Автоматизация выбора объясняющей переменной логистического бинарного классификатора на основе оценки шансов

В.Г.Подлесных, А.А.Ларионов

НИУ ВШЭ

Обсуждается экспресс-метод выбора лучшей объясняющей (входной, предсказывающей) переменной для логистического бинарного классификатора, основанного на оценке шансов. Метод максимального правдоподобия возможен для числовой объясняющей переменной. В случае дихотомической или полихотомической (интервальной) входной переменной более простым для реализации является метод рекуррентного вычисления нескольких (в зависимости от числа интервалов разбиения объясняющей переменной) фиктивных коэффициентов логистической регрессии с помощью оценки шансов и вероятностей.

Однако в этом случае метод оценки шансов не дает возможности построить ROC-диаграмму, характеризующую точность классификации, так как дает единую оценку теоретической вероятности класса для всех примеров данного интервала, в то время как для построения поточечной диаграммы необходима оценка вероятности каждого примера. В работе показана возможность решения этой проблемы.

**Ключевые слова**: бинарный классификатор, логистическая регрессия, отношение шансов, оценка вероятности, метод максимального правдоподобия, объясняющая переменная.

Классификация: JEL:C38,Z13.

**Введение**.

Существует класс задач бинарной классификации, в которых один из входных показателей определяет принадлежность примера к классу. Например, при оценке вероятности некоторого заболевания основной объясняющей переменной служит возраст пациента. В задаче скоринга при оценке вероятности возвращения кредита класс заемщика (надежный/ненадежный) определяется его доходом. В случае определения класса абонента, который может в ближайшее время отказаться от услуг провайдера связи, таким показателем может служить повышенное число международных звонков. Определяющими (входными) показателями могут быть не только числовые, но и интервальные порядковые переменные. Причем, измерение в интервальной шкале является даже более предпочтительным при экспресс-анализе, поскольку числовую переменную всегда можно представить как интервальную с присвоением интервалу категориальных значений “низкое”, “среднее”, “высокое” и т.д., что будет точнее соответствовать цели классификации, чем использование конкретных чисел. Такие задачи можно решать с помощью пороговой дискриминантной функции - логистической регрессии [1]. Качество классификации можно оценивать по каждой входной переменной отдельно. Оценка не зависит от взаимосвязанности входных переменных, что позволяет не проводить проверку коррелированности, а также предварительный отбор значимых переменных. При этом желательно иметь метод для сравнительной оценки классификаторов по различным входным переменным. Одним из методов такого назначения является построение ROC-диаграмм [2]. Этот метод можно с успехом применять для числовых входных переменных, поскольку для построения диаграмм требуется вычислять теоретическую вероятность класса для каждого примера, что возможно при непрерывной функции логистической регрессии от входной переменной. В этом случае коэффициенты регрессии можно вычислить методом максимального правдоподобия или с использованием шансов. Но входная переменная может быть дихотомической или полихотомической (интервальной). Тогда в уравнении логистической регрессии появляются фиктивные интервальные переменные, которые могут принимать только два значения: 0 или 1. Это делает оценку теоретической вероятности для каждого примера в данном интервале при построении ROC-диаграмм одинаковой, что исключает поточечное построение диаграммы. Можно попытаться оценивать вероятность класса для каждого примера интервала с использованием экспериментальных данных, но для этого нужно обоснование, что и является целью данной работы. Прототипом данной работы явилась работа Н.Б.Паклина [1]. В наших примерах использована та же БД, взятая из приложения на СD “Текучесть абонентской базы”, файл churn.xls., чтобы иметь известный контрольный пример с оценкой коэффициентов логистической регрессии методом максимального правдоподобия и шансов. База данных содержит 3333 примера и 18 входных показателей, из которых 5 были исключены как не оказывающие влияния на выходную переменную “уход ” клиента в другую компанию. Положительным исходом (у = 1) выбран “уход”. В указанной работе для приведенных примеров не строились ROC-диаграммы и не рассматривался вопрос выбора лучшей объясняющей переменной с интервальной шкалой измерения (возможно, из-за нерешенности описанной проблемы).

Работа имеет следующую структуру. В первой части рассматривается метод максимального правдоподобия для числовой входной переменной и метод на основе оценки шансов для дихотомической входной переменной, который гораздо проще реализовать при экспресс-анализе, чем предложенный в [3]. Во второй части рассмотрено вычисление коэффициентов регрессии на основе шансов для полихотомической входной переменной. В третьей части рассмотрены проблемы ошибок 1 и 2 рода и правдоподобия модели, а также особенности построения ROC-диаграмм для интервальной шкалы измерения входной переменной. В четвертой части описан предложенный метод расчета вероятности для каждого примера при интервальной входной переменной и приведены экспериментальные данные.

1. **Логистическая регрессия и методы нахождения оценок ее коэффициентов.**

Уравнение для логистической регрессии имеет вид [1]:

где х – входная (объясняющая) переменная; β0,β1 – искомые параметры. Функция определена на бесконечном интервале и принимает значения в диапазоне [0,1]. Требуется найти наилучшие оценки параметров b0 и b1. Для этого нельзя применить метод наименьших квадратов подобно тому, как это делается в случае обычной линейной регрессии. Поэтому используем метод максимального правдоподобия.

Рассмотрим функцию максимального правдоподобия для общего случая числовой входной и дихотомической выходной переменных.

Пусть вероятность того, что выходная переменная *y* приобретет значение 1 для заданного значения *x* (вероятность успеха), будет  *P*(*y=*1*|x*) = ρ(х), а вероятность того, что *y* = 0 при заданном *x* *P*(*y=*0*|x*) = 1 – ρ(х).

Тогда, при *yi*=0 или 1, вклад *i*-го наблюдения может быть выражен как [1]:

.

Предположение, что примеры выборки являются независимыми, позволяет представить функцию правдоподобия как произведение двух отдельных членов:

.

В вычислительном плане применяется логарифмическая функция правдоподобия

,

Оценки коэффициентов максимального правдоподобия могут быть найдены путем дифференцирования относительно каждого параметра β0 и β1 и приравнивания полученных выражений к 0.

Продифференцируем по и приравняем полученное выражение к нулю:

.

Возьмем производную от первого слагаемого выражения в фигурных скобках:

,

.

Теперь продифференцируем второе слагаемое:

,

,

.

Производная по исходного выражения примет вид:

.

Отбросив знаменатель, который не влияет на правую часть уравнения, получим первое уравнение системы:

.

Теперь продифференцируем по и приравняем полученное выражение к нулю:

.

Дифференцирование по происходит по схеме, аналогичной дифференцированию по , с той разницей, что:

.

Производная по исходного выражения примет вид:

.

Отбросив общий знаменатель, получим второе уравнение системы:

.

Итак, получена система уравнений для нахождения коэффициентов β0 и β1 логистической регрессии общего вида для случая числовой входной и дихотомической выходной переменных:

.

Проверим правильность полученной системы уравнений, используя данные, приведенные в книге Н.Б. Паклина для примера связи между возрастом пациента и вероятностью некоторого заболевания (табл.8.16). За положительный исход примем наличие заболевания yi = 1.

Запишем первое уравнение из приведенной выше системы уравнений в следующем виде:

.

В левой части уравнения суммируются , которые равны единице. В правую часть уравнения подставляются из таблицы значения , которым соответствуют , = 0.

Подставив значения переменных из таблицы и приведенные в книге Н.Б. Паклина значения коэффициентов и для данной выборки, (полагая, что они найдены методом максимального правдоподобия), получим тождество, что свидетельствует о правильности выполненных выше преобразований.

Аналогичная проверка второго уравнения также дает тождество. Однако из полученной системы уравнений оказалось невозможным выразить и аналитическим путем. Решение, совпадающее с данным в [1], получено итерационным численным методом в среде Maple. В случае дихотомической или полихотомической (интервальной) входной переменной этот метод не подходит для задачи выбора лучшей объясняющей переменной при экспресс-анализе. На помощь приходит известный метод вычисления коэффициентов логистической регрессии на основе оценки шансов [1].

Таблица 1. Фрагмент выборки «текучесть абонентской базы»

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Международные звонки | Голосовая почта | Количество голосовых сообщений | Использовано дневных минут | Количество звонков днем | Использовано вечерних минут | Количество звонков вечером | Использование ночных минут | Количество ночных звонков | Минут международных разговоров | Число международных звонков | Число обращений в сервисную службу | Уход |
| Нет | Нет | 0 | 211,60 | 70 | 216,9 | 80 | 153,5 | 60 | 7,8 | 1 | 1 | Нет |
| Нет | Нет | 0 | 178,90 | 101 | 169,1 | 110 | 148,6 | 100 | 13,8 | 3 | 4 | Да |
| Нет | Нет | 0 | 241,80 | 93 | 170,5 | 83 | 295,3 | 104 | 11,8 | 7 | 3 | Нет |
| Нет | Да | 46 | 224,90 | 97 | 188,2 | 84 | 254,6 | 61 | 12,1 | 2 | 0 | Нет |
| Нет | Нет | 0 | 248,60 | 83 | 148,9 | 85 | 172,5 | 109 | 8 | 4 | 3 | Нет |
| Нет | Нет | 0 | 203,40 | 146 | 226,7 | 117 | 152,4 | 105 | 7,3 | 4 | 1 | Нет |

Данные для расчета взяты из файла “Текучесть абонентской базы”, фрагмент которой приведен в таблице 1. Предположим, нужно оценить, как влияет использование голосовой почты на уход клиента в другую компанию. Очевидно, нужны статистические данные, для извлечения которых из данной таблицы разработана программа. Результаты совпали с приведенной в [1] таблицей ухода клиентов в зависимости от дихотомической переменной “ голосовая почта” (таблица 2).

Таблица 2. Сводные данные по уходу клиента

|  |  |  |  |
| --- | --- | --- | --- |
| Уход | Голосовая почта = нет, *x*=0 | Голосовая почта = да, *x*=1 | Всего |
| Нет, *y*=0 | 2008 | 842 | 2850 |
| Да, *y*=1 | 403 | 80 | 483 |
| Всего | 2411 | 922 | 3333 |

В таблице 2 указано, какое количество клиентов, использующих и не использующих голосовую почту, ушло и осталось.

Шанс – это отношение вероятности ρ, что событие (уход) произойдет, к вероятности, что оно не произойдет 1 – ρ. Для бинарной логистической регрессии с дихотомической входной переменной “голосовая почта” шанс ухода клиента, использующего почту ( х = 1) равен О = ρ(1)/(1 – ρ(1)) = 80/842 = 0,095. Шанс ухода клиента, не использующего голосовую почту О = ρ(0)/(1 – ρ(0)) = 403/2008 = 0,201. Но для бинарной логистической регрессии с дихотомической входной переменной теоретический шанс ухода при использовании голосовой почты (х = 1) равен



Теоретический шанс ухода при неиспользовании голосовой почты (х = 0) равен



Тогда теоретическое отношение шансов OR = . Экспериментальное отношение шансов OR = 0,095/0,201 = 0,473. Отсюда exp(β1) = 0,473 и находим оценку β1 = - 0,748, что совпадает со значением в [1], найденным с использованием метода максимального правдоподобия. Оценку β0 можно получить с помощью формулы для вероятности ρ(х), если по таблице 2 оценить экспериментальную вероятность ухода клиента, использующего голосовую почту: ρэкс(х=1) = 80/922=0,087 и приравнять теоретическую и экспериментальную вероятности при известном уже значении β1. Итак



Из этого уравнения нужно найти значение вспомогательной переменной t = β0 – 0,748\*1, решив уравнение относительно et. Получим et = 0,095, откуда t = ln(0,095) = - 2,354. Значит – 2,354 = β0 – 0,748, откуда β0 = - 1,606, что также совпадает со значением в [1]. Таким образом, простая рекуррентная процедура вычисления оценок коэффициентов через шансы и экспериментальные вероятности заменяет метод максимального правдоподобия, но, к сожалению, она не подходит для построения ROC-диаграмм для бинарной входной переменной, где требуется оценка теоретической (по модели) вероятности для каждого примера, а модель дает только общую интервальную оценку. Рассмотрим случай полихотомической входной переменной.

1. **Расчет параметров логистической регрессии на основе оценки шансов для полихотомической (интервальной) входной переменной.**

Рассмотрим эту процедуру для входной переменной “количество обращений в сервисную службу”, полагая, что повышенное число таких обращений увеличивает вероятность ухода (у = 1) клиента. Данная переменная является числовой порядковой и ее значения можно отсортировать по возрастанию с привязкой к выходной переменной “уход” и разбить на интервалы, введя категориальные значения интервалов, например, “низкое”, ”среднее” и “высокое”. Анализ и построение ROC-диаграмм следует проводить для каждого интервала отдельно по отношению к опорному значению, например, “низкое”.

***Примечание****. Если входная переменная является номинальной, например, “цвет”, то возможные её значения можно комбинаторным способом разделять на два подмножества, добиваясь максимальной разности вероятности ухода в подмножествах. Нечто подобное делается в известном алгоритме построения бинарного дерева решений CART [4] с использованием функционального критерия К.Джини. В нашем случае можно использовать критерий площади под ROC-кривой (см. ниже) для выбора наиболее удачного разделения значений категориальной переменной на два подмножества, которые будут содержать значения, лучше других объясняющие разделение примеров на два класса.*

В зависимости от выбранного числа интервалов разбиения n нам потребуется ввести вспомогательные переменные, число которых k = n – 1. При n = 3 потребуется ввести две переменные для квантования значений в трех интервалах. Обозначим их как С1 иС2. В матрице квантования в строке с числом обращений “низкое”, С1 = С2 = 0. В следующей строке с числом обращений “среднее” С1 =1, С2 = 0. В третьей строке с числом обращений “высокое” С1 = С2 = 1. Т.е. в общем случае матица С треугольная. Добавление 1 в каждой следующей строке, как будет видно ниже, позволяет рекуррентно рассчитывать коэффициенты в логит-преобразовании g(х) = β0 + β1С1 +β2С2, которое является показателем степени в самом первом уравнении для вероятности ρ(х). Вероятность ухода клиента оценивается по трем интервалам сводной таблицы (таблица 3), относительно опорной категории “низкое”. Уход означает у = 1.

Таблица 3.Сводные данные

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уход | Низкое | Среднее | Высокое | Всего |
| Нет, у=0 | 1664 | 1057 | 129 | 2850 |
| Да, *y*=1 | 214 | 131 | 138 | 483 |
| Всего | 1878 | 1188 | 267 | 3333 |

Шанс ухода клиента для категории «низкое: 214/1664=0,129.

Шанс ухода клиента для категории «среднее»: 131/1057=0,124.

Шанс ухода клиента для категории «высокое»: 138/129=1,069.

Отношение шансов для категории «среднее» по отношению к категории “низкое”: OR=0,124/0,129 = 0,964. Отношение шансов для категории “высокое” по отношению к категории "низкое" OR=1,069/0,129 = 8,318.

Теоретическая вероятность *ρ*(*х*) ухода клиента для категории «низкое» рассчитывается по формуле:

,

где ,

Функция после подстановки в нее значений вспомогательных переменных для категории «низкое» = = 0 примет вид:

.

Тогда *ρ*(g(*х*)) равняется:

.

Экспериментальную вероятность экс можно найти, поделив число положительных исходов на общее количество исходов,ρэкс =214/1878=0,114. Приравнивая вероятности, получим

.

Отсюда можно найти коэффициент по следующей формуле:

,

После подстановки значений, в данном примере будет равен: -2,05, что совпадает с оценкой в [1].

Для категории «среднее» , а , тогда теоретическая вероятность ухода

, где неизвестно значение β1. Экспериментальная вероятность ухода для категории среднее определяется по таблице 3 как ρэкс = 131/ 1188 = 0,110. Приравнивая эти вероятности, находим β1 по рекуррентной формуле



После подстановки значений в данном примере будет равен: -0,04.

Для категории «высокое» , тогда теоретическая вероятность ухода

где неизвестно только Экспериментальное значение вероятности ухода для категории “высокое” определяется по таблице 3 как ρэкс = 138/267 = 0,512. Приравнивая эти вероятности, найдем следующую рекуррентную формулу для вычисления β2.



После подстановки значений в данном примере будет равен: 2,138.

Полученные изложенным путем коэффициенты логистической регрессии для контрольных примеров с дихотомической и полихотомической входными переменными совпадают со значениями этих коэффициентов, полученными методом максимального правдоподобия в [1]. Это доказывает корректность использования экспериментальных данных для оценки шансов и вероятностей. Именно с этой целью и были взяты в качестве контрольных примеры из книги Н.Б.Паклина. Но полученные формулы для g(x), включают вспомогательные (фиктивные) переменные С1 и С2, и не пригодны для оценки вероятности для каждого примера, так как фиктивные переменные являются интервальными, а для построения ROC-диаграмм нужна поточечная оценка. Для решения этой проблемы предлагается исходить из доказанной экспериментально хорошей сходимости оценок коэффициентов методом правдоподобия и с использованием шансов. Это позволяет для каждого примера оценивать вероятность ухода экспериментально.

Рассмотренная процедура последовательного вычисления коэффициентов логистической регрессии реализована в виде программы с удобным графическим интерфейсом (рис.1) как для дихотомической, так и для полихотомической входной переменной.

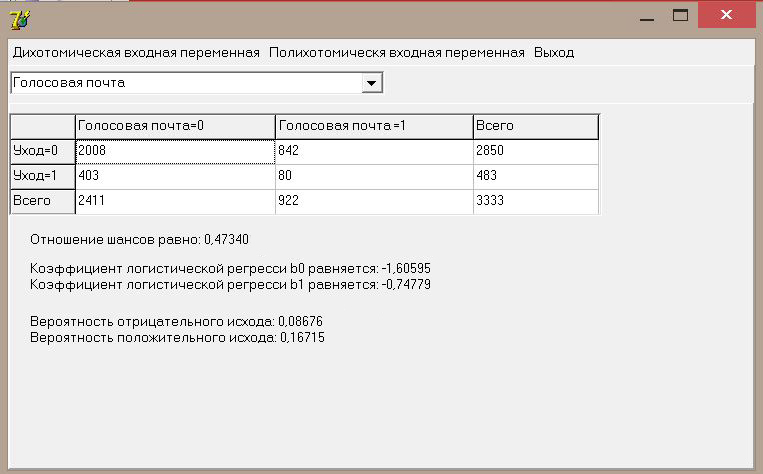


Рисунок 1. Окно расчета параметров логистической регрессии.

1. **Проблемы ошибок модели и особенности построения ROC-диаграмм для интервальной входной переменной.**

При использовании бинарной классификационной модели могут возникать ошибки. Предсказание, соответствующее положительному исходу, может быть распознано моделью как отрицательный исход. Такие ошибки называются ложноотрицательными или ошибками I рода. Предсказание, соответствующее отрицательному исходу, наоборот, может быть распознано моделью как положительный исход. Такие ошибки называются ложноположительными или ошибками II рода. Поэтому в бинарной классификации каждое предсказание может иметь следующие четыре исхода, которые приведены в матрице классификации (табл. 4): истинно положительный TP, истинно отрицательный TN, ложноположительный FP и ложноотрицательный FN.

Таблица 4. Матрица классов

|  |  |  |
| --- | --- | --- |
| Фактически класс | Предсказанный класс | |
| Да | Нет |
| Да, y = 1 | *TP* | *FN (1 род ошибок)* |
| Нет, y = 0 | *FP (2 род ошибок)* | *TN* |

 В таблице 2 под фактическим классом понимается то, что конкретное событие по данным выборки действительно имеет положительный или отрицательный исход. В то же время, предсказанный моделью класс может быть ошибочно указан как FN или FP.

Для построения ROC-диаграмм используются чувствительность и специфичность. Чувствительность (sensitivity, Se) определяется как отношение числа истинно положительных наблюдений к числу фактически положительных наблюдений:

Se=TP/(TP+FN).

Специфичность (specificity, Sp) определяется как отношение числа истинно отрицательных наблюдений к числу фактически отрицательных наблюдений:

Sp=TN/(TN+FP).

ROС-анализ представляет собой графическую методику оценивания эффективности моделей с помощью двух показателей — Se и Sp.

В задачах бинарной классификации, когда модель предсказывает вероятность того, что наблюдение относится к одному из двух классов, важен выбор точки отсечения (cut-of-point) на вероятностной кривой, то есть порога вероятности, разделяющего два класса. Такая точка отсечения показывает, после какого значения вероятности один класс сменяется другим. Выбирая точку отсечения, мы управляем вероятностью правильного распознавания положительных и отрицательных примеров. При уменьшении порога отсечения увеличивается вероятность ошибочного распознавания положительных наблюдений (ложноположительных исходов), а при увеличении возрастает вероятность неправильного распознавания отрицательных наблюдений (ложноотрицательных исходов).

0

100

100

50

50

Se, %

100-Sp, %

.

Рисунок 2. Пример ROC-диаграммы.

Диаграмма строится в координатах Se(%) vs (100 – Sp),%. Для построения ROC-кривой требуется изменять порог отсечения в диапазоне от 0 до 1 с заданным шагом, например 0,01. При каждом значении порога с ним сравнивается значение вероятности для примера (в нашем случае, группы примеров с одинаковым значением входной перемнной), задаваемое моделью. В зависимости от результатов сравнения на каждом шаге пополняются счетчики ошибок Sp или Se. В конце прохода по БД суммируются значения счетчиков, подсчитываются Se и Sp и строится точка на диаграмме. Ясно что, для оценки Se и Sp необходимо иметь модельные значения вероятностей для каждого примера или группы примеров с одинаковыми значениями входной переменной. Модель с фиктивными переменными С1 и С2 не подходит, так как дает общую оценку для интервала.

На рисунке 2 показан пример ROC-диаграммы. Пунктирная линия соответствует идеальному (безошибочному) классификатору. Диагональная линия соответствует бесполезному классификатору (не распознающему классы). Кривая линия – типичный график. Чем дальше удалена кривая от диагонали и больше площадь под ней (параметр AUC), тем выше качество классификации. Рассмотрим процесс построения ROC-кривой более детально.

**4**. **Расчет вероятности положительного исхода для каждого примера при интервальной входной переменной.**

Поскольку входная переменная интервальная, для нее невозможен точный расчет теоретической вероятности по выведенным выше формулам с фиктивными коэффициентами для ρ(х). Однако поскольку выше показано, что значения коэффициентов логистической модели можно с высокой точностью вычислять по экспериментальным оценкам шансов и вероятностям, то это дает основания предложить соответствующий способ вычисления модельных оценок вероятности для каждого примера.

Для этого в программе выполняются следующие действия. Сортируется столбец таблицы данных по выбранной входной переменной с привязкой к ней столбца выходной переменной “уход” =да/нет. Определяются значения границ интервалов “низкое”, “среднее”, “высокое”. ROC-диаграммы будут строиться для каждого из этих интервалов отдельно. Для каждого неповторяющегося значения входной переменной в данном интервале вероятность будет принимать значение 0/1 в зависимости от соответствующего значения выходной переменной “уход”. Эта оценка будет столь же груба, как и интервальная, поэтому одиночные примеры предлагается исключить. В группе повторяющихся значений входной переменной вероятность ухода определяется как отношение числа примеров со значением “уход” = да, к общему числу примеров в данной группе. Полученные значения оценок вероятностей записываются в специальный массив.

Задаем значение точки отсечения (порога). Вероятность для каждой группы примеров интервала сравниваем с пороговым значением. Если вероятность больше порога и пример положительный (“уход”=да), то событие будет распознано, как истинно положительное (*TP*) и пополняет единицей соответствующий счетчик. В случае, если вероятность меньше порога и пример отрицательный (“уход”=нет), событие будет распознано, как истинно отрицательное с пополнением счетчика TN. Если вероятность больше порога, но пример отрицательный, событие будет распознано, как ложноположительное (*FP*) (ошибка 2 рода). В случае если вероятность меньше порога, но пример положительный, событие будет распознано, как ложноотрицательное (*FN*) (ошибка 1 рода).

По полученным значениям *TP*, *TN*, *FN*, *FP* рассчитываются специфичность (*Sp*) и чувствительность (*Se*) для данного примера выборки. На графике откладывается точка с координатами: по оси ординат - *Se*\*100%, по оси абсцисс - (100-*Sp*)\*100%.

При следующем значении порога делается новый проход по БД, выполняются аналогичные действия и на диаграмме строится следующая точка.

Визуальное сравнение интервальных ROC-кривые для различных входных переменных дает качественную оценку классификатора, но их можно сравнивать с помощью подсчета площади под кривой (AUC). AUC =1 соответствует идеальному классификатору, AUC = 0,8 - 0,9 принято считать очень хорошим качество классификации, а AUC=0,5-0,6 соответствует бесполезному классификатору. Площадь под кривой AUC рассчитана в данном случае методом трапеций.

В опытах сравнивали между собой ROC-диаграммы для различных входных переменных: число обращений в сервисную службу, число международных звонков, число ночных, дневных и вечерних звонков, количество голосовых сообщений. Для каждой переменной построено по три ROC-диаграммы для интервалов низкое, среднее, высокое. Лучшей объясняющей переменной и лучшим интервалом оказалось “число международных звонков” и интервал “высокое” (рис.3). Для сравнения приведена диаграмма для той же переменной и интервала “средний” (рис.4).

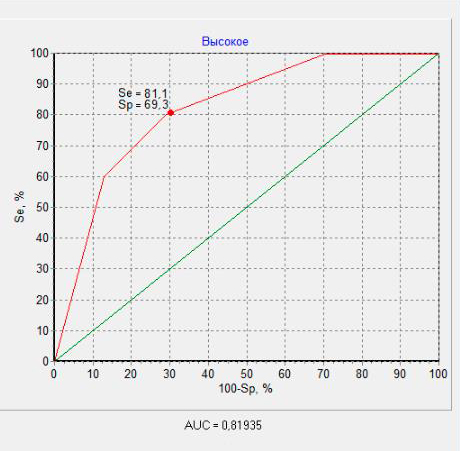
****

Рис.3. ROC-диаграмма для переменной “число международных звонков” и интервала “высокое”. AUC близок к 0,82. Отмеченная точка (наиболее отдаленная от диагонали) имеет минимальное число ошибок 1 и 2 рода.

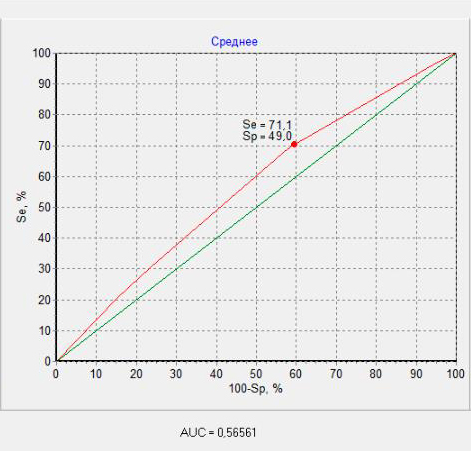


Рис.4. ROC-диаграмма для переменной на рис.3 для интервала “средний”. AUC = 0,56.

Для остальных переменных и интервалов AUC в районе 0,6 – 0,7.

**Заключение**

Сходимость результатов расчета коэффициентов логистической регрессии для бинарного классификатора методом максимального правдоподобия и с использованием шансов показала, что метод шансов не уступает по точности аналитическому методу, но проще в реализации для экспресс-оценок. Он наиболее подходит для числовой входной (объясняющей) переменной, где не представляет сложности поточечное построение ROC-диаграмм с целью оценки правдоподобия классификации.

В случае дихотомической и полихотомической (интервальной) входной переменной предлагается использовать процедуру рекуррентного вычисления фиктивных коэффициентов интервальной логистической регрессии. Однако оценка правдоподобия по шансам затруднена, так как для соответствующего построения ROC-диаграмм требуется оценка вероятности класса для каждого примера, а не интервальная оценка. Показана возможность замены теоретической интервальной оценки вероятности класса на экспериментальную оценку для каждой группы примеров с одинаковым значением входной переменной, кроме одиночных примеров. Для решения этой инженерной задачи разработано соответствующее программное обеспечение.

**Литература**

1. Н.Паклин, В.Орешков.(2010). Бизнес-аналитика: от данных к знаниям. 2-е изд., Питер,410-420, 586-590..
2. Сайт BaseGroup.ru:: Логистическая регрессия и ROC-анализ – математический аппарат.
3. Е.Р.Горяинова, Т.И.Слепнева. Методы бинарной классификации объектов с номинальными показателями. Ж. Новой экономической ассоциации, №2(14),27-49.
4. Сайт ВaseGroup.ru:: Деревья решений – CART математический аппарат.

Automation of explanatory variable choice for logistic binary classifier based on the odds ratio

V.G.Podlesnykh, A.A.Larionov

National Research University “Higher School of Economics”

We have examine the express method for the best choice of dichotomic or polychotomous type explanatory variable in respect to logistic binary classifier based on the odds ratio. We have considered the maximum likelihood method which makes necessary computational solution of derived system of equations in Maple toolkit. This method we can use for numerical explanatory variable. In case of dichotomic or polychotomous variable the more convenient method of sequential computation of several dummy logistic regression coefficients (in depend of explanatory variable intervals number) with help of the odds ratio and estimated probability is used. However in this case the odds ratio method should use one estimated probability in process construction ROC-diagram while as the estimated probability for each case is more accurate. We substantiate that instead of model probability calculation the experimental estimated probability is used for case groups with the same input variable amount. We recommend delete the single cases. It allows improve the classifier estimated likelihood accuracy. We developed software for coefficients logistic regression calculation based on odds ratio and likelihood estimate for each explanatory variable and elected variation interval for it using ROC-diagrams.

**Keywords**: binary classifier, logistic regression, odds ratio, estimated probability,

likelihood method, explanatory variable.

JEL Classification: C38,Z13.

Статья поступила в редакцию 6 сентября 2013 г

Статья опубликована в журнале “Качество Инновации Образование” №1, 2014 г,

с.56-63. (www.quality-journal.ru)