Методическая разработка

для проведения лекции

Занятие 11. Проверка гипотезы о модели закона распределения. Критерии согласия

Учебные вопросы занятия:

1. Критерий Колмогорова
2. Критерий Смирнова (ω^2)
3. Критерий К. Пирсона (χ^2)

Заключительная часть

Введение

Итак, в параметрических системах закон распределения признаков считают нормальным. Свойство нормальности сильно упрощает процесс анализа. Показатели качества (либо вероятности ошибок) выражаются определенными интегралами, содержащими функции вида $\exp\{-x^2\}$ в различных комбинациях с другими функциями. Поэтому, учитывая простоту решающих функций при нормальности признаков, желательно проверить это свойство на этапе обучения. По обучающей выборке объема k надо выяснить, могла ли она быть получена из совокупности с нормальным законом распределения. Необходимо, чтобы процедура давала всегда однозначный ответ с заданным уровнем значимости ζ. Значение $0<\zeta<1$ равно вероятности того, что в результате проверки мы сочтем обучающую выборку отличной от нормальной, когда она на самом деле извлечена из таковой (вероятность ошибки 2-го рода).

Какой критерий избрать для проверки нормальности?

В настоящее время существует ряд методов для решения этой задачи, однако на практике наибольшее распространение получили методы А.Н. Колмогорова, Н.В. Смирнова (ω^2) и К. Пирсона, отличающееся видом меры рассогласования между статистическим и гипотетическим законами распределения. В методах А.Н. Колмогорова и Н.В. Смирнова такой мерой является функция разности между статической функцией распределения $\tilde{F}(x)$ и функцией распределения гипотетического закона $F(x)$, т.е.

$$d = f[\tilde{F}(x) - F(x)], \qquad (1)$$

а в методе К. Пирсона – функция разности между частотой и вероятностью попадания случайной величины в заданные интервалы, т.е.

$$d = f[\tilde{p}_i(x) - p_i(x)],\ i = \overline{1(1)I}, \qquad (2)$$

где i – номер интервала, I – число интервалов.

1. Критерий Колмогорова

При проверке гипотез о законе распределения по методу Колмогорова в качестве показателя согласованности гипотезы используется случайная величина

$$\hat{u} = \sqrt{k} \max_x |\tilde{F}(x) - F(x)|. \qquad (3)$$

Правило проверки гипотезы о законе распределения заключается в следующем:

1. Назначается уровень значимости ζ, в соответствии с которым определяется критическое значение u_ζ.

2. По результатам наблюдений строится статистическая функция распределения $\tilde{F}(x)$.

3. На этом же графике строится предполагаемая теоретическая функция распределения $F(x)$.

4. По графику определяется максимальная величина модуля разности ординат статистической и теоретической функций распределения и вычисляется значение $\hat{u}$ по формуле (3).

5. Проверяется условие $u > u_\zeta$. Если оно выполняется, то гипотеза о предполагаемом законе распределения отвергается, в противном случае делается вывод, что результаты эксперимента не противоречат гипотезе о том, что наблюдаемая случайная величина $\hat{x}$ подчинена закону распределения с функцией распределения $F(x)$.

Достоинствами метода Колмогорова являются его простота и отсутствие сложных расчетов. Однако он обладает существенными недостатками, а именно:

1. применение метода требует значительной априорной информации о гипотетическом законе распределения, так как кроме вида закона распределения должны быть указаны значения всех параметров распределения;

2. метод учитывает только максимальное отклонение статистической функции распределения от теоретической, а не закон изменения этого отклонения по всему размаху случайной выборки.

В связи с этим при принятии гипотезы может быть допущена ошибка в тех случаях, когда функция распределения сдвинута по оси абсцисс.

2. Критерий Смирнова (ω^2)

При проверке гипотезы о законе распределения по методу Н.В. Смирнова в качестве меры рассогласования теоретического и статистического законов распределения, как и в предыдущем методе, используется функция разности статистической и теоретической функций распределения. Однако в качестве показателя согласованности гипотезы применяется не максимальное значение этой разницы, а среднее значение ее по всей области определения функции

распределения, что исключает недостаток, присущий методу Колмогорова. Суть этого критерия заключается в следующем: по обучающей выборке $\{x_i\}_k$ вычисляется значение ω^2 и далее сравнивается с допустимым значением ω_d^2 (где $d = 1-\zeta$). Если выполняется неравенство $\omega^2 < \omega_d^2$, то гипотеза нормальности принимается с достоверностью $d = 1-\zeta$. В противном случае гипотеза отвергается.

Сравнение вычисленных значений ω^2 с допустимым значением осуществляется табличным, либо графическим методом. В различных литературных источниках (например "Таблица математической статистики" Большев Л.Н., Смирнов Н.В., "Критерии омега-квадрат" Мартынов Г.В.) приведены таблицы и графики допустимых значений ω^2 при нормальных выборках. Рассмотрим последовательность действий при проверке нормальности одномерной выборки.

1. В обучающую систему вводится выборка $\{x_i\}_k = (x_1, x_2, ..., x_k)$.
2. Выборка упорядочивается по возрастанию элемента $x_{(1)} \le x_{(2)} \le ... \le x_{(k)}$.
3. Вычисляются оценки $\tilde{M}$ и $\tilde{\sigma}^2$.
4. Данные оценки подставляются в выражение

$$F(x) = \frac{1}{(2\pi)^{1/2}\tilde{\sigma}} \int_{-\infty}^{x} \exp\left[-\frac{(x-\tilde{M})^2}{2\tilde{\sigma}^2} \right] dx. \qquad (4)$$

5. По формуле (5) вычисляется ω^2:

$$\omega^2 = \sum_{i=1}^{k} \left[F(x_i, \tilde{M}, \tilde{\sigma}^2) - \tilde{F}(x_i) \right]^2 p(x_i), \qquad (5)$$

где $\tilde{F}(x_i)$ - эмпирическая функция распределения случайной величины x_i.

6. По таблице отыскивается значение ω_d^2 для заданной достоверности и сравнивается с ω^2.
7. Если $\omega^2 < \omega_d^2$, то выборка $\{x_i\}_k$ считается нормальной.

Погрешность расчета будет иметь порядок $\left|\Delta\omega^2/\omega^2\right| \approx 1/k^2$. Отсюда следует, что при $k > 10$ погрешность наблюдений не превысит 1%.

Значение показателя согласованности также может рассчитываться по формуле

$$\omega^2 = \frac{1}{2k} + \sum_{i=1}^{k} \left[F\left(x_i, \tilde{M}, \tilde{\sigma}^2\right) - \frac{2i-1}{2k} \right]^2 \qquad (6)$$

При малых объемах выборок (k<40) значение ω_d^2 пересчитывается в соответствии с выражением

$$\left(\omega_d^2\right)' = \left(\omega_d^2 - \frac{0,4}{k} + \frac{0,6}{k^2} \right)\left(1 + \frac{1}{k} \right). \qquad (7)$$

3. Критерий К. Пирсона (χ^2)

В методе К. Пирсона мерой расхождения теоретического и статистического законов распределения служит сумма квадратов разностей между частотой $\tilde{p}(x_i)$ и вероятностью $p(x_i)$ попадания случайной величины $\{x_i\}_k$ в интервалы, на которые разбивается множество возможных значений этой величины. При этом Пирсон показал, что при больших значениях k закон распределения показателя согласованности

$$\hat{u} = \sum_{i=1}^{r} \frac{(\tilde{p}(x_i) - p(x_i))^2}{p(x_i)}, \text{ где } r \text{ - число интервалов,} \qquad (8)$$

обладает весьма важным свойством: он практически зависит не от вида закона распределения случайной величины $\{x_i\}_k$ и объема выборки k, а только от числа интервалов r, причем при увеличении k закон распределения случайной величины $\hat{u}$ приближается к распределению χ^2.

Как известно, распределение χ^2 зависит от числа степеней свободы $v = r - s$, равного числу интервалов минус число независимых условий («связей»), наложенных на частоты $\tilde{p}(x_i)$. Во всех случаях накладывается одно обязательное условие

$$\sum_{i=1}^{r} \tilde{p}(x_i) = 1.$$

В случае, когда теоретическое распределение подбирается так, чтобы совпадали математическое ожидание теоретического распределения и оценка математического ожидания, полученная по результатам наблюдения, число связей увеличивается на единицу. Тогда $s = 2$ и $v = r - 2$.

Если условие совпадения параметров теоретического и статистического законов распределения распространяется и на дисперсию, то $s = 3$ и $v = r - 3$.

Таким образом, число степеней свободы хи-квадрат распределения при проверке гипотез зависит от условий проведения проверки, что необходимо учитывать, используя показатель согласованности.

Порядок проверки гипотезы о виде закона распределения включает:

1. Назначается уровень значимости ζ и по соответствующей таблице определяется критическая граница u_ζ. Входами в таблицу служат уровень значимости ζ и число степеней свободы v.

2. Результаты эксперимента представляются в виде статистического ряда, включающего числа и частоты попадания исследуемой величины $\{x_i\}_k$ в каждый интервал соответственно.

3. Вычисляются вероятности $p(x_i)$ попадания случайной величины $\{x_i\}_k$, следующей гипотетическому закону распределения, в каждый интервал

$$p(x_i) = P(x_{i-1} < \hat{x} < x_{i+1}) = \int_{x_{i-1}}^{x_{i+1}} \omega(x)dx, \qquad (9)$$

где $\omega(x)$ - плотность распределения гипотетического закона.

4. Рассчитывается значение $\hat{u}$ показателя согласованности гипотезы.

5. Проверяется условие $\hat{u} \le u_\zeta$. Если оно выполняется, то расхождение между экспериментальными данными и предполагаемым распределением

полагается несущественным. В противном случае указанное предположение отвергается.

Существенное достоинство метода К. Пирсона состоит в возможности его применения тогда, когда априорно известен лишь вид гипотетического распределения, но неизвестны его параметры. В этом случае параметры распределения заменяются оценками, полученными по экспериментальным данным и используемыми в дальнейшем для вычисления вероятностей $p(x_i)$, а число степеней свободы уменьшается на число заменяемых параметров. Метод К. Пирсона имеет следующие недостатки:

а) он применим только при большой выборке ($k > 100$), так как показатель согласованности следует распределению хи-квадрат лишь при достаточно большом k;

б) результаты проверки в значительной степени зависят от способа разбиения выборки на интервалы, причем их число должно быть не менее десяти, а количество попаданий случайной величины $\{x_i\}_k$ в любой из интервалов – не менее пяти.

Выводы: Обязательным условием применения параметрического метода распознавания является этап проверки нормальности обучающей выборки. Процедура проверки нормальности сводится к вычислению показателя согласованности и сравнению его с допустимым значением по заранее разработанным таблицам и графикам.

Если значение показателя согласованности не превышает допустимое значение, то с заданной достоверностью $d = 1 - \zeta$ выборку можно считать нормальной и использовать параметрические методы обучения. Если нет, используются непараметрические методы обучения.

Заключительная часть.

Подвожу итоги занятия.

Рекомендованная литература:

1. Хемминг Р.В. Численные методы. – М.: Наука, 1972.

2. Левин Б.Р. Теоретические основы статистической радиотехники. – М.: Сов. радио и связь, 1974.

3. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978.