ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования
Омский государственный университет
им. Ф. М. Достоевского

Исторический факультет
Кафедра социологии

Нагорный Олег Станиславович

Методы text mining в социологии

КУРСОВАЯ РАБОТА
СТУДЕНТА 4 КУРСА

Научный руководитель
канд. соц. наук
К.В. Павленко

Омск 2014 г.

Содержание

Введение

В грамм добыча, в годы труды.
Изводишь единого слова ради
Тысячи тонн словесной руды.

В. В. Маяковский

Глава 1

Место text mining в структуре исследовательских методов

1.1 История развития статистических методов

1.1.1 Дуальность статистики

Дуальность статистики берёт своё начало из философского спора Аристотеля и Платона [1, стр. 7]. Аристотель считал, что реальность может быть познана только эмпирически и что исследователь должен тщательно изучать вещественный мир вокруг себя. Он пришёл к убеждению, что можно разложить сложную систему на элементы, детально описать эти элементы, соединить их вместе и, затем, понять целое. Именно таким механистичным путём долгое время следовала наука. Однако в дальнейшем стало понятно, что не всегда целое можно представить как простую сумму частей, его составляющих. Часто, будучи соединёнными вместе, совокупность этих частей приобретает новое качество.

В отличие от своего ученика, Платон считал что свойством подлинного бытия обладают только идеи, а человек может лишь воспринимать и воплощать в вещах их смутные очертания. Для Платона идея (целое) была большим, чем сумма её материальных проявлений.

Эта дихотомия восприятия реальности проявляется во многих аспектах человеческой мысли, в том числе и в сфере статистического знания. С XVIII в. теория статистического вывода развивается в двух основных направлениях: классическом, связанном с именами Дж. Неймана и Е. С. Пирсона, (и являющимся его развитием частотным) и байесовском. Сторонники классического подхода исходят из того, что истинные параметры модели не случайны, а аппроксимирующие их оценки случайны, поскольку они являются функциями наблюдений, содержащих случайный элемент. [2, стр. 5-6] Параметры модели считаются не случайными из-за того, что классическое определение вероятности исходит из предположения равновозможности как объективного свойства изучаемых явлений, основанного на их реальной симметрии [3, стр. 24]. Суждение вида «Вероятность выпадения шестёрки при бросании игрального кубика равняется 1/6» основывается на том, что любая из шести граней при подбрасывании на удачу не имеет реальных преимуществ перед другими, и это не подлежит формальному определению. Таким образом, вероятностью случайного события A в её классическом понимании будет называться отношение числа несовместимых (не могущих произойти одновременно) и равновозможных элементарных событий m к числу всех возможных элементарных событий n:

$$P(A) = \frac{m}{n} \tag{1.1}$$

Однако такое определение наталкивается на некоторые непреодолимые препятствия, связанные с тем, что не все явления подчиняются принципу симметрии. Например, из соображений симметрии невозможно определить вероятность рождения ребёнка определённого пола. Для преодоления подобных трудностей был предложен статистический или частотный способ приближённой оценки неизвестной вероятности случайного события, основанный на длительном наблюдении над проявлением или непроявлением события A при большом числе независимых испытаний и поиске устойчивых закономерностей числа проявлений этого события. Если в результате достаточно многочисленных наблюдений замечено, что частота события A колеблется около некоторой постоянной, то мы скажем, что это событие имеет вероятность. Данные тип вероятности был выражен Р. Мизесом в следующей математической формуле:

$$p = \lim_{x \to \infty} \frac{\mu}{n} \tag{1.2}$$

где μ — количество успешных испытаний, n — количество всех испытаний [3, стр. 46-47].

Совершенно другим взглядом на вероятность отличался байсовский подход, названный так по имени английского математика Томаса Байеса

Литература

1. Nisbet Robert, Elder John, Miner Gary. Handbook of statistical analysis and data mining applications. Academic Press, 2009. с. 864.

2. Зельнер А. Байесовские методы в эконометрии. Москва: «Статистика», 1980.

3. Гнеденко Б. В. Курс теории вероятностей. 8 изд. Москва: Едиториал УРСС, 2005. с. 448.