Математические основы искусственного интеллекта. Генеральная совокупность и выборка

Солодушкин Святослав Игоревич

Кафедра вычислительной математики и компьютерных наук,
УрФУ имени первого Президента России Б.Н. Ельцина

Февраль 2022

Что такое «Прикладная статистика»

Прикладная статистика — раздел математики, в котором разрабатываются методы регистрации, описания и анализа данных наблюдений и экспериментов с целью построения вероятностных моделей массовых случайных явлений.

Предметом прикладной статистики является изучение закономерностей, которым подчиняются массовые случайные явления и процессы, с применением методов теории вероятностей.

Задачи прикладной статистики

Задачи статистики

1. указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или специально поставленных экспериментов;
2. разработать методы анализа статистических данных в зависимости от целей исследования: оценка неизвестной вероятности события, оценка неизвестной функции распределения, оценка параметров распределения, оценка зависимости случайной величины от одной или нескольких случайных величин и т. д.

Итак, задача прикладной статистики заключается в разработке методов сбора и обработки статистических данных для получения научных и практических выводов.

Основные понятия выборочного метода

Основные понятия выборочного метода: генеральная совокупность и выборка.

Пусть требуется изучить, как в совокупности однородных объектов распределен некоторый качественный или количественный признак, характеризующий эти объекты.

Определение

Выборочной совокупностью, или просто *выборкой*, называют совокупность случайно отобранных объектов.

При этом выборку осуществляют из генеральной совокупности.

Определение

Генеральной совокупностью называют совокупность всех объектов, относительно которых предполагается делать выводы при изучении конкретной задачи.

Вопрос отбора объектов из генеральной совокупности не является тривиальным. От способа организации этого отбора зависит качество выборки.

Для того чтобы по данным выборки можно было достаточно уверенно судить об интересующем признаке генеральной совокупности, необходимо, чтобы объекты выборки правильно его представляли. Другими словами, выборка должна правильно представлять пропорции генеральной совокупности. Это требование коротко формулируют так: выборка должна быть репрезентативной (представительной).

Репрезентативность выборки

В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществить случайно: каждый объект выборки отобран из генеральной совокупности случайно, т. е. никаким объектам при отборе не отдается предпочтение.

Репрезентативность выборки

Одним из наиболее известных исторических примеров нерепрезентативной выборки считается случай, происшедший во время президентских выборов в США в 1936 г. Журнал «Литерари Дайджест», успешно прогнозировавший события нескольких предшествующих выборов, ошибся в своих предсказаниях, разослав 10 млн пробных бюллетеней своим подписчикам, а также людям, выбранным по телефонным книгам всей страны, и людям из регистрационных списков автомобилей. В 25 % вернувшихся бюллетеней (почти 2,5 млн) голоса были распределены следующим образом:

1. 57 % отдавали предпочтение кандидату-республиканцу А. Лэндону;
2. 40 % выбрали действующего в то время президента-демократа Ф. Рузвельта.

Репрезентативность выборки

На выборах же победил Рузвельт, набрав более 60 % голосов. Ошибка «Литерари Дайджест» заключалась в следующем: желая увеличить репрезентативность выборки, работники журнала, которым было известно, что большинство их подписчиков считают себя республиканцами, расширили выборку за счет людей, выбранных из телефонных книг и регистрационных списков. Однако они не учли современных реалий и набрали еще больше республиканцев: во время Великой депрессии обладать телефонами и автомобилями могли себе позволить в основном представители среднего и высшего класса (т. е. большинство республиканцев, а не демократов).

Репрезентативность выборки

Одна и та же выборка может рассматриваться как репрезентативная и как нерепрезентативная в зависимости от того, на какую генеральную совокупность исследователь желает распространить свои выводы.

Выборка составлена по результатам периодического медицинского осмотра работников Богословского алюминиевого завода (выявление бронхолегочной патологии). Но если ставится задача исследования структуры бронхолегочной патологии жителей Свердловской области, то такую выборку следует считать нерепрезентативной. Однако при исследовании структуры бронхолегочной патологии работников алюминиевого производства в Российской Федерации та же самая выборка может считаться репрезентативной.

Генеральная совокупность

Теперь проведем формализацию понятий.

Пусть проводятся наблюдения за случайной величиной X, распределение которой нам частично или полностью неизвестно. В математической статистике принято следующее определение:

Определение

Генеральной совокупностью случайной величины X (или просто генеральной совокупностью X) называется множество возможных значений случайной величины X. Законом распределения (распределением) генеральной совокупности X называется закон распределения случайной величины X.

Выборка

Исходным материалом для изучения свойств генеральной совокупности (т. е. некоторой случайной величины) являются экспериментальные (статистические) данные, под которыми понимают значения случайной величины, полученные в результате повторений случайного эксперимента (наблюдений за случайной величиной).

Предполагается, что эксперимент хотя бы теоретически может быть повторен сколько угодно раз в одних и тех же условиях. Под словами «в одних и тех же условиях» будем понимать, что распределение случайной величины X_i, $i = 1, 2, \ldots$, заданной на множестве исходов i-го эксперимента, не зависит от номера испытания и совпадает с распределением генеральной совокупности X. В этом случае принято говорить о независимых повторных экспериментах (испытаниях) или о независимых повторных наблюдениях над случайной величиной.

Случайная выборка и выборка

Определение

Случайной выборкой $\overrightarrow{X_n} = (X_1, \ldots, X_n)$ объема n из генеральной совокупности X называется набор из n независимых случайных величин $X_1, \ldots, X_n$, каждая из которых имеет то же распределение, что и случайная величина X.

Очевидно, что случайная выборка — объект абстрактный, в эксперименте не наблюдаемый.

Определение

Выборкой $\overrightarrow{x_n} = (x_1, \ldots, x_n)$ из генеральной совокупности X называется любое возможное значение случайной выборки $\overrightarrow{X_n}$.

Шкалы измерений в статистике

Грамотное использование статистических методов обработки данных во многом зависит от четкого понимания исследователем того, как интерпретировать числа, внесенные в базу данных.

Иинтерпретация данных, внесенных в базу, должна проводиться в соответствии с тем, к какой шкале данные были отнесены. Выделяются четыре вида шкал.

Номинальная шкала

Проводится клиническое исследование и в базу данных внесены сведения о поле пациентов. Для удобства работы можно использовать кодировку: 0 — мужской пол, 1 — женский.

Очевидно, что обозначение цифрами 0 и 1 соответственно лиц мужского и женского пола абсолютно произвольно, цифры можно было поменять местами, а можно для кодирования использовать другие цифры, например, 1 и 2.

Такая же ситуация и с переменной «раса». Пусть в базе она имеет четыре значения: европеоидная раса — 1, негроидная — 2, восточноафриканская — 3, монголоидная — 4.

Номинальная шкала

Возможности обработки переменных, относящихся к номинальной шкале, очень ограничены.

Можно провести только частотный анализ таких переменных. К примеру, расчет среднего значения для расы совершенно бессмыслен.

Переменные, относящиеся к номинальной шкале, часто используются для группировки, с помощью которой совокупная выборка разбивается по категориям этих переменных. В частичных выборках проводятся одинаковые статистические тесты, результаты которых затем сравниваются друг с другом.

Порядковая шкала

Классификация хронической болезни почек

Код	Стадия ХБП	Признаки	СКФ, мл/мин/1.73 м^2
1	1	Признаки нефропатии	> 90
2	2	Признаки нефропатии	60 — 89
3	3А	Умеренное снижение СКФ	45 — 59
4	3Б	Выраженное снижение СКФ	30 — 44
5	4	Тяжелое снижение СКФ	15 — 29
6	5	Терминальная почечная недостаточность	< 15

Классическим примером, где переменная относится к порядковой шкале, являются всевозможные рейтинги, а также места, занятые участниками соревнований.

Порядковая шкала

Кроме частотного анализа, переменные с порядковой шкалой допускают также вычисление определенных статистических характеристик, таких как медианы.

Если должна быть исследована связь с другими переменными такого рода, то для этой цели можно использовать коэффициент ранговой корреляции Спирмена или Кендалла.

Для сравнения различных выборок переменных, относящихся к порядковой шкале, могут применяться непараметрические тесты, формулы которых оперируют рангами.

Интервальная шкала

Рассмотрим коэффициент интеллекта IQ.

Его абсолютные значения отображают порядковое отношение между респондентами, и разница между двумя значениями также имеет содержательный смысл. Например, если у Владимира IQ равен 90, у Ивана — 120, а у Владислава — 150, можно сказать, что Иван настолько же интеллектуальнее Владимира, насколько Владислав интеллектуальнее Ивана (а именно на 30 единиц).

Однако тот факт, что у Владислава значение IQ в 1.25 раза больше, чем у Ивана, не позволяет на основании определения IQ сделать вывод, что Владислав на 25 % умнее Ивана.

Шкала отношений

Примером переменной, относящейся к такой шкале, является возраст: так, если Владиславу 25 лет, а Ивану 50, можно сказать, что Владислав вдвое младше Ивана. Шкала, к которой относятся данные, называется шкалой отношений.

Эта шкала включает все интервальные переменные, которые имеют абсолютную нулевую точку. Поэтому переменные, относящиеся к интервальной шкале, как правило, имеют и шкалу отношений.

С помощью таких шкал могут быть измерены масса, длина, концентрация. Шкала Кельвина (температуры, отсчитанные от абсолютного нуля, с выбранной по соглашению специалистов единицей измерения Кельвин) является примером шкалы отношений.

Шкалы измерений в статистике

1. Номинальная шкала. Числа, хранимые в базе данных, являются условным кодом, например, чистота кредитной истории (1 — есть случаи невозврата кредитов, 0 — нет таких случаев).
2. Порядковая шкала. Числа, хранимые в базе данных, выражают степень развития признака, например, уровень компетенций сотрудника (1 — junior, 2 — middle, 3 — senior).
3. Интервальная шкала. Числа, хранимые в базе данных, характеризуют физическую и/или экономическую величину в единицах ее измерения, при этом можно оценивать, на сколько одно значение больше другого, но нельзя оценивать во сколько раз одно значение больше другого (например, температура тела в градусах Цельсия, уровень IQ в баллах).
4. Шкала отношений. Числа, хранимые в базе данных, характеризуют физическую и/или экономическую величину в единицах ее измерения, при этом можно оценивать, во сколько раз одно значение больше другого (например, стоимость в рублях, длина в метрах, вес в килограммах).

Шкалы измерений в статистике

Типы шкал и их свойства согласно классификации Стэнли Смита Стивенса					
		Номинальная шкала	**Порядковая** шкала	**Интервальная** шкала	шкала **Отношений**
Логические/ математические операции	× ÷	✗	✗	✗	✓
	+ -	✗	✗	✓	✓
	< >	✗	✓	✓	✓
	= ≠	✓	✓	✓	✓

Работа с пропущенными значениями

На практике в реальных данных очень часто встречаются пропуски (англ. *missing data*).
Например, при проведении клинических исследований некоторым пациентам не назначают анализы.
При проведении социологических опросов респонденты могут отказаться отвечать на некоторые вопросы. Причинами пропусков могут быть ошибки ввода данных, утеря или сокрытие информации.

Работа с пропущенными значениями: пример

Гемостаз после эндопротезирования тазобедренного сустава наступает обычно на вторые сутки, а потому важным лабораторным показателем, позволяющим оценить состояние пациента в послеоперационный период, является уровень гемоглобина на вторые сутки.

Если объем кровопотери низкий и пациент чувствует себя хорошо, данный анализ иногда не проводят. Соответственно, *ничего* неизвестно про уровень гемоглобина у этих пациентов.

Работа с пропущенными значениями: пример

Исследователь рассуждает: «*Ничего* — это, как мы знаем со школы, ноль» — и ставит нули во всех ячейках, где не было данных о гемоглобине на вторые сутки.

При проведении расчетов алгоритмы, встроенные в пакеты программ, воспринимают эти нули как обычные данные, и все оценки получаются смещенными, а статистические выводы — неверными.

В частности, на основе таких неверных данных можно прийти к выводу, что риски развития послеоперационных осложнений не связаны с уровенем гемоглобина на вторые сутки.

Работа с пропущенными значениями

Как же обрабатывать пропущенные значения? Методы обработки зависят от типа пропусков:

1. полностью случайные пропуски,
2. случайные пропуски,
3. неслучайные пропуски.

Rubin D. Inference and Missing Data // Biometrika. 1976. No 3. P. 581–592.

Missing Completely at Random

1. MCAR (Missing Completely at Random). Полностью случайные пропуски имеют место в тех случаях, когда подвыборка значений по переменной(-ым), подлежащей изучению, по-прежнему является моделью генеральной совокупности.

Проводится опрос населения с целью выяснения политических предпочтений. У некоторых респондентов данные о политических предпочтениях пропущены, однако эти данные не зависят от других переменных (например, образование, возраст, уровень доходов и т. д.). Кроме того, вероятность пропуска не зависит от значения самой переменной, т. е. не возникает ситуаций, когда респонденты с определенной политической позицией чаще других не дают ответа на соответствующий вопрос. Причина пропусков случайна: респондент не понял суть вопроса, невнимательно читал анкету и не заметил вопрос.

Missing at Random

2. MAR (Missing at Random). Случайные пропуски имеют место в тех случаях, когда вероятность пропуска зависит от значений других переменных, но не зависит от самих пропущенных значений.

Пожилые респонденты более склонны скрывать свои политические предпочтения, чем молодые люди, но внутри старшей возрастной группы пропуски распределены случайно. Иными словами, скрытность пожилых респондентов не зависит от их политической ориентации (консерваторы, либералы и т. д.), в то же время молодые респонденты открыто заявляют о любых своих политических предпочтениях.

Missing at Random

В этом случае возможно смещение результатов оценивания параметров. Например, если молодые респонденты более либерально настроены, а пожилые — более консервативно (но скрывают это), то оценка среднего значения будет смещаться в сторону либеральных настроений. Если же политические предпочтения от возраста не зависят, то смещения результатов оценивания не произойдет.

Missing Not at Random

Неслучайные пропуски имеют место в тех случаях, когда вероятность пропуска зависит от самих пропущенных значений. Например, люди с левыми политическими взглядами более склонны скрывать свои политические предпочтения. Такие пропуски обязательно вносят систематические ошибки в результаты анализа.

Задания

Известно, что у представителей негроидной расы нормальный уровень креатинина сыворотки крови выше, чем у других рас. Это, в частности, учитывается при определении скорости клубочковой фильтрации.

Исследователь, желая выяснить связь расы, закодированной, как описано на слайде 14, с уровнем креатинина, нашел коэффициент корреляции Пирсона. Коэффициент оказался статистически незначимо отличным от нуля. На основании этого исследователь сделал вывод, что уровень креатинина не связан с расами.

Нет ли в этом анализе ошибки? Если да, то в чем она состоит? Какой альтернативный подход можно предложить?

Задания

Верны ли следующие утверждения о пропущенных данных?

А. Замена пропущенных данных (англ. *imputation*) на сгенерированные данные — это на самом деле просто выдумывание данных для искусственного повышения значимости результатов. Лучше не включать в анализ строки с пропущенными данными, чем проводить искусственные вставки.

Б. Недостающие данные можно заменить на среднее или медианное значение. Это не сделает оценки смещенными, но зато уменьшит стандартную ошибку среднего и повысит мощность используемых критериев.

В. Отсутствие данных на самом деле не проблема, если требуется провести простой тест, такой как Хи-квадрат и t-тест.

Г. Худшее, к чему приводят пропущенные данные, — это уменьшение размера выборки и уменьшение мощности.

Задания

У исследователя есть данные о росте и весе пациентов. Желая изучить связь между уровнем глюкозы натощак и ожирением пациентов, исследователь определяет группы согласно значениям индекса массы тела (ИМТ) (см. таблицу).

Группа	ИМТ	Описание
1	18.5 — 25	Нормальный вес
2	25 — 30	Избыточный вес
3	30 — 35	Ожирение I степени
4	35 — 40	Ожирение II степени
5	Более 40	Ожирение III степени

Не происходит ли потери информации при переходе от объективной величины ИМТ[1] к группам ожирения? Не является ли это примером суррогатной группировки?

[1] Индекс массы тела рассчитывается по формуле: $I = m/h^2$, где m — масса тела в килограммах, h — рост в метрах, и измеряется в кг/м^2.