Оценивание значимости переменных для ранговой регрессии

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

«Математические методы распознавания образов»
(ММРО-18), г. Таганрог, 9–13 октября 2017 г.

Анализ информативности переменных

Цель работы: распространение понятия ROC–кривой на задачу регрессионного анализа, с сохранением свойств из задачи классификации.

Подцели:

- визуальный анализ ROC–кривой,
- оценка информативности переменных,
- критерий продолжения уточнения решения,
- критерий останова для методов бустинга.

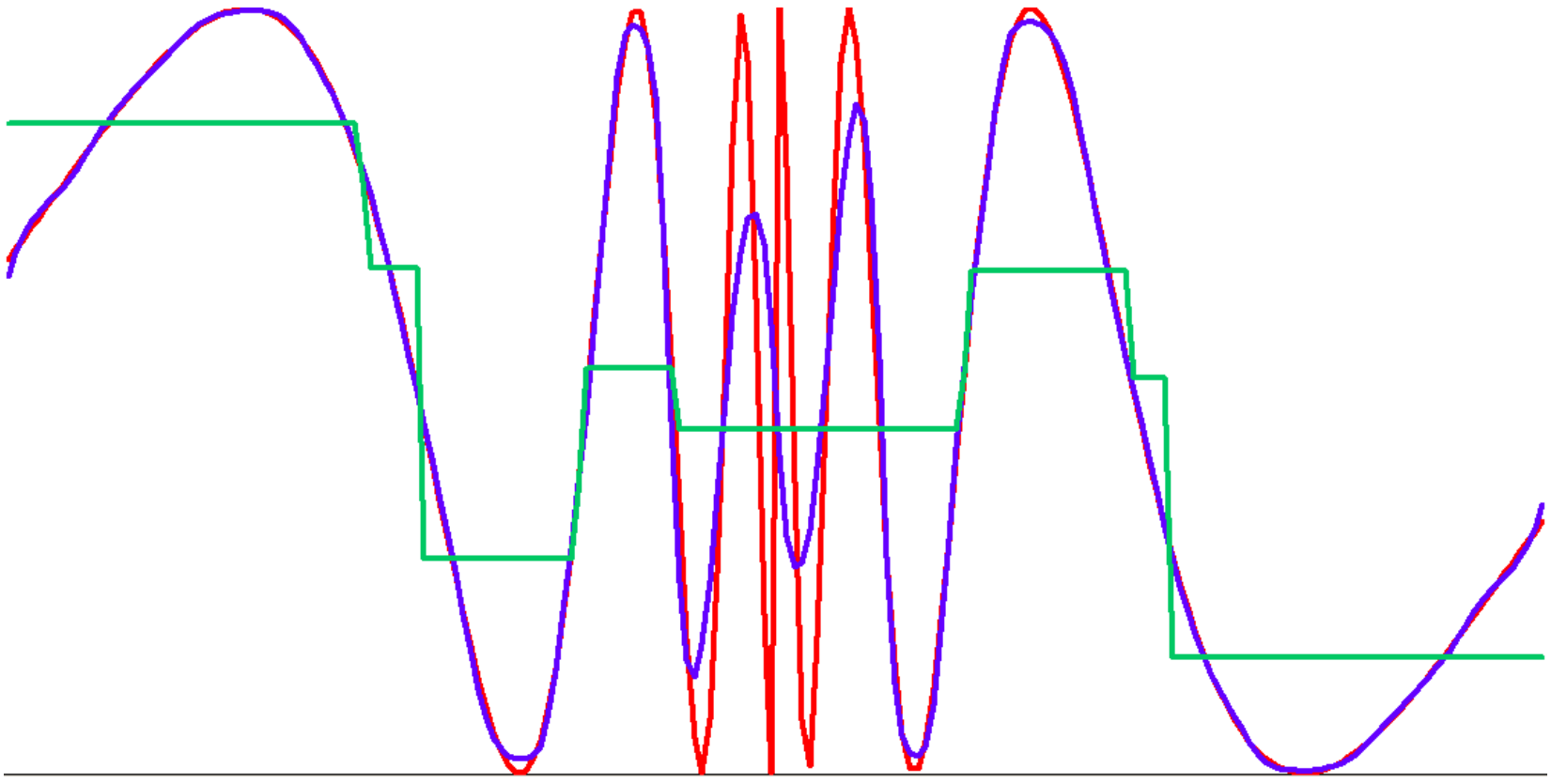
Бустинг как метод аппроксимации

Решение методом бустинга можно представить в форме логистической регрессии, имеющей аналогию с рядом Бахадура, который даёт возможность учитывать зависимости между переменными, последовательно добавляя парные зависимости, зависимости в тройках и т.д.

$$P(y=1|x) = \sigma\left(u_0 + \sum_{j=1}^{n} u_j s_j(x_j) + \sum_{j,k} u_{jk} s_{jk}(x_j, x_k) + \right.$$

$$\left. + \sum_{j,k,l} u_{jkl} s_{jkl}(x_j, x_k, x_l) + \ldots\right).$$

Аппроксимация функции условной вероятности



Кубический сплайн на 20 интервалов.
AdaBoost 10 итераций.

Критерий останова при наращивании композиции

Для выбора сложности композиции (в т.ч. для предотвращения переобучения бустинга):

- скользящий экзамен,
- статистический критерий.

Статистический критерий должен оценить значимость зависимости регрессионных остатков для заданной переменной.

Постановка задачи

Пусть X — пространство значений переменных, используемых для прогноза,
$Y = (-\infty, +\infty)$ — пространство значений прогнозируемых переменных.
Решающей функцией называется соответствие $\lambda\colon X \to Y$.
Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$.
В данной работе будем рассматривать функции потерь вида $\mathcal{L}(y, y') = \mathcal{L}_R(F(y), F(y'))$, где $F(y)$ – функция распределения.
С таким критерием задачу построения решающих функций будем называть задачей ранговой регрессии.

Свойства ROC–кривой в задаче классификации

- ROC–кривая вычисляется на основе известных значений целевой переменной и заданного порядка объектов.
- Из двух решений лучше то, для которого ROC кривая лежит выше.
- Для случайного (наугад) решения ROC–кривая близка к прямой.
- Для оптимального решения ROC–кривая состоит из вертикального и горизонтального отрезков.

Кривая REC

Рассмотрим известные способы определения кривой ошибок в задаче регрессионного анализа.
Кривая REC (regression error characteristic) — это по сути эмпирическая функция распределения для величины ошибки (потерь). Единственное отличие в том, что эмпирическая функция распределения ступенчатая, а REC кривая соединяет вершины «ступенек» отрезками.
Данная кривая обладает некоторыми свойствами, схожими со свойствами ROC кривой в задаче классификации. В частности, из двух решений лучше то, для которого REC кривая лежит выше.
Однако REC кривая вряд ли может рассматриваться как аналог ROC кривой, поскольку её поведение в целом существенно иное.

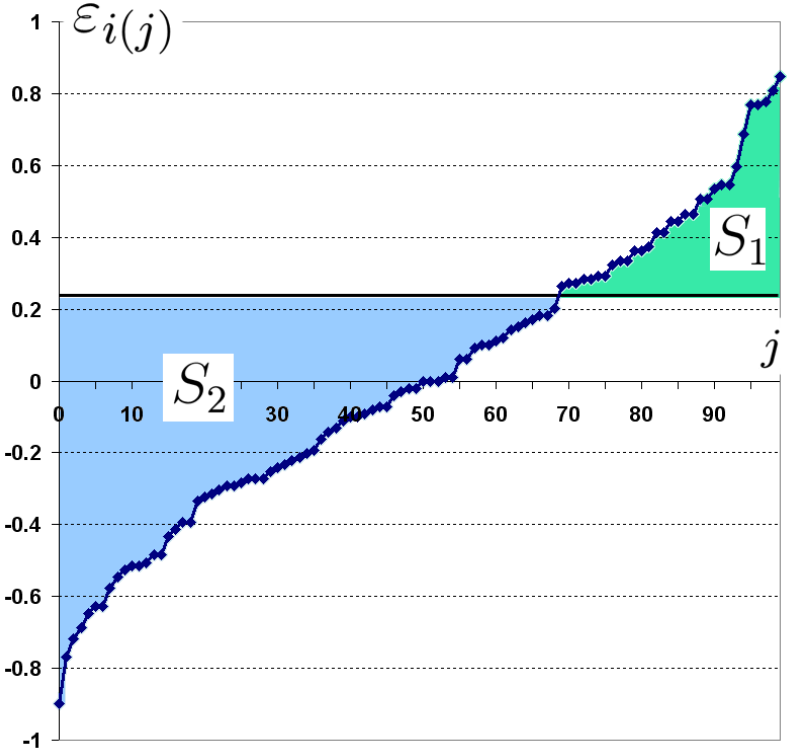
Кривая RROC

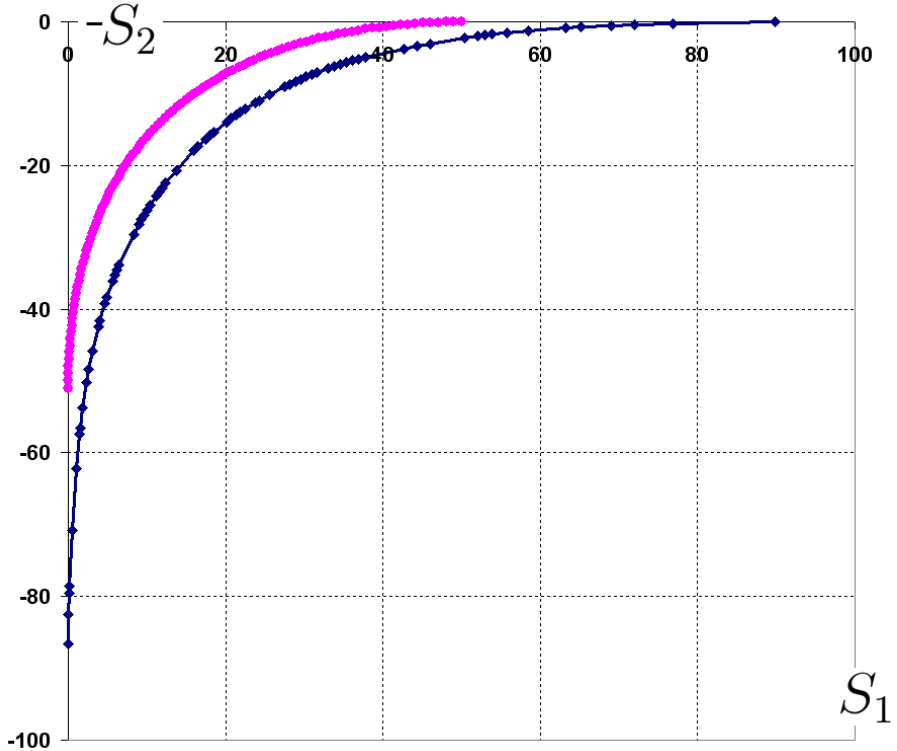
Обозначим $\varepsilon_i = y^i - \hat{y}^i$ – последовательность регрессионных остатков (отклонений истинных значений от прогнозируемых). Здесь $\hat{y}^i = \lambda(x^i)$.
Пусть $i(j)$, $j = 1, ..., N$, – перестановка индексов, которая упорядочивает значения $\varepsilon_{i(j)}$ по неубыванию.
Введём величины

$$S_1(j) = \sum_{k=j}^{N} \left(\varepsilon_{i(k)} - \varepsilon_{i(j)}\right), \quad -S_2(j) = \sum_{k=1}^{j} \left(\varepsilon_{i(k)} - \varepsilon_{i(j)}\right).$$

Кривая RROC (Regression ROC curve) определяется как ломаная с вершинами в точках $(S_1(j), -S_2(j))$, $j = 1, ..., N$.

Построение RROC кривой

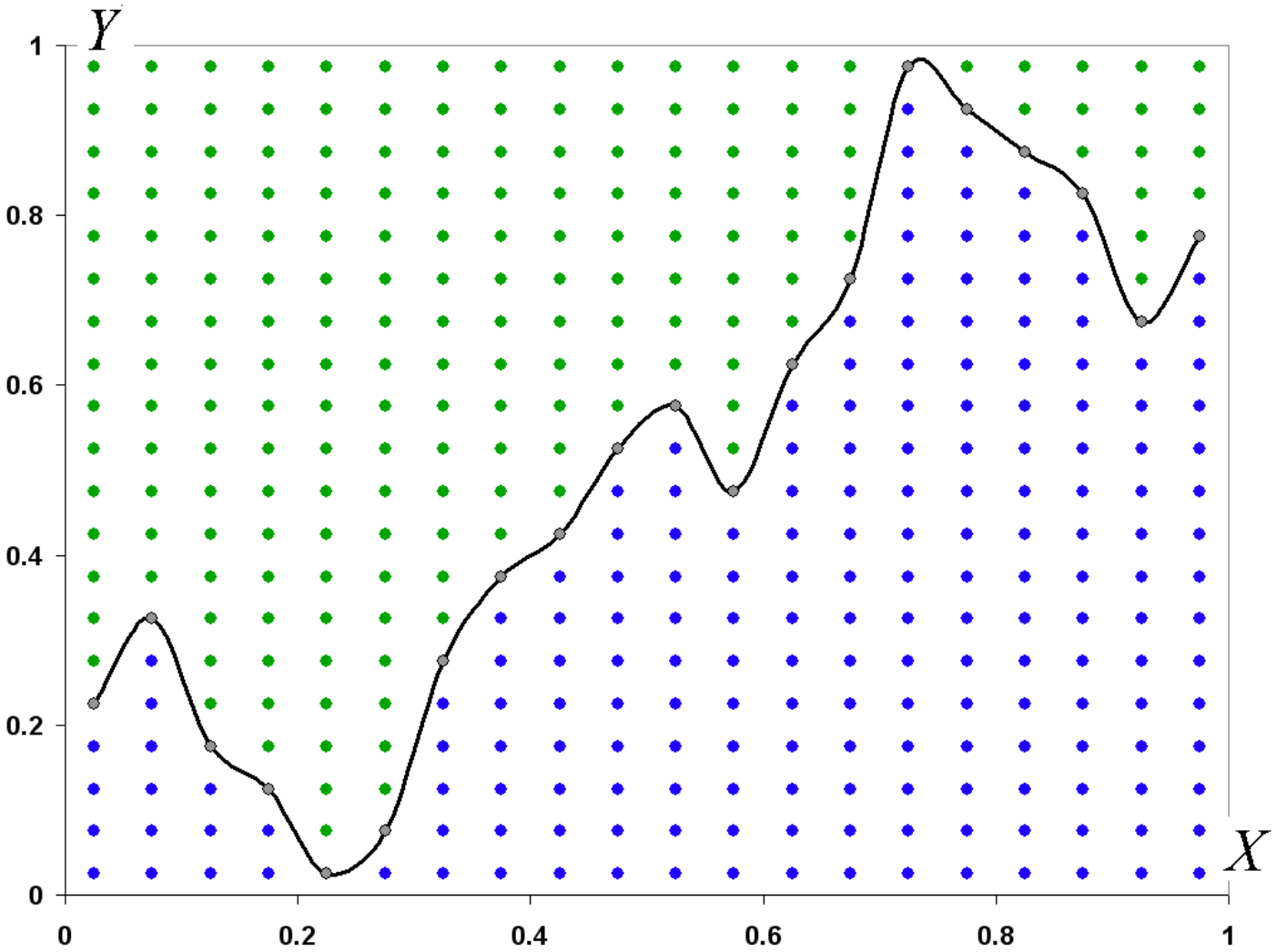




ROC кривая для регрессии

Для построения ROC кривой представим задачу построения регрессии как задачу классификации.
Сконструируем по исходной выборке фиктивную выборку с бинарной целевой переменной.
Рассмотрим случай, когда все y^i различны. Обозначим $r(i)$ – ранг значения y^i. Иными словами, $r(i)$ – это перестановка индексов, которая упорядочивает значения y^i в выборке по возрастанию.
Для каждой пары (x^i, y^i) исходной выборки включим в новую выборку $r(i) - 1$ пар $(X^i, -1)$ и $N - r(i)$ пар $(X^i, 1)$.

Фиктивная выборка



Критерий Андерсона – Дарлинга

Статистика критерия Ω^2 Мизеса (статистика Андерсона – Дарлинга) определяется выражением

$$S_\Omega = -n - 2\sum_{i=1}^{n}\left(\frac{2i-1}{2n}\ln F(x_i) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_i))\right).$$

Данная статистика совпадает с функцией правдоподобия для фиктивной выборки.

Ранговая регрессия как классификация

Можно было создать нечёткую выборку, где значение ранга бы интерпретировалось как степень принадлежности первому классу.

После такого представления можно строить регрессию обычными методами классификации.

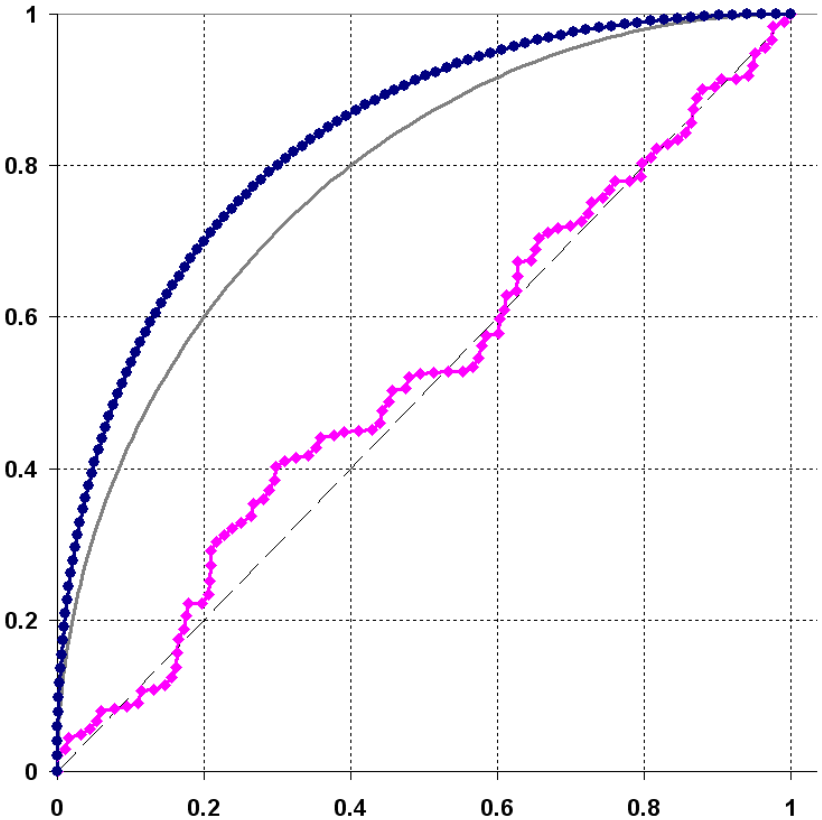
ROC_r кривая

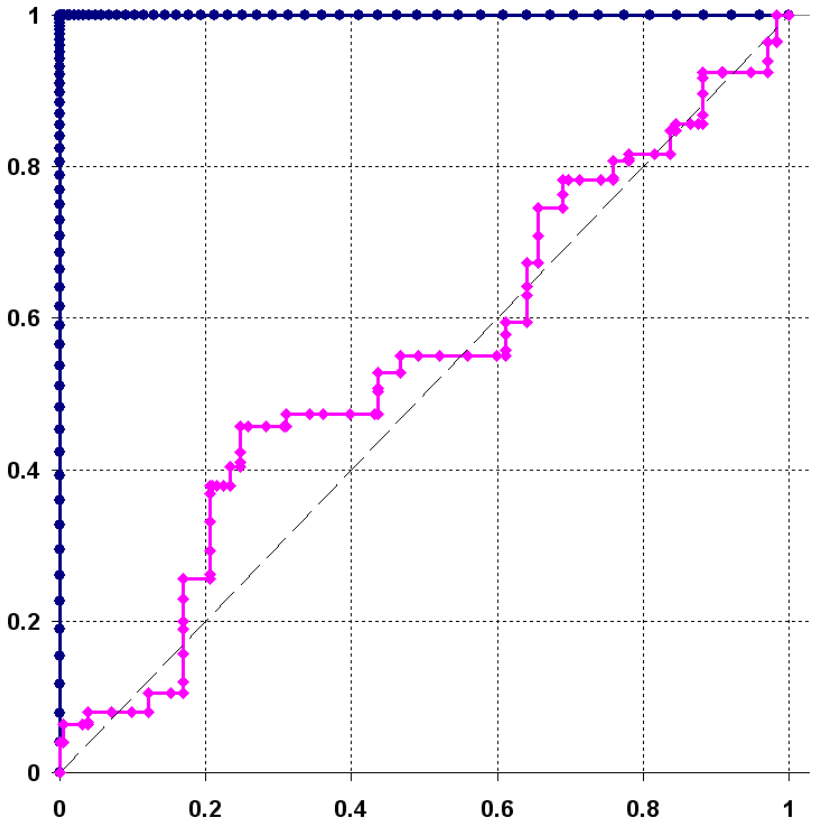
Для полученной выборки строим обычную ROC-кривую, которую назовём ROC_r кривой.
Алгоритм построения ROC_r кривой можно описать следующим образом.
Ставим «перо» в начало координат.
Далее перебираем (в заданном порядке) все объекты выборки, и для каждого объекта перемещаем «перо» на $\frac{r(i)-1}{N(N-1)}$ вправо и на $\frac{N-r(i)}{N(N-1)}$ вверх.
В итоге «перо» оказывается в точке $(1,1)$.

ROC_d кривая

Чтобы добиться более полного сходства с ROC кривой введём модификацию, которую назовём ROC_d кривой. Для каждой пары (x^i, y^i) исходной выборки включим в новую выборку $|N+1-2r(i)|$ пар $(X^i, sign(N+1-2r(i)))$. Для полученной выборки строим обычную ROC-кривую, которую назовём ROC_d кривой. В отличие от предыдущего варианта, здесь мы «сокращаем» объекты противоположных классов (с одинаковыми X^i).
Визуально ROC_d кривые среди всех рассмотренных в большей степени соответствуют поведению ROC-кривых в задаче классификации. Однако ROC_d кривая имеет существенный недостаток, заключающийся в нечувствительности к изменению прогнозируемого ранга для объекта, у которого $N+1-2r(i)=0$.

Кривые ROC_r и ROC_d





Эмпирический мост

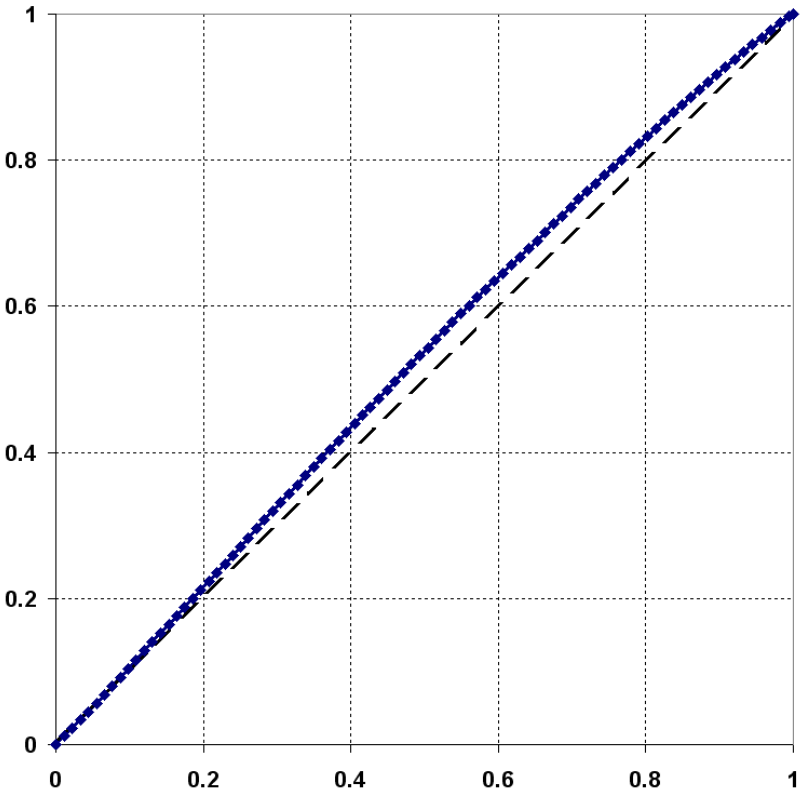
Для анализа регрессионных зависимостей известна конструкция, названная эмпирическим мостом [Гусарова Г. В., Ковалевский А. П., Макаренко А. Г. Критерии наличия разладки // Сиб. журн. индустр. матем., 2005. Т. 8. № 4. С. 18–33.].
Эмпирический мост — это ломаная, соединяющая последовательность точек

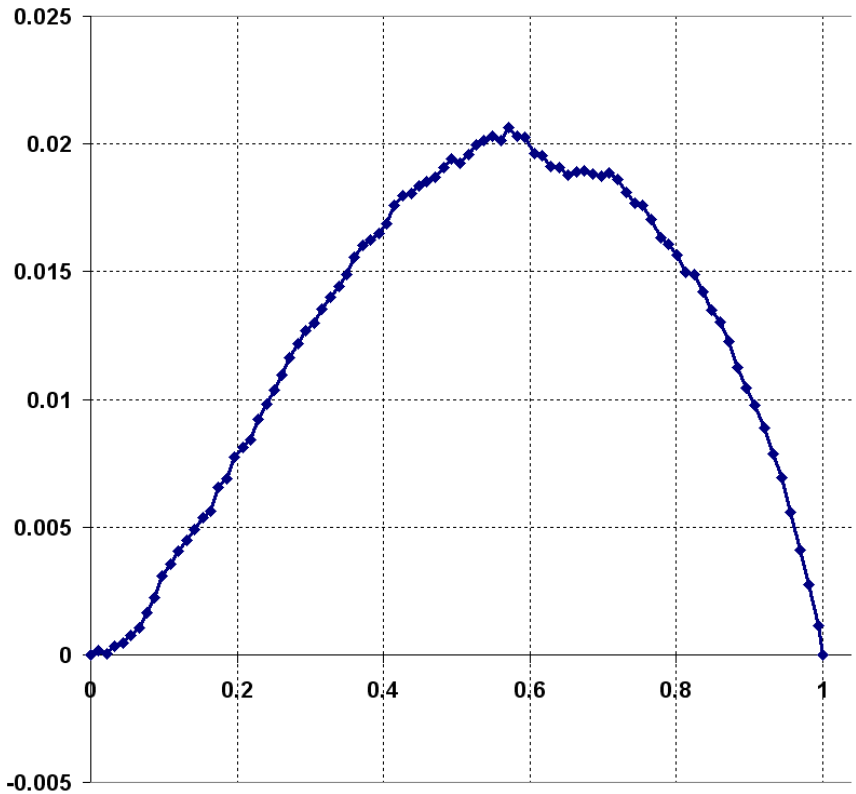
$$\left(\frac{j}{N}, \frac{\Delta_j - \frac{j}{N}\Delta_N}{\sigma\sqrt{N}} \right),$$

где $\Delta_j = \sum_{k=1}^{j} \varepsilon_k$ – частичная сумма регрессионных остатков, а σ – стандартное отклонение остатков.

Кривая ROC_r и эмпирический мост

Пример из прикладной задачи.





Связь понятий

Выясним, что представляет собой эмпирический мост, если в качестве значений целевой переменной взять ранги $r(j)$, а в качестве прогнозируемых значений — константу, равную среднему рангу, т.е. $\frac{N+1}{2}$.
В этом случае $\varepsilon_k = r(k) - \frac{N+1}{2}$, а $\Delta_N = 0$.
При повороте на 45° кривая ROC_r превращается в эмпирический мост для медианного прогноза с точностью до масштабирующих коэффициентов по осям координат.

Эмпирический мост в задаче классификации

Эмпирический мост можно построить и в задаче классификации, когда $Y \in \{-1, 1\}$.
Для этого достаточно в качестве решения взять среднее по выборке значение целевой переменной.
При повороте на 45° ROC кривая превращается в эмпирический мост для константного прогноза с точностью до масштабирующих коэффициентов по осям координат.

Эмпирический мост как универсальная конструкция

Эмпирический мост для построения требует наличия некоторого текущего решения и некоторого порядка на объектах.
Помогает ответить на вопрос, можно ли улучшить решение, за счёт использования данного порядка.
ROC–кривая является (с точностью до поворота) частным случаем эмпирического моста. При этом в качестве «текущего решения» выступает константный прогноз.

Выводы

Предложеные обобщения понятия ROC кривой на случай регрессионного анализа более полно воспроизводят свойства ROC кривой по сравнению с известными вариантами ROC кривой для регрессии, такими как RROC и REC кривые.
ROC кривые при случайном прогнозе приближаются к прямой, а отклонения от прямой позволяют оценить информативность «объясняющей» переменной.
Предложенные варианты ROC кривой оказались близкими понятию эмпирического моста.
Кривая ROC_d наиболее полно соответствует визуальному образу ROC кривой в задаче классификации, но не вполне отражает информацию о качестве прогноза.
Вопрос построения регрессионного аналога ROC кривой, который бы полностью отражал все свойства этой кривой в задаче классификации, пока остаётся открытым.

Выводы

- RROC и REC кривые не воспроизводят привычные свойства ROC–кривой.
- Задача ранговой регрессии может рассматриваться как задача классификации.
- ROC–кривая является (с точностью до поворота) частным случаем эмпирического моста.
- Эмпирическй мост позволяет строить критерии информативности переменных и оценивать «остаточную предсказуемость».