взято с сайта: <https://basegroup.ru/community/articles/logistic>

Логистическая регрессия и ROC-анализ — математический аппарат

Введение

Логистическая регрессия – полезный классический инструмент для решения задачи регрессии и классификации. Без логистической регрессии и ROC-анализа – аппарата для анализа качества моделей – немыслимо построение моделей в медицине и проведение клинических исследований. В последние годы логистическая регрессия получила распространение в скоринге для расчета рейтинга заемщиков и управления кредитными рисками. Поэтому, несмотря на свое "происхождение" из статистики, логистическую регрессию и ROC-анализ почти всегда можно увидеть в наборе Data Mining алгоритмов.

Логистическая регрессия

Логистическая регрессия – это разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной. Бинарная логистическая регрессия, как следует из названия, применяется в случае, когда зависимая переменная является бинарной (т.е. может принимать только два значения). Иными словами, с помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого (больной/здоровый, возврат кредита/дефолт и т.д.).

Как известно, все регрессионные модели могут быть записаны в виде формулы:



Например, в множественной линейной регрессии предполагается, что зависимая переменная является линейной функцией независимых переменных, т.е.:



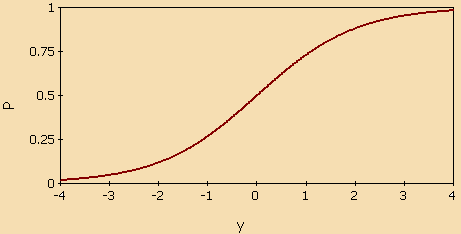
Можно ли ее использовать для задачи оценки вероятности исхода события? Да, можно, вычислив стандартные коэффициенты регрессии. Например, если рассматривается исход по займу, задается переменная y со значениями 1 и 0, где 1 означает, что соответствующий заемщик расплатился по кредиту, а 0, что имел место дефолт. Однако здесь возникает проблема: множественная регрессия не "знает", что переменная отклика бинарна по своей природе. Это неизбежно приведет к модели с предсказываемыми значениями большими 1 и меньшими 0. Но такие значения вообще не допустимы для первоначальной задачи. Таким образом, множественная регрессия просто игнорирует ограничения на диапазон значений для y.

Для решения проблемы задача регрессии может быть сформулирована иначе: вместо предсказания бинарной переменной, мы предсказываем непрерывную переменную со значениями на отрезке [0,1] при любых значениях независимых переменных. Это достигается применением следующего регрессионного уравнения (логит-преобразование):



где P – вероятность того, что произойдет интересующее событие; e – основание натуральных логарифмов 2,71…; y – стандартное уравнение регрессии.

Зависимость, связывающая вероятность события и величину y, показана на следующем графике (рис. 1):



*Рисунок 1 – Логистическая кривая*

Поясним необходимость преобразования. Предположим, что мы рассуждаем о нашей зависимой переменной в терминах основной вероятности P, лежащей между 0 и 1. Тогда преобразуем эту вероятность P:



Это преобразование обычно называют логистическим или логит-преобразованием. Теоретически P’ может принимать любое значение. Поскольку логистическое преобразование решает проблему об ограничении на 0-1 границы для первоначальной зависимой переменной (вероятности), то эти преобразованные значения можно использовать в обычном линейном регрессионном уравнении. А именно, если произвести логистическое преобразование обеих частей описанного выше уравнения, мы получим стандартную модель линейной регрессии.

Существует несколько способов нахождения коэффициентов логистической регрессии. На практике часто используют *метод максимального правдоподобия*. Он применяется в статистике для получения оценок параметров генеральной совокупности по данным выборки. Основу метода составляет *функция правдоподобия* (likehood function), выражающая плотность вероятности (вероятность) совместного появления результатов выборки Y1, Y2,…,Yk:



Согласно методу максимального правдоподобия в качестве оценки неизвестного параметра принимается такое значение , которое максимизирует функцию L.

Нахождение оценки упрощается, если максимизировать не саму функцию L, а натуральный логарифм ln(L), поскольку максимум обеих функций достигается при одном и том же значении θ:



В случае бинарной независимой переменной, которую мы имеем в логистической регрессии, выкладки можно продолжить следующим образом. Обозначим через Piвероятность появления единицы: Pi=Prob(Yi=1). Эта вероятность будет зависеть от **XiW**, где **Xi** – строка матрицы регрессоров, **W** – вектор коэффициентов регрессии:



Логарифмическая функция правдоподобия равна:



где I0, I1 – множества наблюдений, для которых Yi=0 и Yi=1 соответственно.

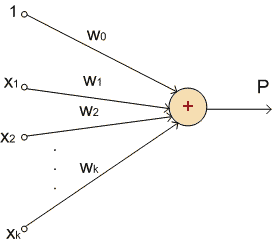
Можно показать, что градиент **g** и гессиан H функции правдоподобия равны:



Гессиан всюду отрицательно определенный, поэтому логарифмическая функция правдоподобия всюду вогнута. Для поиска максимума можно использовать метод Ньютона, который здесь будет всегда сходиться (выполнено условие сходимости метода):



На самом деле, логистическую регрессию можно представить в виде однослойной нейронной сети с сигмоидальной функцией активации, веса которой есть коэффициенты логистической регрессии, а вес поляризации – константа регрессионного уравнения (рис. 2).



*Рисунок 2 – Представление логистической регрессии в виде нейронной сети*

Как известно, однослойная нейронная сеть может успешно решить лишь задачу линейной сепарации. Поэтому возможности по моделированию нелинейных зависимостей у логистической регрессии отсутствуют. Однако для оценки качества модели логистической регрессии существует эффективный инструмент ROC-анализа, что является несомненным ее преимуществом.

Для расчета коэффициентов логистической регрессии можно применять любые градиентные методы: метод сопряженных градиентов, методы переменной метрики и другие.

ROC-анализ

ROC-кривая (Receiver Operator Characteristic) – кривая, которая наиболее часто используется для представления результатов бинарной классификации в машинном обучении. Название пришло из систем обработки сигналов. Поскольку классов два, один из них называется классом с положительными исходами, второй – с отрицательными исходами. ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. В терминологии ROC-анализа первые называются истинно положительным, вторые – ложно отрицательным множеством. При этом предполагается, что у классификатора имеется некоторый параметр, варьируя который, мы будем получать то или иное разбиение на два класса. Этот параметр часто называют порогом, или точкой отсечения (cut-off value). В зависимости от него будут получаться различные величины *ошибок I и II рода*.

В логистической регрессии порог отсечения изменяется от 0 до 1 – это и есть расчетное значение уравнения регрессии. Будем называть его рейтингом.

Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу сопряженности (confusion matrix), которая строится на основе результатов классификации моделью и фактической (объективной) принадлежностью примеров к классам.

|  |  |  |
| --- | --- | --- |
|  | **Фактически** | |
| **Модель** | **положительно** | **отрицательно** |
| положительно | TP | FP |
| отрицательно | FN | TN |

* TP (*True Positives*) – верно классифицированные положительные примеры (так называемые истинно положительные случаи);
* TN (*True Negatives*) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);
* FN (*False Negatives*) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый "ложный пропуск" – когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);
* FP (*False Positives*) – отрицательные примеры, классифицированные как положительные (ошибка II рода); Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи. Например, если мы прогнозируем вероятность наличия заболевания, то положительным исходом будет класс "Больной пациент", отрицательным – "Здоровый пациент". И наоборот, если мы ходим определить вероятность того, что человек здоров, то положительным исходом будет класс "Здоровый пациент", и так далее.

При анализе чаще оперируют не абсолютными показателями, а относительными – долями (rates), выраженными в процентах:

Доля истинно положительных примеров (True Positives Rate):



Доля ложно положительных примеров (False Positives Rate):



Введем еще два определения: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

Чувствительность (Sensitivity) – это и есть доля истинно положительных случаев:



Специфичность (Specificity) – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:



Заметим, что FPR=100-Sp.

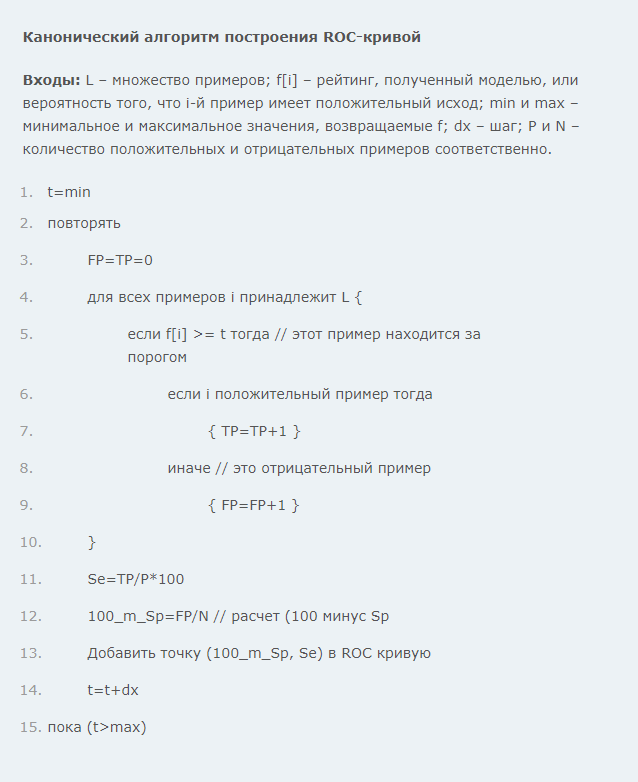
Попытаемся разобраться в этих определениях.

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры). Если рассуждать в терминах медицины – задачи диагностики заболевания, где модель классификации пациентов на больных и здоровых называется диагностическим тестом, то получится следующее:

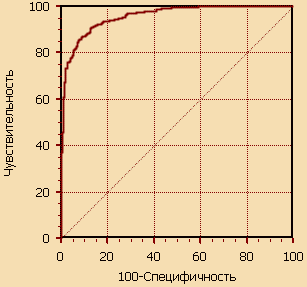
* Чувствительный диагностический тест проявляется в гипердиагностике – максимальном предотвращении пропуска больных;
* Специфичный диагностический тест диагностирует только доподлинно больных. Это важно в случае, когда, например, лечение больного связано с серьезными побочными эффектами и гипердиагностика пациентов не желательна.

ROC-кривая получается следующим образом:

1. Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом dx (например, 0.01) рассчитываются значения чувствительности Se и специфичности Sp. В качестве альтернативы порогом может являться каждое последующее значение примера в выборке.
2. Строится график зависимости: по оси Y откладывается чувствительность Se, по оси X – 100%–Sp (сто процентов минус специфичность), или, что то же самое, FPR – доля ложно положительных случаев.



В результате вырисовывается некоторая кривая (рис. 3).



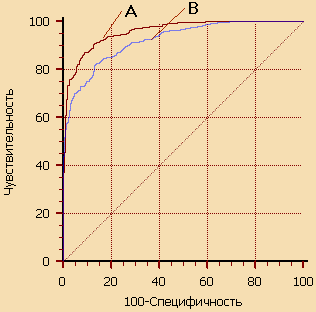
*Рисунок 3 – Кривая ROC*

График часто дополняют прямой y=x.

Заметим, что имеется более экономичный способ расчета точек ROC-кривой, чем тот, который приводился выше, т.к. его вычислительная сложность нелинейная и равна O(n2): для каждого порога необходимо "пробегать" по записям и каждый раз рассчитывать TP и FP. Если же двигаться вниз по набору данных, отсортированному по убыванию выходного поля классификатора (рейтингу), то можно за один проход вычислить значения всех точек ROC-кривой, последовательно обновляя значения TP и FP.

Для идеального классификатора график ROC-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100% или 1.0 (идеальная чувствительность), а доля ложно положительных примеров равна нулю. Поэтому чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели. Наоборот, чем меньше изгиб кривой и чем ближе она расположена к диагональной прямой, тем менее эффективна модель. Диагональная линия соответствует "бесполезному" классификатору, т.е. полной неразличимости двух классов.

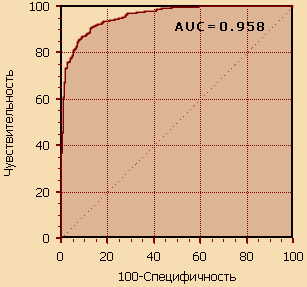
При визуальной оценке ROC-кривых расположение их относительно друг друга указывает на их сравнительную эффективность. Кривая, расположенная выше и левее, свидетельствует о большей предсказательной способности модели. Так, на рис. 4 две ROC-кривые совмещены на одном графике. Видно, что модель "A" лучше.



*Рисунок 4 – Сравнение ROC-кривых*

Визуальное сравнение кривых ROC не всегда позволяет выявить наиболее эффективную модель. Своеобразным методом сравнения ROC-кривых является оценка площади под кривыми. Теоретически она изменяется от 0 до 1.0, но, поскольку модель всегда характеризуются кривой, расположенной выше положительной диагонали, то обычно говорят об изменениях от 0.5 ("бесполезный" классификатор) до 1.0 ("идеальная" модель). Эта оценка может быть получена непосредственно вычислением площади под многогранником, ограниченным справа и снизу осями координат и слева вверху – экспериментально полученными точками (рис. 5). Численный показатель площади под кривой называется AUC (Area Under Curve). Вычислить его можно, например, с помощью численного метода трапеций:

AUC=∫f(x)dx=∑i[Xi+1+Xi2]⋅(Yi+1−Yi)AUC=∫f(x)dx=∑i[Xi+1+Xi2]⋅(Yi+1−Yi)



*Рисунок 5 – Площадь под ROC-кривой*

С большими допущениями можно считать, что чем больше показатель AUC, тем лучшей прогностической силой обладает модель. Однако следует знать, что:

* показатель AUC предназначен скорее для сравнительного анализа нескольких моделей;
* AUC не содержит никакой информации о чувствительности и специфичности модели.

В литературе иногда приводится следующая экспертная шкала для значений AUC, по которой можно судить о качестве модели:

|  |  |
| --- | --- |
| **Интервал AUC** | **Качество модели** |
| 0.9-1.0 | Отличное |
| 0.8-0.9 | Очень хорошее |
| 0.7-0.8 | Хорошее |
| 0.6-0.7 | Среднее |
| 0.5-0.6 | Неудовлетворительное |

Идеальная модель обладает 100% чувствительностью и специфичностью. Однако на практике добиться этого невозможно, более того, невозможно одновременно повысить и чувствительность, и специфичность модели. Компромисс находится с помощью порога отсечения, т.к. пороговое значение влияет на соотношение Se и Sp. Можно говорить о задаче нахождения *оптимального порога отсечения* (optimal cut-off value).

Порог отсечения нужен для того, чтобы применять модель на практике: относить новые примеры к одному из двух классов. Для определения оптимального порога нужно задать критерий его определения, т.к. в разных задачах присутствует своя оптимальная стратегия. Критериями выбора порога отсечения могут выступать:

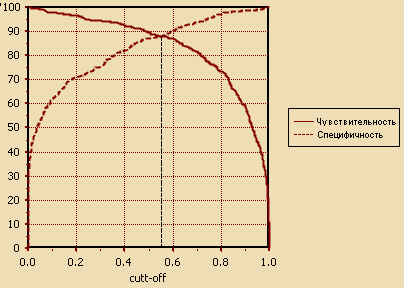
* Требование минимальной величины чувствительности (специфичности) модели. Например, нужно обеспечить чувствительность теста не менее 80%. В этом случае оптимальным порогом будет максимальная специфичность (чувствительность), которая достигается при 80% (или значение, близкое к нему "справа" из-за дискретности ряда) чувствительности (специфичности).
* Требование максимальной суммарной чувствительности и специфичности модели, т.е.

Cutt−offo=maxk(Sek+Spk)Cutt\_offo=maxk(Sek+Spk)

* Требование баланса между чувствительностью и специфичностью, т.е. когда Se≈SpSe≈Sp:

Cutt−offo=mink∣∣Sek−Spk∣∣Cutt\_offo=mink|Sek−Spk|

Второе значение порога обычно предлагается пользователю по умолчанию. В третьем случае порог есть точка пересечения двух кривых, когда по оси X откладывается порог отсечения, а по оси Y –чувствительность или специфичность модели (рис. 6).



*Рисунок 6 – "Точка баланса" между чувствительностью и специфичностью*

Существуют и другие подходы, когда ошибкам I и II рода назначается вес, который интерпретируется как цена ошибок. Но здесь встает проблема определения этих весов, что само по себе является сложной, а часто не разрешимой задачей.

В следующей статье будет показано применение логистической регрессии и ROC-анализа на практике: в [медицине и скоринге](http://www.basegroup.ru/library/practice/logis_medic_scoring/).

**ЛИТЕРАТУРА**

* Цыплаков А. А. Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии. Учебное пособие.
* Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers // 2004 Kluwer Academic Publishers.
* Zweig M.H., Campbell G. ROC Plots: A Fundamental Evaluation Tool in Clinical Medicine // Clinical Chemistry, Vol. 39, No. 4, 1993.
* Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves // Proc. Of 23 International Conference on Machine Learning, Pittsburgh, PA, 2006