RUDN university

5100

Программирование в среде R

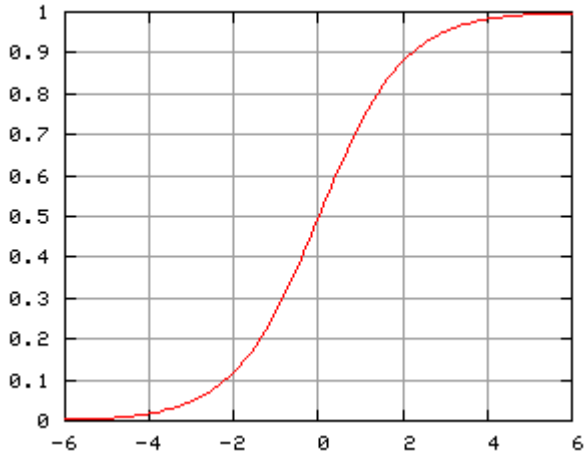
Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

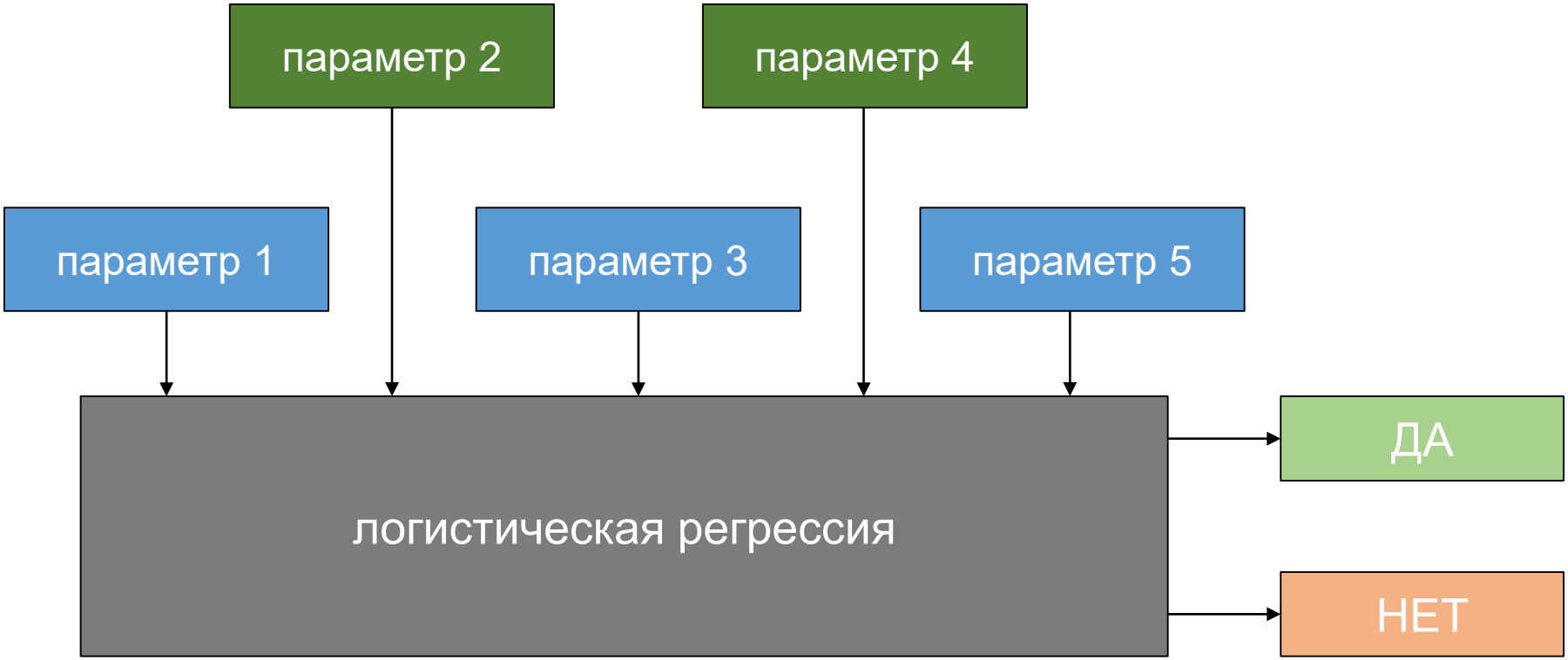
Логистическая регрессия

Определение

Логистическая регрессия или логит-модель (англ. logit model) — это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его сравнения с логистической кривой. Эта регрессия выдаёт ответ в виде вероятности бинарного события (1 или 0).

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится зависимая переменная y, принимающая лишь одно из двух значений — как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами) — вещественных $x_1,x_2,...,x_n$, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной. Как и в случае линейной регрессии, для простоты записи вводится фиктивный признак x_0



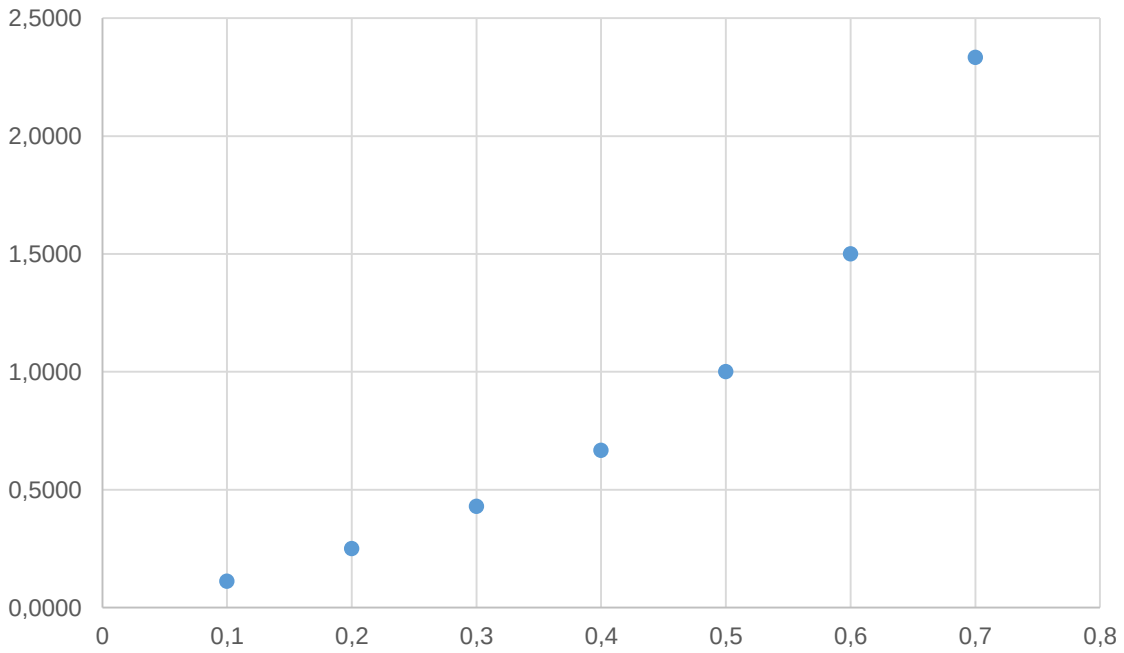


- выдача кредита
- сдача зачета
- принятие на работу
- финансирование проекта

p - вероятность

Отношение шансов

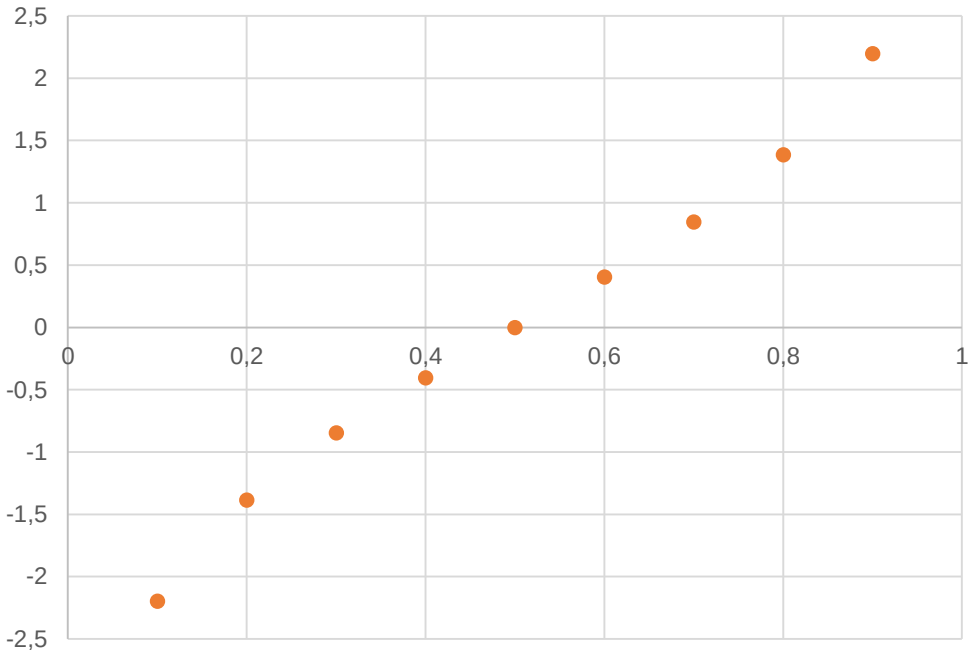
p	odds
0,1	0,1111
0,2	0,2500
0,3	0,4286
0,4	0,6667
0,5	1,0000
0,6	1,5000
0,7	2,3333
0,8	4,0000
0,9	9,0000



odds - отношение шансов: $\frac{p}{1-p}$

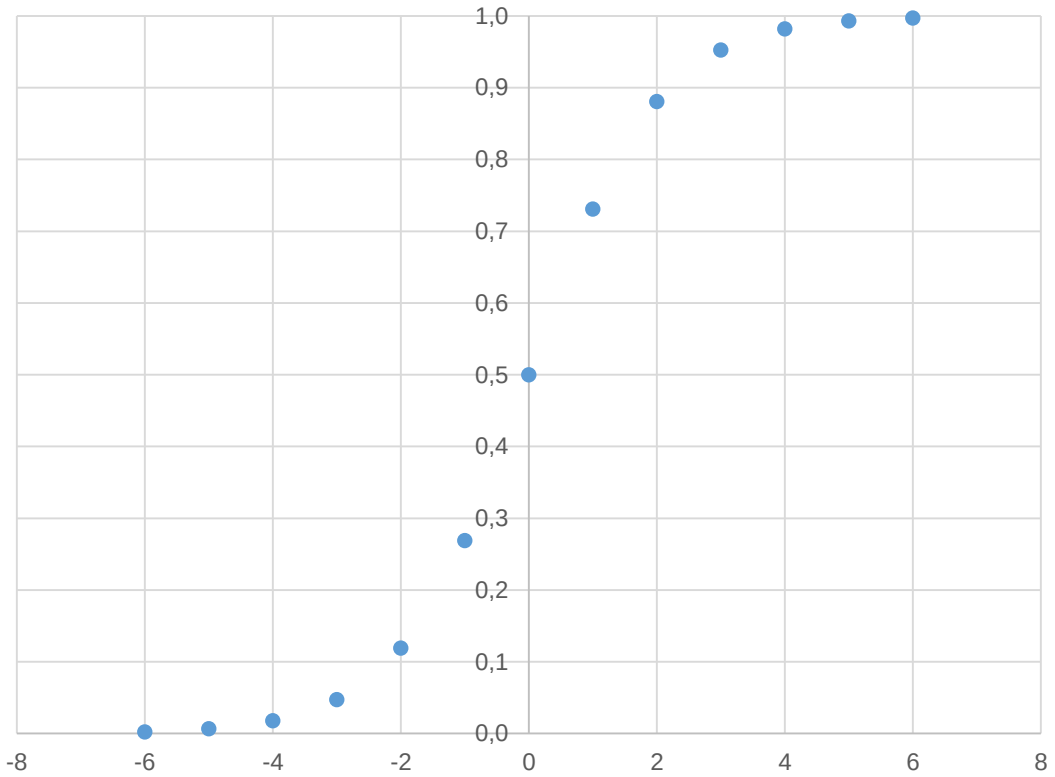
Поведение натурального логарифма

p	odds	ln
0,1	0,1111	-2,19722
0,2	0,2500	-1,38629
0,3	0,4286	-0,8473
0,4	0,6667	-0,40547
0,5	1,0000	0
0,6	1,5000	0,405465
0,7	2,3333	0,847298
0,8	4,0000	1,386294
0,9	9,0000	2,197225



Итоговая функция

logit(p)	p
-6	0,0025
-5	0,0067
-4	0,0180
-3	0,0474
-2	0,1192
-1	0,2689
0	0,5000
1	0,7311
2	0,8808
3	0,9526
4	0,9820
5	0,9933
6	0,9975



$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

$$p = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

Эффективность логистической регрессии

1. AIC (Akaike Information Criteria). Аналогом R^2 в логистической регрессии является AIC. AIC - это мера соответствия, которая штрафует за использование излишнего количества параметров модели. Поэтому мы всегда предпочитаем модель с минимальным значением AIC.

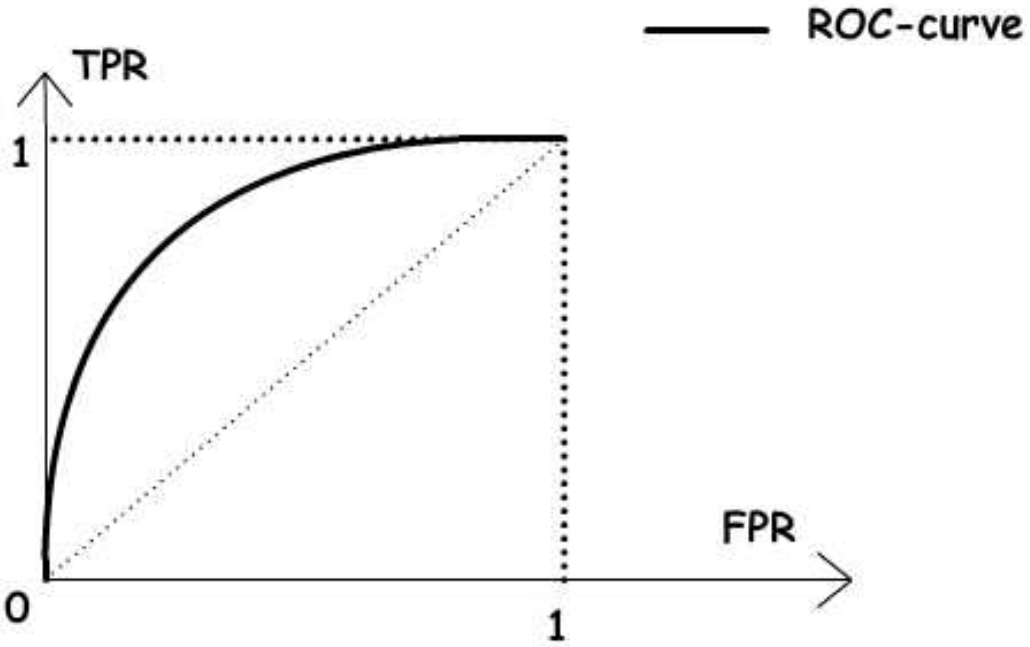
2. Null Deviance и Residual Deviance - Null Deviance указывает на отклонение модели без параметров, только со свободным коэффициентом. Чем меньше значение, тем лучше модель. Residual Deviance указывает указывает на отклонение модели при добавлении независимых переменных. Чем меньше значение, тем лучше модель.

3. Матрица ошибок (Confusion Matrix): это не что иное, как табличное представление фактических и прогнозируемых значений. Она помогает нам оценить точность модели и избежать переобучения.

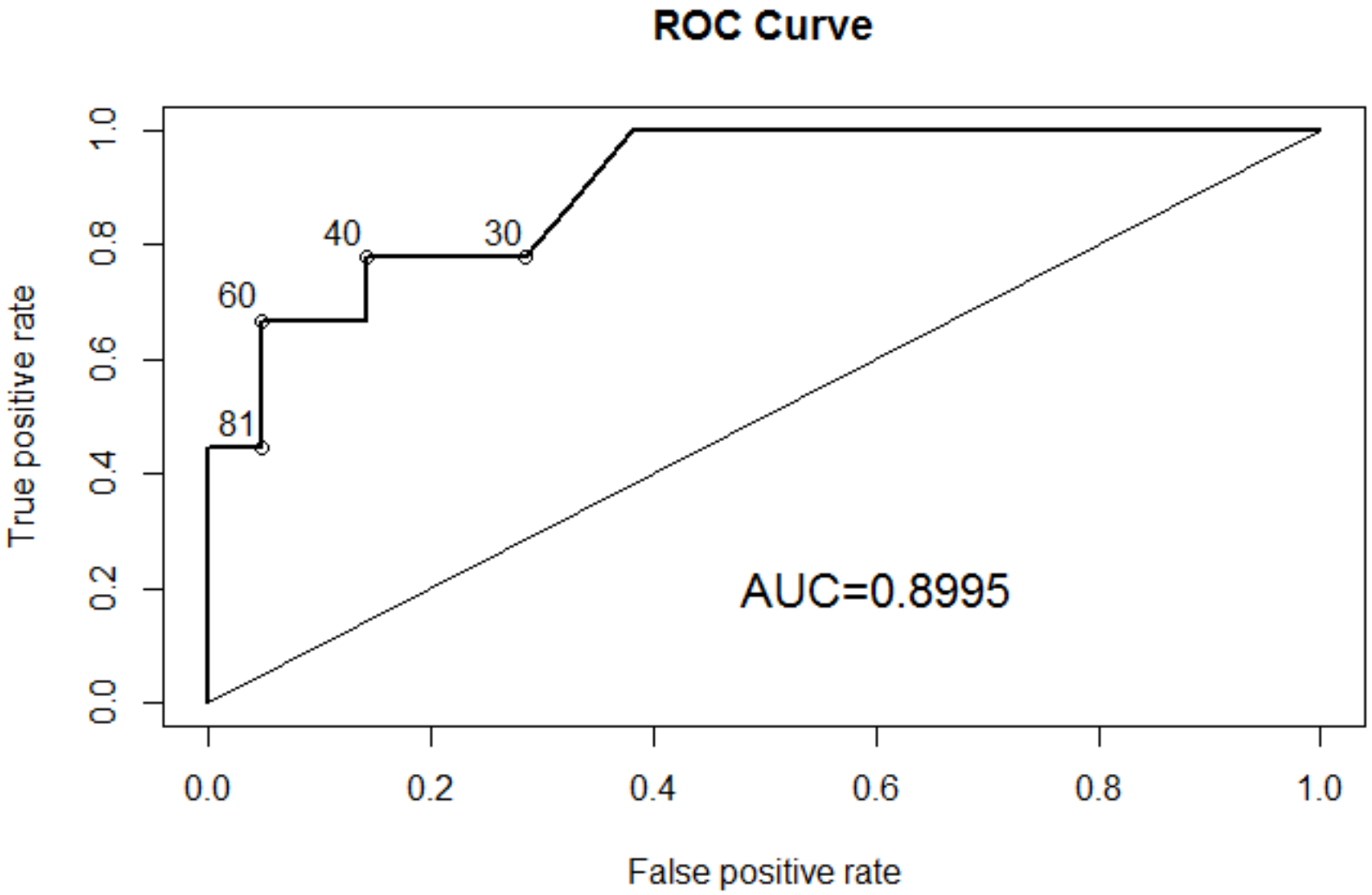
4. ROC-кривая (Receiver Operating Characteristic) суммирует эффективность модели, оценивая компромисс между TPR (чувствительностью) и FPR (1- специфичность). Для построения ROC рекомендуется принять p> 0,5, так как нас больше интересует вероятность успеха. ROC суммирует предсказательную силу для всех возможных значений при p> 0,5. Площадь под кривой (AUC), называемая индексом точности (A) или индексом согласованности, является идеальной метрикой эффективности для ROC-кривой. Чем больше область под кривой, тем лучше предсказательная сила модели. ROC идеальной предсказательной модели имеет TP, равное 1, а FP равное 0. Эта кривая будет касаться верхнего левого угла графика.

ROC-кривая

ROC-кривая (Receiver Operating Characteristic) суммирует эффективность модели, оценивая компромисс между TPR (чувствительностью) и FPR (1- специфичность). Для построения ROC рекомендуется принять p> 0,5, так как нас больше интересует вероятность успеха. ROC суммирует предсказательную силу для всех возможных значений при p> 0,5. Площадь под кривой (AUC), называемая индексом точности (A) или индексом согласованности, является идеальной метрикой эффективности для ROC-кривой. Чем больше область под кривой, тем лучше предсказательная сила модели. ROC идеальной предсказательной модели имеет TP, равное 1, а FP равное 0. Эта кривая будет касаться верхнего левого угла графика.



Реальная ROC-кривая



ROCR

ROCR представляет собой пакет системы R для оценки и визуализации качества классификации на два класса. Есть возможность подсчета различных мер качества классификации и построения 2D графиков для отдельных мер или для зависимости одной меры от другой. Пакет содержит всего три интуитивно понятные функции и два класса для хранения промежуточных данных.

Для использования средств данного пакета необходимы данные об истинных (labels) и предсказанных каким-либо образом (predictions) метках классов выборки объектов.
При этом данные могут представлять собой описание как одной выборки, так и нескольких, например нескольких подвыборок при использовании кросс-валидации. Наборы истинных и предсказанных меток представляются в виде векторов или списков одинаковой длины.

Истинные метки могут принимать только два значения, на которых должно быть задано отношение сравнения (заданы по умолчанию: 0<1 – да и для любых чисел, ‘a’<’b’, FALSE<TRUE). Если предсказанные и истинные метки являются числами, то предсказанные метки могут принимать сколько угодно различных значений (например, если истинные 0 и 1, то предсказанные вполне могут быть заданы промежуточными значениями между 0 и 1), если же предсказанные или истинные метки не являются числами, то предсказанные метки могут принимать только 2 значения, причем те же, что и истинные.

Класс, для которого метка больше, будем называть положительным, а другой – отрицательным.

prediction class

Объекты этого класса предназначены для внутреннего представления исходных данных: истинных и предсказанных каким-то образом меток классов.

predictions	Список, каждый элемент которого представляет собой вектор предсказанных меток.
labels	Список, каждый элемент которого представляет собой вектор истинных меток.
cutoffs	Список, каждый элемент которого представляет собой вектор всех отсечек – всех возможных предсказанных меток (при этом добавляется значение Inf, метки сортируются в порядке убывания и удаляются повторы).
fp	Список, каждый элемент которого представляет собой вектор, который состоит из количеств неправильно классифицированных объектов положительного класса при разделении объектов на основе предсказанных меток на классы по отсечкам из соответствующего вектора cutoffs.
tp	То же, что fp, но для правильно классифицированных объектов положительного класса.
tn	То же, что fp, но для правильно классифицированных объектов отрицательного класса.
fn	То же, что fp, но для неправильно классифицированных объектов отрицательного класса.
n.pos	Список, каждый элемент которого содержит число объектов положительного класса при истинных метках.
n.neg	То же, что n.pos, но для объектов отрицательного класса.
n.pos.pred	Список, каждый элемент которого представляет собой вектор, который состоит из количеств объектов, отнесенных к положительному классу при разделении объектов на основе предсказанных меток на классы по отсечкам из соответствующего вектора cutoffs.
n.neg.pred	То же, что n.pos.pred, но для отрицательного класса.

performance class

Объекты этого класса предназначены для хранения результатов оценки качества классификации в форме предназначенной для построения графика (отдельно рассматриваются меры качества для осей и параметризация).

x.name	Название меры качества, используемой для оси x.
y.name	Название меры качества, используемой для оси y.
alpha.name	Название элемента, используемого для создания параметризованной кривой. Обычно это “none” или “cutoff”.
x.values	Список, каждый элемент которого представляет собой вектор, который состоит из значений меры качества x в точках соответствующего вектора alpha.values.
y.values	То же, что x.values, но для меры качества y.
alpha.values	Список, каждый элемент которого представляет собой вектор, который состоит из значений заданного параметра кривой.

Функции

prediction
Функция для создания объекта класса prediction из исходных данных.

Вызов:
prediction(predictions, labels, label.ordering = NULL)

Аргументы:

predictions	Вектор, матрица, список или фрейм, содержащий предсказанные метки выборки объектов.
labels	Вектор, матрица, список или фрейм, содержащий истинные метки выборки объектов.
label.ordering	Отношение сравнения между метками класса по умолчанию можно изменить, поставив в аргумент вектор, содержащий метки отрицательного и положительного класса.

Функции

performance
Функция для создания объекта класса performance из объекта класса prediction.

Вызов:
performance(prediction.obj, measure, x.measure="cutoff", ...)

Аргументы:

prediction.obj	Объект класса prediction.
measure	Мера качества, используемая для оси y.
x.measure	Мера качества, используемая для оси x.
...	Дополнительные аргументы, которые определены для некоторых мер.

Функции

plot

Функция для визуализации объекта класса performance.

Вызов:

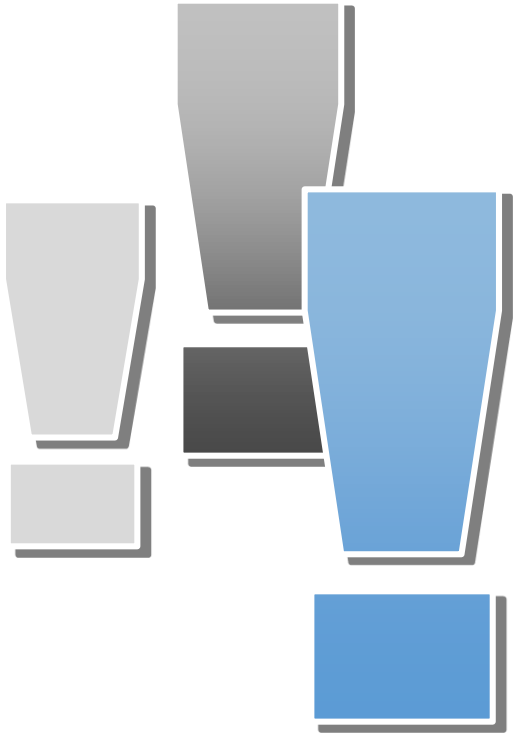
plot(x, y, ..., avg="none", spread.estimate="none", spread.scale=1, show.spread.at=c(), colorize=FALSE, colorize.palette=rev(rainbow(256,start=0, end=4/6)), colorkey=colorize, colorkey.relwidth=0.25, colorkey.pos="right", print.cutoffs.at=c(), cutoff.label.function=function(x) { round(x,2) }, downsampling=0, add=FALSE)

Аргументы:

x	Объект класса performance.
y	Не используется.
…	Дополнительные графические параметры для настройки различных компонент графика. Для обращения к параметру некоторой компоненты нужно пользоваться следующей записью: component.parameter. Доступны следующие компоненты: xaxis, yaxis, coloraxis, box, points, text, plotCI (погрешности), boxplot.При настройке параметров самого холста и кривых префикс указывать не нужно.
avg	Объект, описывающий отображение нескольких кривых (например, если данные содержат несколько выборок объектов, полученных при кросс-валидации, то и кривых получается несколько). Кривые можно усреднять различными способами: • none – кривые рисуются отдельно без усреднения, • horizontal – горизонтальное усреднение, • vertical – вертикальное усреднение, • threshold - усреднение по отсечкам.
spread.estimate	При включенном усреднении кривых, отклонение от средней кривой может быть визуализировано как: stderror – окно стандартной ошибки, stddev – окно стандартного отклонения, boxplot – окно разброса.

spread.scale	Константа, на которую домножаются длины окон stderror и stddev.
show.spread.at	При вертикальном усреднении этот вектор задает позиции x, в которых производится визуализация. По умолчанию она производится в 11 равномерно распределенных по всему пространству значений x точках.
colorize	Логическое значение, показывающее, должна ли кривая быть раскрашена в соответствии с отсечками.
colorize.palette	Если colorize включено, то определяет цветовую палитру, в которой отображается диапазон отсечек.
colorkey	Логическое значение. Если TRUE, то в 4% граничной зоне рисуется цветовой ключ, показывающий отображение отсечек в цветовую палитру.
colorkey.relwidth	Константа от 0 до 1, определяющая часть 4% граничной зоны, которая отводится под цветовой ключ.
colorkey.pos	Определяет, где как рисуется цветовой ключ: вертикально справо или горизонтально сверху.
print.cutoffs.at	Вектор значений отсечек, которые нужно напечатать вдоль кривой в соответствующих точках.
cutoff.label.function	По умолчанию значения отсечек, выводимые на кривой и цветовом ключе, округляются до двух знаков после запятой. Используя этот параметр, можно задать некоторое преобразование отсечек перед выводом (например, округление или взятие логарифма).
downsampling	При очень больших размерах выборки, построение графиков может быть медленным, а их размеры слишком большими. В таких случаях можно строить графики только по части выборки. Данный параметр задает константу от 0 до 1, которая показывает, по какой части объектов нужно строить графики. Если значение больше 1, то графики строятся по всей выборке.
add	Если TRUE, то кривые добавляются к уже существующему графику, иначе создается новый график.

Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57