Математические и инструментальные методы машинного обучения

4. Понятие описательных статистик

Методы структурирования данных

Диаграммы рассеяния

Диаграмма рассеяния (разброса) показывает взаимосвязь между двумя видами связанных данных и подтверждает их зависимость. Такими двумя видами данных могут быть характеристика качества и влияющий на неё фактор, две различных характеристики качества, два фактора, влияющих на одну характеристику качества, и т.д.

Описательные характеристики

Цель описательной (дескриптивной) статистики — обработка эмпирических данных, их систематизация, наглядное представление в форме графиков и таблиц, а также их количественное описание посредством основных статистических показателей. Основные статистические показатели можно разделить на две группы:

- меры среднего уровня и
- меры рассеяния.

Меры среднего уровня

Меры среднего уровня дают усредненную характеристику совокупности объектов по определенному признаку.

- Среднее значение
- Стандартная ошибка
- Стандартное отклонение
- Эксцесс
- Асимметрия
- Интервал
- Минимум
- Максимум
- Счёт
- Медиана
- Мода
- Квантиль
- Математическое ожидание
- Доверительный интервал

Меры рассеяния

Меры рассеяния показывают, насколько хорошо данные значения представляют данную совокупность.

Дисперсия случайной величины

Среднеквадратическое отклонение

Размах вариации

Интерквантильный размах

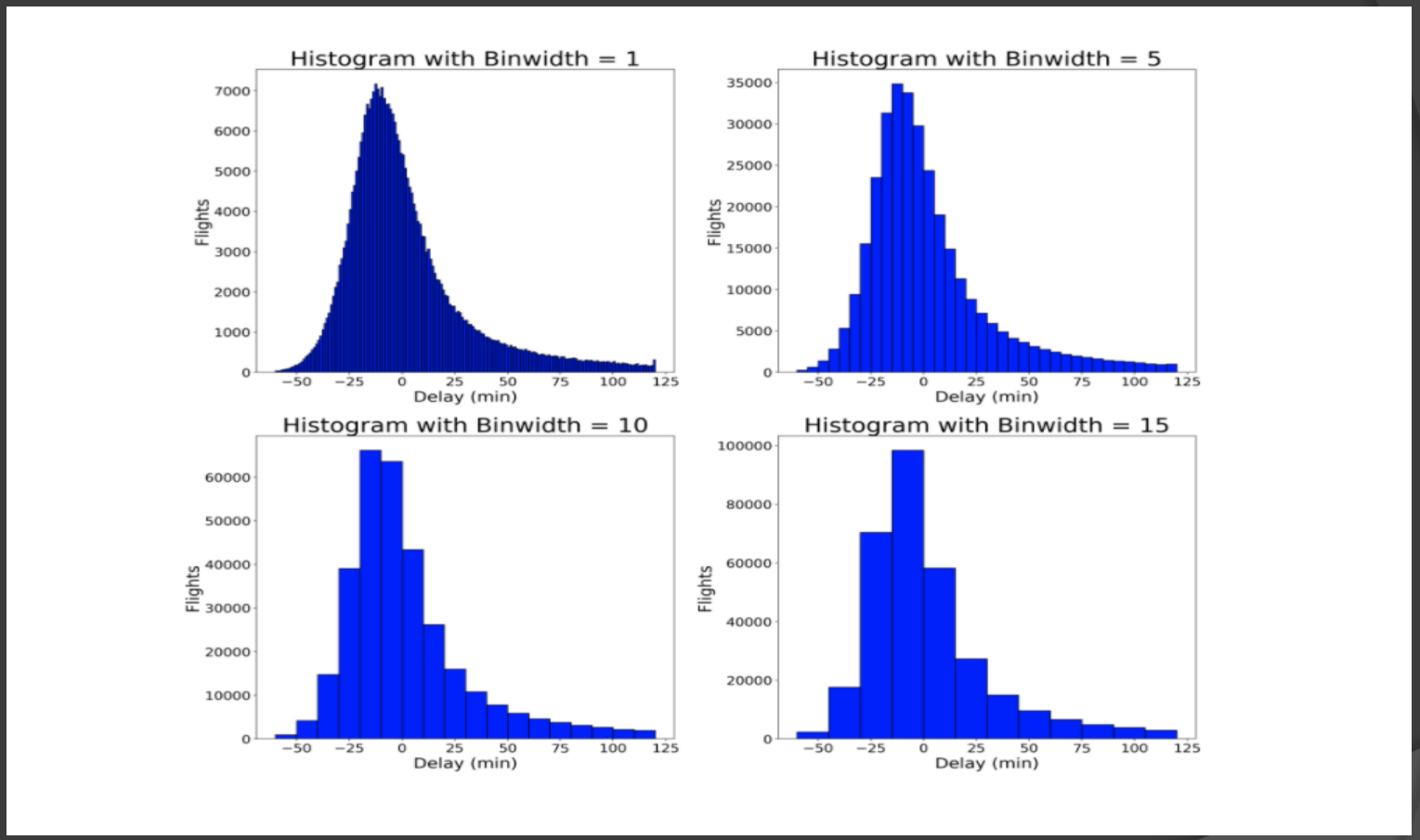
Частотные распределения

Частотное распределение — метод статистического описания данных (измеренных значений, характерных значений). Математически распределение частот является функцией, которая в первую очередь определяет для каждого показателя идеальное значение, так как эта величина обычно уже измерена. Такое распределение можно представить в виде таблицы или графика, моделируя функциональные уравнения. В описательной статистике частота распределения имеет ряд математических функций, которые используются для выравнивания и анализа частотного распределения.

Частотные распределения

В статистике гистограмма — геометрическое изображение эмпирической функции плотности вероятности некоторой случайной величины, построенное по выборке. Гистограмма строится следующим образом. Сначала множество значений, которое может принимать элемент выборки, разбивается на несколько интервалов (bins). Чаще всего эти интервалы берут одинаковыми, но это не является строгим требованием. Эти интервалы откладываются на горизонтальной оси, затем над каждым рисуется прямоугольник.

Гистограммы



Виды анализа статистических связей

Анализ связи между номинальными признаками

- Таблицы сопряженности
- Ранговый корреляционный анализ

Анализ связи между количественными признаками

- Корреляционный анализ
- Регрессионный анализ

Анализ связи между номинальными и количественными признаками

- Номинальный регрессионный анализ
- Дисперсионный анализ (ANOVA)

Кросс-табуляция

Статистический метод, который одновременно характеризует две или больше переменных и заключается в создании таблиц сопряженности признаков, отражающих совместное распределение двух или больше переменных с ограниченным числом категорий или определенными значениями.

Кросс-табуляция представляет собой процесс объединения распределений частот значений двух или больше переменных в одну таблицу. Она объясняет, как одна переменная, например лояльность торговой марке, связана с другой переменной, такой как пол. В таблицах сопряженности признаков показывается совместное распределение значений двух или больше переменных, обладающих ограниченным числом категорий или принимающих определенные значения.

Категории одной переменной помещают в таблицу так, чтобы они размещались в ней (сопрягались) в соответствии с категориями другой или другими несколькими переменными. Таким образом, распределение частот одной переменной подразделяется на группы в зависимости от категорий других переменных.

Кросс-табуляция

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
405	Nikola Pekovic	Minnesota Timberwolves	14.0	C	30.0	6-11	307.0	NaN	12100000.0
302	Boban Marjanovic	San Antonio Spurs	40.0	C	27.0	7-3	290.0	NaN	1200000.0
330	Al Jefferson	Charlotte Hornets	25.0	C	31.0	6-10	289.0	NaN	13500000.0
395	Jusuf Nurkic	Denver Nuggets	23.0	C	21.0	7-0	280.0	NaN	1842000.0
188	Andre Drummond	Detroit Pistons	0.0	C	22.0	6-11	279.0	Connecticut	3272091.0
41	Kevin Seraphin	New York Knicks	1.0	C	26.0	6-10	278.0	NaN	2814000.0
23	Brook Lopez	Brooklyn Nets	11.0	C	28.0	7-0	275.0	Stanford	19689000.0
56	Jahlil Okafor	Philadelphia 76ers	8.0	C	20.0	6-11	275.0	Duke	4582680.0
155	Cristiano Felicio	Chicago Bulls	6.0	PF	23.0	6-10	275.0	NaN	525093.0
176	Timofey Mozgov	Cleveland Cavaliers	20.0	C	29.0	7-1	275.0	NaN	4950000.0

Преимущества кросс-таблиц

Таблица кросс-табуляции состоит из ячеек, в которых приведены комбинации категорий двух переменных. Рассматриваемые данные должны быть качественными или категориальными, поскольку предполагается, что каждая переменная должна измеряться только по номинальной шкале .Таблицами сопряженности широко пользуются при проведении прикладных маркетинговых исследований, поскольку:

Преимущества кросс-таблиц

- менеджеры, которые недостаточно владеют статистическими методами, легко интерпретируют и понимают процедуру кросс-табуляции и ее результаты;
- очевидность трактовки результатов анализа ясно свидетельствует о возможных управленческих действиях;
- ряд операций кросс-табуляции позволяет лучше понять сложное явление, чем это сделал бы один многовариантный анализ;
- кросс-табуляция облегчает проблему разбросанных ячеек, которая затрудняет дискретный многовариантный анализ;
- анализ методом кросс-табуляции прост для выполнения и поэтому обращен к исследователям, менее искушенным в вопросах статистики.

Тест Хи-квадрат

При проведении теста хи-квадрат проверяется взаимная независимость двух переменных таблицы сопряженности и благодаря этому косвенно выясняется зависимость обоих переменных. Две переменные считаются взаимно независимыми, если наблюдаемые частоты (f_0) в ячейках совпадают с ожидаемыми частотами (f_e). Обычно для вычисления критерия хи-квадрат используется формула Пирсона.

Корректность проведения теста хи-квадрат определяется двумя условиями: во-первых, ожидаемые частоты < 5 должны встречаться не более чем в 20 % полей таблицы; во-вторых, суммы по строкам и столбцам всегда должны быть больше нуля.

Тест Хи-квадрат. Описание

Пусть дана случайная величина X .

Гипотеза H_0 : с. в. X подчиняется закону распределения $F(x)$.

Для проверки гипотезы рассмотрим выборку, состоящую из n независимых наблюдений над с.в. X: $X^n = (x_1, \cdots x_n)$, $x_i \in [a,b]$, $\forall i = 1 \cdots n$. По выборке построим эмпирическое распределение $F^*(x)$ с.в X. Сравнение эмпирического $F^*(x)$ и теоретического распределения $F(x)$ (предполагаемого в гипотезе) производится с помощью специально подобранной функции — критерия согласия. Рассмотрим критерий согласия Пирсона (критерий χ^2):

Гипотеза H_0^* : X^n порождается функцией $F^*(x)$.

Разделим [a,b] на k непересекающихся интервалов $(a_i, b_i]$, $i = 1 \cdots k$;

Пусть n_j - количество наблюдений в j-м интервале: $n_j = \sum_{i=1}^{n} [a_j < x_i \le b_j]$;

$p_j = F(b_j) - F(a_j)$ - вероятность попадания наблюдения в j-ый интервал при выполнении гипотезы H_0^*;

$E_j = n p_j$ - ожидаемое число попаданий в j-ый интервал;

Статистика: $\chi^2 = \sum_{j=1}^{k} \frac{(n_j - E_j)^2}{E_j} \sim \chi^2_{k-1}$ - Распределение хи-квадрат с k-1 степенью свободы.

Тесты Стьюдента

Т-критерий Стьюдента — общее название для статистических тестов, в которых статистика критерия имеет распределение Стьюдента. Наиболее часто t-критерии применяются для проверки равенства средних значений в двух выборках. Нулевая гипотеза предполагает, что средние равны (отрицание этого предположения называют гипотезой сдвига).

Все разновидности критерия Стьюдента являются параметрическими и основаны на дополнительном предположении о нормальности выборки данных. Поэтому перед применением критерия Стьюдента рекомендуется выполнить проверку нормальности. Если гипотеза нормальности отвергается, можно проверить другие распределения, если и они не подходят, то следует воспользоваться непараметрическими статистическими тестами.

Тесты Стьюдента. Сравнение двух выборочных средних при известных дисперсиях

Заданы две выборки $x^m = (x_1, \ldots, x_m),\ x_i \in \mathbb{R};\quad y^n = (y_1, \ldots, y_n),\ y_i \in \mathbb{R}.$

Дополнительные предположения:

- обе выборки простые и нормальные;
- значения дисперсий σ_x^2, σ_y^2 известны априори; это означает, что дисперсии были оценены заранее не по этим выборкам, а исходя из какой-то другой информации самим выборкам, описан ниже.

Нулевая гипотеза H_0: $\overline{x} = \overline{y}$ (средние в двух выборках равны).

Статистика критерия:

$$z = (\overline{x} - \overline{y}) \left(\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n} \right)^{-1/2}$$

имеет стандартное Нормальное распределение $\mathcal{N}(0,1)$, где

$$\overline{x} = \frac{1}{m} \sum_{i=1}^{m} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$ — выборочные средние.

Критерий (при уровне значимости α):

- против альтернативы H_1: $\overline{x} \neq \overline{y}$
  если $|z| > \Phi_{1-\alpha/2}$, то нулевая гипотеза отвергается;
- против альтернативы H_1': $\overline{x} < \overline{y}$
  если $z < \Phi_\alpha$, то нулевая гипотеза отвергается;
- против альтернативы H_1'': $\overline{x} > \overline{y}$
  если $z > \Phi_{1-\alpha}$, то нулевая гипотеза отвергается;

где Φ_α есть α-квантиль стандартного нормального распределения.

Критерий Манна-Уитни

U-критерий Манна-Уитни – непараметрический статистический критерий, используемый для сравнения двух независимых выборок по уровню какого-либо признака, измеренного количественно.

Метод основан на определении того, достаточно ли мала зона перекрещивающихся значений между двумя вариационными рядами (ранжированным рядом значений параметра в первой выборке и таким же во второй выборке).

Чем меньше значение критерия, тем вероятнее, что различия между значениями параметра в выборках достоверны.

Критерий Манна-Уитни. Расчёт

Заданы две выборки $x^m = (x_1, \ldots, x_m),\ x_i \in \mathbb{R};\quad y^n = (y_1, \ldots, y_n),\ y_i \in \mathbb{R}.$

Дополнительные предположения:

- обе выборки простые, объединённая выборка независима;
- выборки взяты из неизвестных непрерывных распределений $F(x)$ и $G(y)$ соответственно.

Нулевая **гипотеза** H_0: $\mathbb{P}\{x<y\} = 1/2$.

Статистика критерия:

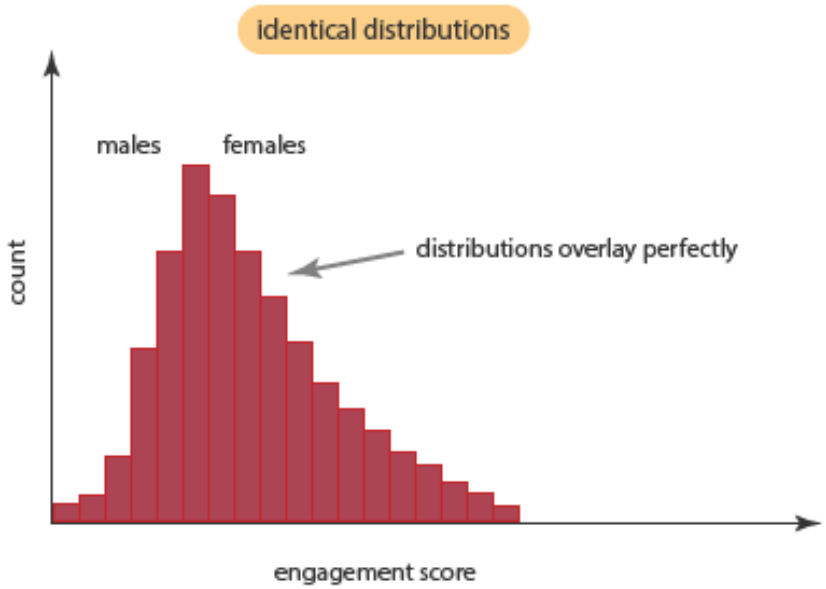
1. Построить общий вариационный ряд объединённой выборки $x^{(1)} \leq \ldots \leq x^{(m+n)}$ и найти ранги $r(x_i),\ r(y_i)$
2. Вычислить суммарные ранги обеих выборок и статистику Манна-Уитни U:

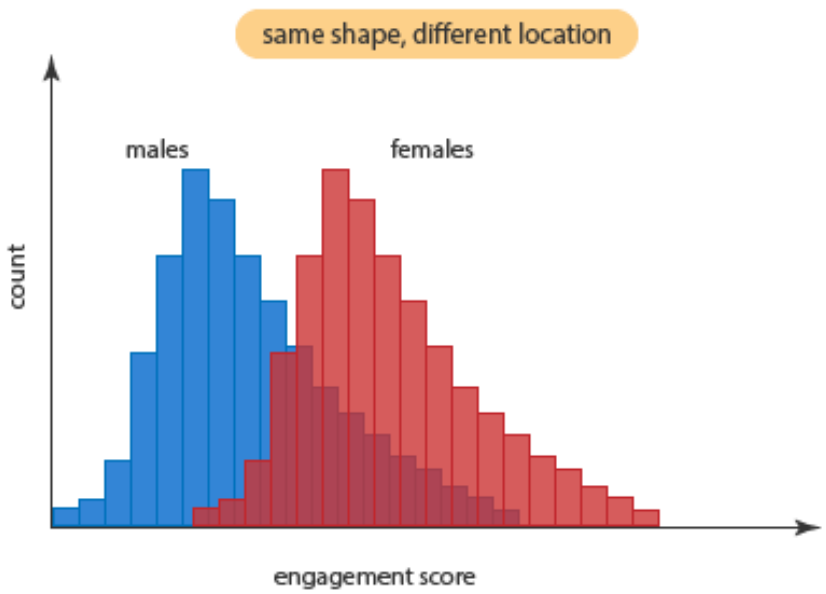
$$R_x = \sum_{i=1}^{m} r(x_i); \qquad U_x = mn + \tfrac{1}{2}m(m+1) - R_x;$$

$$R_y = \sum_{i=1}^{n} r(y_i); \qquad U_y = mn + \tfrac{1}{2}n(n+1) - R_y;$$

$$U = \min\{U_x, U_y\}.$$

Критерий Манна-Уитни. Расчёт





Дисперсионный анализ

Основной целью дисперсионного анализа (ANOVA) является исследование значимости различия между средними с помощью сравнения (анализа) дисперсий. Разделение общей дисперсии на несколько источников, позволяет сравнить дисперсию, вызванную различием между группами, с дисперсией, вызванной внутригрупповой изменчивостью. При истинности нулевой гипотезы (о равенстве средних в нескольких группах наблюдений, выбранных из генеральной совокупности), оценка дисперсии, связанной с внутригрупповой изменчивостью, должна быть близкой к оценке межгрупповой дисперсии. Сравнивая компоненты дисперсии друг с другом посредством F—критерия Фишера, можно определить, какая доля общей вариативности результативного признака обусловлена действием регулируемых факторов.

Дисперсионный анализ

Исходным материалом для дисперсионного анализа служат данные исследования трех и более выборок: , которые могут быть как равными, так и неравными по численности, как связными, так и несвязными. По количеству выявляемых регулируемых факторов дисперсионный анализ может быть однофакторным (при этом изучается влияние одного фактора на результаты эксперимента), двухфакторным (при изучении влияния двух факторов) и многофакторным (позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие). Дисперсионный анализ относится к группе параметрических методов и поэтому его следует применять только тогда, когда доказано, что распределение является нормальным.

Однофакторный дисперсионный анализ

1. Определение независимых и зависимых переменных

2. Разложение полной дисперсии (SS)

$$SS = SS_{between} + SS_{within}$$

3. Изменение эффекта (η^2)

$$\eta^2 = \frac{SS_{between}}{SS}$$

4. Проверка значимости (F)

$$F = \frac{MS_{between}}{MS_{within}}$$

$$SS_{between} = \sum_{t=1}^{k} n_i \cdot (X_i - X)^2$$

$$SS_{within} = \sum_{t=1}^{k} (n_i - 1) \cdot \sigma_i^2$$

Однофакторный дисперсионный анализ