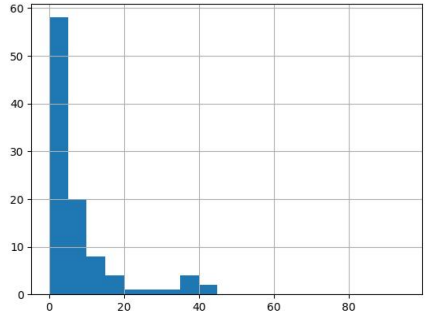
Пример “Процесс подключения клиента к системе интернет-платежей”

Рассмотрим количество дней, которое прошло с момента поступления заявки на подключение и момента подключения магазина клиента к его счету. Вот совсем свежие данные за 8 месяцев этого года:
[27. 37. 6. 6. 6. 7. 32. 4. 3. 2. 1. 11. 0. 16. 1. 8. 2. 2. 2. 6. 2. 1. 3. 2. 15. 15. 0. 1. 0. 0. 0. 4. 16. 9. 10. 11. 3. 40. 1. 1. 1. 1. 0. 1. … 1. 5. 1. 1. 7. 16. 3. 3. 2. 2. 2. 1. 9. 6. 3. 7. 0. 0. 1. 1.]

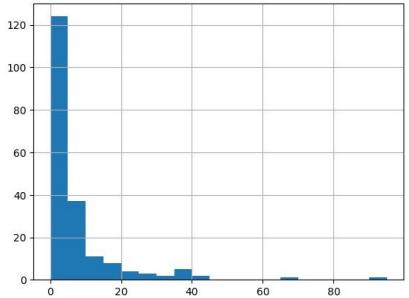
При этом поступило всего 858 заявок. Из них 198 были доведены до подключения. При этом из них только 121 подключение стало активно (пошли транзакции). Доля подключенных заявок составила **23.1%**.

Если рассмотреть отдельно первую половину заявок и отдельно вторую половину заявок, то этот показатель станет равным: **24%** и **22.1%** соответственно.

Возможно, что летом активность клиентов по заполнению ЛК немного упала, а может это просто случайные колебания. А пока мы будем считать, что это достаточно устойчивый показатель нашего взаимодействия с клиентами. И мы можем говорить, что вероятность того, что произвольный заявка будет доведена до подключения составляет 0,23.



Время подключения клиентов (четные дни)



Время подключения клиентов (все дни)

Вероятность исхода

Под **вероятностью какого-либо исхода** (исход = значение параметра) будем понимать предел относительной частоты наступления данного исхода в серии бесконечного числа испытаний.
Следствие: 0 <= p(исход) <= 1.

Если у нас нет никаких данных, то все исходы будем считать равновероятными.
(Отсюда шутка: Какова вероятность встретить бородатую женщину на Ямайке? Ясное дело ½ - либо встретишь, либо нет).

Закон распределения вероятностей

Отображение, которое каждому возможному исходу (значению параметра) ставит в соответствие его вероятность, называется **Законом распределения вероятностей.**
Условие нормировки: сумма вероятностей всех исходов равна 1.

Пример “Знакомство …”
Отлуп : 0.5;
Убежала: 0.25;
Телефончик оставила: 0.25.

Пример “Подключение …”
подключение клиента: 0,23
неподключение: 0,77

Пока мы считаем наши параметры принимающими дискретные значения (категориальные, порядковые или числовые, но дискретные). С непрерывными разберемся позже.

Summary

- Работа с данными начинается с построения Параметрической Модели;
- Для получения Данных - значений параметров в каждом «эксперименте» определяется канал наблюдения для каждого параметра;
- Сами данные не дают нам информации, если у нас нет порождающей их Модели; порождающая Модель дает нам важные инварианты в описании Данных;
- Источниками неопределенности в новых данных являются неидеальность Модели и Каналов наблюдений.

Спасибо за внимание!

Вопросы?