Лекция 4

Статистические методы распознавания, Распознавание при заданной точности для некоторых классов, ROC-анализ

Лектор – *Сенько Олег Валентинович*

Структура распознающего алгоритма

Каждый алгоритм распознавания классов $K_1, \ldots, K_l$ может быть представлен как последовательное выполнение распознающего оператора $\mathbf{R}$ и решающего правила :

$$A = \mathbf{R} \otimes \mathbf{C}.$$

Оператор оценок вычисляет для распознаваемого объекта s вещественные оценки $\gamma_1, \ldots, \gamma_L$ за классы $K_1, \ldots, K_l$ соответственно.

Пример модели распознавания - Линейная машина

Множество (модель) алгоритмов $\widetilde{W} = \{A : \widetilde{X} \to \widetilde{Y}\}$ внутри которого производится поис оптимального алгоритма прогнозирования вместе со способом решения оптимизационной задачи будем называть методом прогнозирования или методом распознавания, если прогнозируемая величина принадлежит конечному множеству. В качестве примера рассмотрим известный метод решения задачи распознавания – Линейная машина

Метод «Линейная машина» предназначен для решения задачи распознавания с классами $K_1, \ldots, K_L$. Алгоритм распознавания имеет следующий вид. В процессе обучения классам $K_1, \ldots, K_L$ ставятся в соответствие линейные функции от переменных $X_1, \ldots, X_n$:

$$\gamma_1(X_1, \ldots, X_n) = w_0^1 + w_1^1 X_1 + \ldots + w_n^1 X_n$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\gamma_L(X_1, \ldots, X_n) = w_0^L + w_1^L X_1 + \ldots + w_n^L X_n.$$

Таким образом алгоритм распознавания задаётся параметров

$$w_0^1, \ldots, w_n^1$$

$$\ldots\ldots\ldots$$

$$w_0^L, \ldots, w_n^L$$

.

Пусть требуется распознать объект s^* , описание которого задаётся вектором $\boldsymbol{x}^*$. Вычисляются значения функций $\gamma_1, \ldots, \gamma_L$ в точке $\boldsymbol{x}^*$. Объект s^* будет отнесён классу K_i, если выполняется набор неравенств

$$\gamma_i(\boldsymbol{x}^*) > \gamma_j(\boldsymbol{x}^*),$$

где $j \in \{1, \ldots, L\} \backslash \{i\}$.

Линейная машина

Максимальная точность на выборке $\tilde{S}_t$ соответствует выполнению максимального числа блоков неравенств:

$$\gamma_{J(1)}(\boldsymbol{x}_1) > \gamma_i(\boldsymbol{x}_1), i \in \{1, \ldots, L\} \backslash \{J(1)\} \tag{1}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\gamma_{J(m)}(\boldsymbol{x}_m) > \gamma_i(\boldsymbol{x}_m), i \in \{1, \ldots, L\} \backslash \{J(m)\}.$$

Каждый из блоков соответствует одному из объектов выборки $\tilde{S}_t$ и включает $L-1$ неравенств. Таким образом суммарное число неравенств во всех блоках составляет $m(L-1)$. Каждое из неравенства из системы (1) соответствует сравнению оценки вектора $\boldsymbol{x}_r$ за класс $K_{J(r)}$ с оценкой за класс $K_i \neq K_{J(r)}$

Линейная машина. Пример

Имеется задача распознавания с 3-я классами и 2-я признаками. Предполагается, что с использованием метода ЛМ для каждого класса найдены линейные разделяющие функции:

- $\gamma_1(X_1, X_2) = 4 + 2X_1 - X_2$;
- $\gamma_2(X_1, X_2) = -2 + X_1 - 3X_2$;
- $\gamma_3(X_1, X_2) = 1 + X_1 - 2X_2$.

Область, где одновременно выполняются неравенства

- $\gamma_1(X_1, X_2) > \gamma_2(X_1, X_2)$;
- $\gamma_1(X_1, X_2) > \gamma_3(X_1, X_2)$;

соответствует классу 1.

Линейная машина. Пример

Последняя система эквивалентна неравенствам

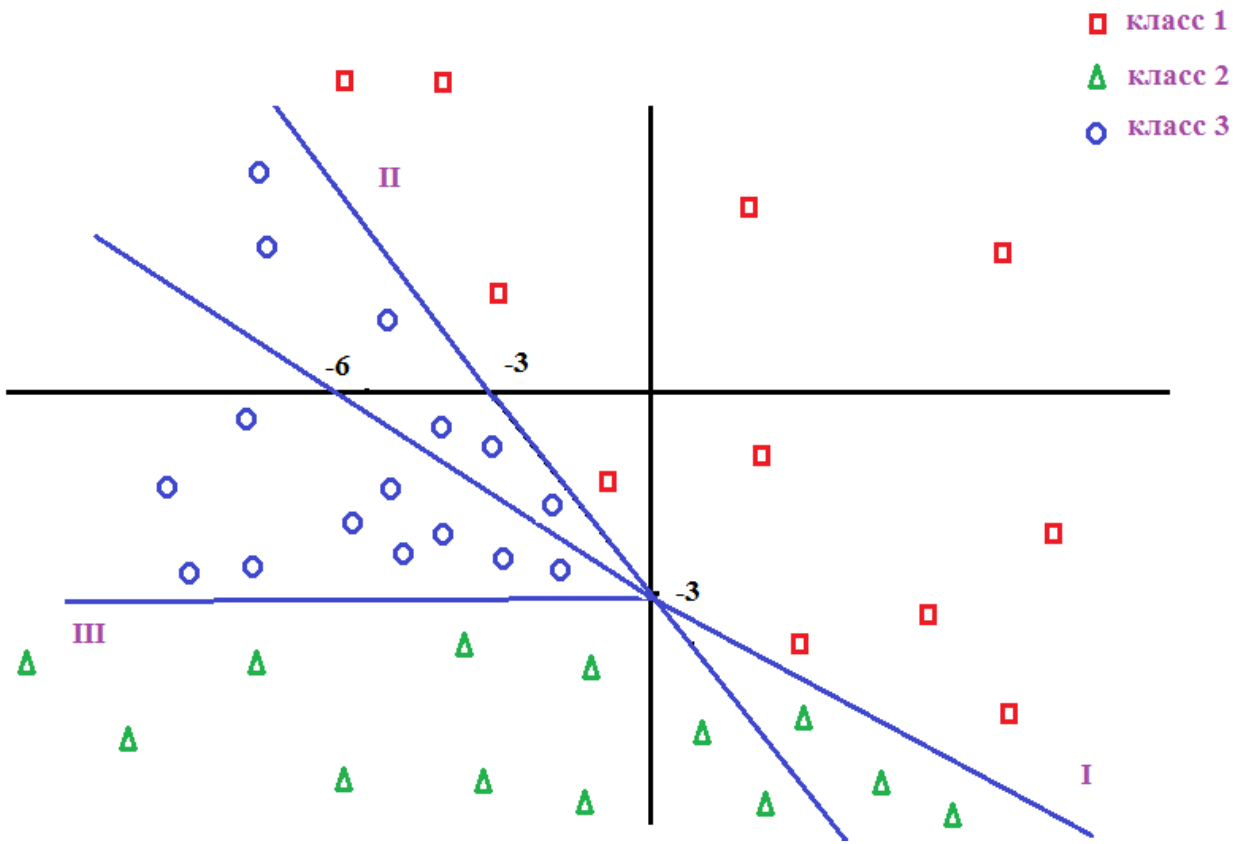
- $6 + X_1 + 2X_2 > 0$ (I),
- $3 + X_1 + X_2 > 0$ (II).

Данные неравенства задают граничные прямые на плоскости, которые обозначены римскими цифрами (I) и (II) соответственно. Область на плоскости, соответствующая классу 1, обозначена красными квадратиками. Предположим, что точка на плоскости не принадлежит классу 1. Тогда она принадлежит классу 2, если выполняется неравенство:

$$\gamma_1(X_1, X_2) > f_3(X_1, X_2),$$

которое эквивалентно неравенству $X_2 < -3$. Область на плоскости, соответствующая классу 2, обозначена зелёными треугольниками. Область, соответствующая классу 3 обозначена синими кружками.

Линейная машина. Пример



Линейная машина. Обучение

Был разработан эффективный метод поиска оптимальных весов

$$w_0^1, \ldots, w_n^1$$

$$\ldots\ldots\ldots$$

$$w_0^L, \ldots, w_n^L,$$

. который называется релаксационным алгоритмом.

Распознавание при заданной точности распознавания некоторых классов

Байесовский классификатор обеспечивает максимальную общую точность распознавания. Однако при решении конкретных практических задач потери, связанные с неправильной классификацией объектов, принадлежащих к одному из классов, значительно превышают потери, связанные с неправильной классификацией объектов других классов. Для оптимизации потерь необходимо использование методов распознавания с учётом предпочтительной точности распознавания для некоторых классов. Одним из возможных подходов является фиксирование порога для точности распознавания одного из классов. Оптимальное решающее правило в задаче распознавания с двумя классами K_1 и K_2 , обеспечивающее максимальную точность распознавания K_2 при фиксированной точности распознавания K_1 , описывается критерием Неймана-Пирсона.

Распознавание при заданной точности распознавания некоторых классов

Критерий Неймана-Пирсона Максимальная точность распознавания K_2 при точности распознавания K_1 равной α обеспечивается правилом: Объект с описанием $\boldsymbol{x}$ относится в класс K_1 , если

$$P(K_1 \mid \boldsymbol{x}) \geq \eta P(K_2 \mid \boldsymbol{x})$$

где параметр η определяется из условия

$$\int_{\Theta} P(K_1 \mid \boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \alpha$$

$$\Theta = \{\boldsymbol{x} \mid P(K_1 \mid \boldsymbol{x}) \geq \eta P(K_2 \mid \boldsymbol{x})\}$$

$p(\boldsymbol{x})$ - плотность распределения $K_1 \bigcup K_2$ в точке $\boldsymbol{x}$. Критерий Неймана-Пирсона может быть использован, если известны плотности распределения распознаваемых классов. Плотности могут быть восстановлены в рамках Байесовских методов обучения на основе гипотез о виде распределений. ,

Оценка эффективности распознавания

Основные, используемые в машинном обучении критерии эффективности диагностики и прогноза, должны вычисляться на контрольной выборке или в режиме скользящего контроля. Для оценки эффективности работы может использоваться точность распознавания в смысле доли правильных отнесений в классы. В англоязычной литературе данную характеристику принято обозначать термином accuracy.

$$\text{accuracy} = \frac{\text{число правильных классификаций}}{\text{общее число объектов}}$$

хорошо описывает ситуацию, когда классы (категории) примерно одинаково представлены в контрольной выборке. В противном случае оценки могут оказаться неадекватными.

Оценка эффективности распознавания

Часто при использовании методов МО ставится цель правильно диагностировать какие-либо нежелательные случаи, соответствующие, например, тяжёлому заболеванию или неблагоприятному исходу из совокупности случаев со сходной симптоматикой. Такого рода случаи мы далее будем называть целевыми. Совокупность всех целевых случаев назовём целевым классом. Долю правильно диагностированных целевых случаев принято называть чувствительностью. В англоязычной литературе наряду с термином чувствительность («sensitivity») используется также термин «recall»

$$\text{recall} = \frac{\text{число правильных классификаций из целевого класса}}{\text{общее число объектов целевого класса}}$$

Оценка эффективности распознавания

Долю правильно диагностированных случаев, не принадлежащих целевому классу, принято называть специфичностью. Наряду с критериями чувствительность и специфичность в англоязычной литературе используется также термин «precision», обозначающий долю правильно диагностированных случаев, среди всех случаев, классифицированных как целевые.

$$\text{precision} = \frac{\text{число правильных классификаций из целевого класса}}{\text{общее число объектов, отнесённых в целевой класс}}$$

Оценка эффективности распознавания

Другим общепринятым критерием эффективности диагностического алгоритма в теории машинного обучения является F-мера, определяемая как среднегармоническое критериев «recall» и «precision», то есть

$$F = \frac{2 * recall * precision}{recall + precision}$$

Критерии чувствительности, специфичности или F-меры более верно характеризуют успешность распознавания целевых случаев, чем критерий точности в смысле термина accuracy. Однако полной картины эффективности использования метода машинного обучения указанные критерии также не дают.

ROC анализ.

Решающее правило производит отнесение объекта s по вектору оценок $\gamma_1, \ldots, \gamma_L$ к одному из классов $K_1, \ldots, K_L$. Распространённым решающим правилом является простая процедура, относящая объект в тот класс, оценка за который максимальна.
В случае распознавания двух классов K_1 и K_2 распознаваемый объект s будет отнесён к классу K_1 , если $\gamma_1(s) - \gamma_2(s) > 0$ и классу K_2 в противном случае. Назовём приведённое выше правило правилом $\boldsymbol{C}(0)$. Однако точность распознавания правила $\boldsymbol{C}(0)$ может оказаться слишком низкой для того, чтобы обеспечить требуемую величину потерь, связанных с неправильной классификацией объектов, на самом деле принадлежащих классу K_1 . Для достижения необходимой величины потерь может быть использовано пороговое решающее правило $\mathbf{C}(\delta)$.

ROC анализ.

Прравило $\mathbf{C}(\delta)$: распознаваемый объект s будет отнесён к классу K_1 , если $\gamma_1(s) - \gamma_2(s) > \delta$ и классу K_2 в противном случае. Обозначим через $p_{ci}(\delta, s)$ вероятность правильной классификации правилом $\mathbf{C}(\delta)$ объекта s , на самом деле принадлежащего K_i , $i \in \{1, 2\}$. При $\delta < 0$ $p_{c1}(\delta, s) \geq p_{c1}(0, s)$, но $p_{c2}(\delta, s) \leq p_{c2}(0, s)$. Уменьшая δ , мы увеличиваем $p_{c1}(\delta, s)$ и уменьшаем $p_{c2}(\delta, s)$. Напротив, увеличивая δ , мы уменьшаем $p_{c1}(\delta, s)$ и увеличиваем $p_{c2}(\delta, s)$. Зависимость между $p_{c1}(\delta, s)$ и $p_{c2}(\delta, s)$ может быть приближённо восстановлена по обучающей выборке $\widetilde{S}_t$, включающей описания объектов $\{s_1, \ldots, s_m\}$.

Распознавание при заданной точности распознавания некоторых классов

Пусть
$\gamma_1(s_1) \ldots \gamma_1(s_m)$
$\gamma_2(s_1) \ldots \gamma_2(s_m)$
является матрицей оценок за классы K_1 и K_2 объектов из $\widetilde{S}_t$. Пусть

$$\gamma(s_1) = \gamma_1(s1) - \gamma_2(s_1), \ldots, \gamma(s_m) = \gamma_1(s_m) - \gamma_2(s_m).$$

Предположим, что величины $[\gamma(s_1), \ldots, \gamma(s_m)]$ принимают r значений $\Gamma_1, \ldots, \Gamma_r$. Рассмотрим r пороговых решающих правил $[\boldsymbol{C}(\Gamma_1), \ldots, \boldsymbol{C}(\Gamma_r)]$ Для каждого из правил $\mathbf{C}(\Gamma_i)$ обозначим через $\nu_{c1}(\Gamma_i)$ долю K_1 среди объектов обучающей выборки, удовлетворяющих условию $\gamma(s_*) \geq \Gamma_i$, а через $\nu_{c2}(\Gamma_i)$ обозначим долю K_2 среди объектов обучающей выборки, удовлетворяющих условию $\gamma(s_*) < \Gamma_i$.

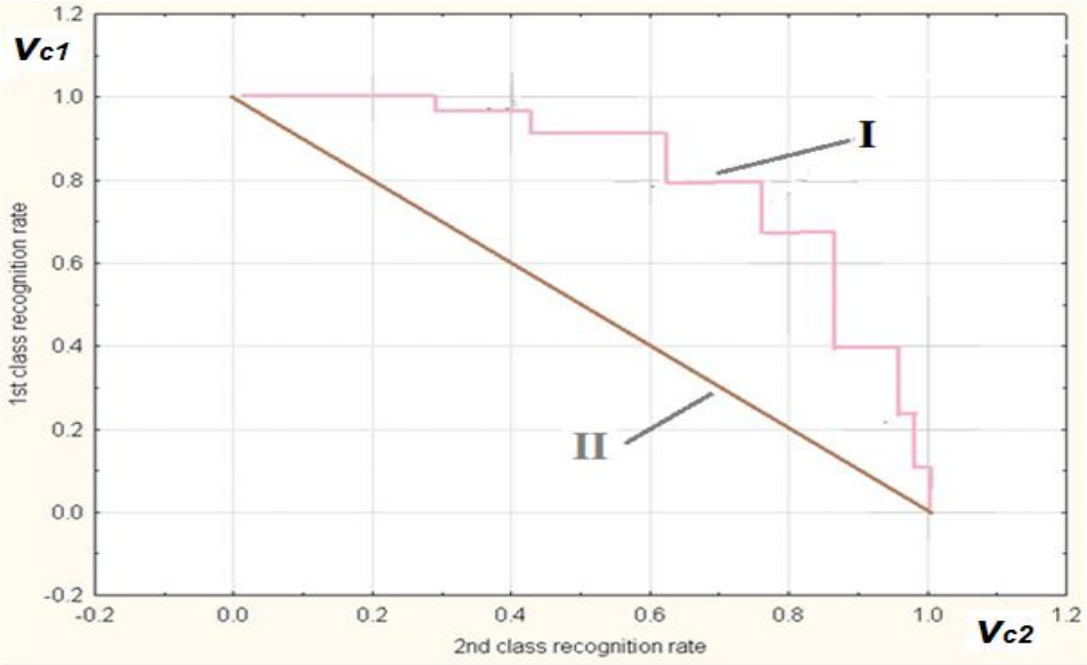
Отобразим результаты расчётов

$$\{[\nu_{c1}(\Gamma_1), \nu_{c2}(\Gamma_1)] \ldots, [\nu_{c1}(\Gamma_r), \nu_{c2}(\Gamma_r)]\}$$

как точки на в декартовой системе координат. Соединив точки отрезками прямых, получим ломаную линию (l), соединяющую точки (1,0) и (0,1). Данная линия графически отображает аппроксимацию по обучающей выборке взаимозависимости между $p_{c1}(\delta, s)$ и $p_{c2}(\delta, s)$ при всевозможных значениях δ . Соответствующий пример представлен на рисунке 2.

ROC анализ.

Взаимозависимость между ν_{c1} и ν_{c2} наиболее полно оценивает эффективность распознающего оператора $\mathbf{R}$. Отметим, что ν_{c1} постепенно убывает по мере роста ν_{c2}.

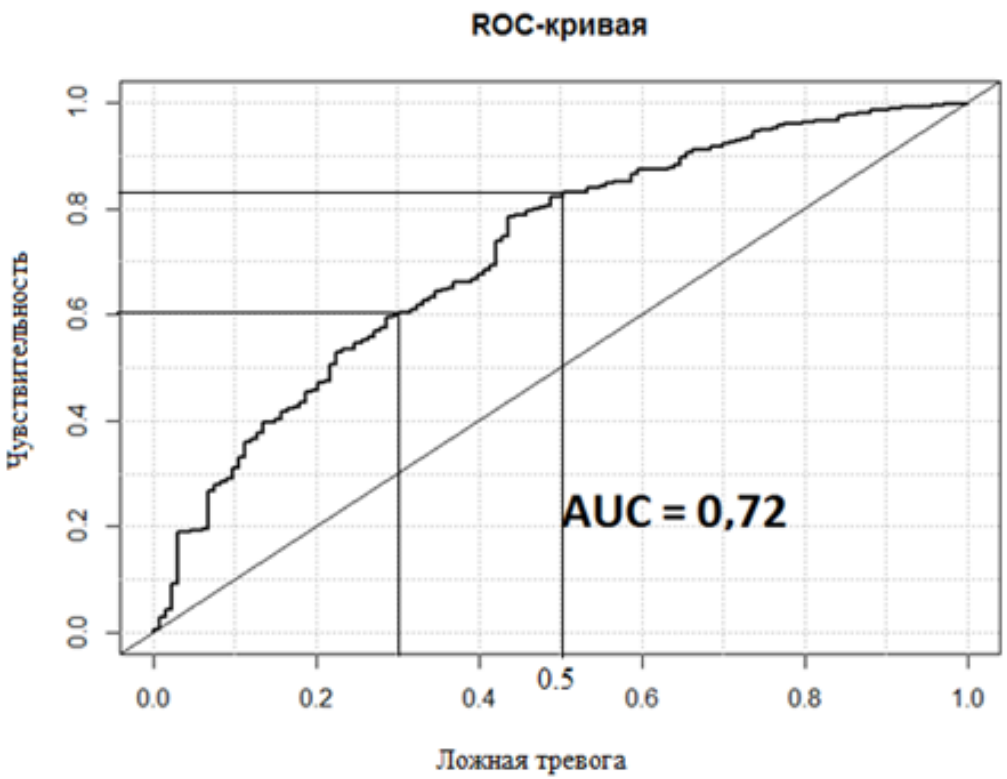


Кривая, связывающая точность распознавания классов K_1 и K_2.

ROC анализ.

Обычно эффективность метода распознавания иллюстрируется с помощью кривой, связывающей «Чувствительность» - доля правильно распознанных объектов целевого класса «Ложная тревога» - доля объектов ошибочно отнесённых в целевой класс. Пример кривой, связывающей параметры «Чувствительность» и «Ложная тревога» представлен на рисунке 4. Анализ, основанный на построении и анализе линий, связывающих параметры «Чувствительность» и «Ложная тревога» принято называть анализом Receiver Operating Characteristic или ROC-анализом. Линии, связывающих параметры «Чувствительность» и «Ложная тревога» принято называть ROC-кривыми.

ROC анализ.



Пример ROC кривой для задачи прогнозирования повторения ОКС в течение полугода после выписки их стационара.
Из рисунка 1 видно, что 60% всех случаев возникновения осложнений в течение полугода выявляется при уровне ложной тревоги, равном 30%.

ROC анализ.

ROC–кривая позволяет увидеть, при каких уровнях ложной тревоги может достигаться тот или иной уровень чувствительности. Кроме того, ROC-анализ позволяет оценить эффективность использования метода машинного обучения для рассматриваемой задачи в целом, то есть безотносительно к конкретному решающему правилу. Очевидно, что чем выше уровень чувствительности, при каждом заданном уровне ложной тревоги, тем эффективнее используемый метод машинного обучения. В свою очередь более высокое прохождение ROC–кривой соответствует большей площади под ней. Поэтому в качестве меры эффективности того или иного метода машинного обучения для рассматриваемой задачи принято использовать параметр AUC (area under curve), равный площади под ROC–кривой. Для задачи прогнозирования повторного возникновения острого коронарного синдрома (ОКС) в течение полугода параметр AUC равен 0,72.

ROC анализ.

Наряду с показателем AUC принято также использовать коэффициент Джини $Gini = 2 * AUC - 0.5$

ROC анализ.

Параметр AUC может быть вычислен по контрольной выборке. Для этого достаточно знать оценки за целевой класс и информацию об истинной принадлежности целевому классу представленных в контрольной выборке случаев. Подчеркнём ещё раз, что перечисленные выше параметры «точность» («accuracy»), «чувствительность», «специфичность», AUC отображают истинную точность алгоритма распознавания или метода машинного обучения только в том случае, если они рассчитаны на новых случаях, которые не использовались для настройки (обучения) алгоритмов.

Использование ROC анализа для максимизации дохода банка

Банк использует 2 метода распознавания для повышения прибыли при кредитовании. Используемая технология основана на распознавании в заёмщиков, для которых риск отказа от выплат по кредиту является высоким. Предполагается, что доход банка с одного добросовестного заёмщика составляет d=10000 условных единиц (у.е.). Потери банка при отказе от выплат по кредиту составляет L=45000 у.е. Доля заёмщиков, отказывающихся от выплат по кредиту составляет p=0.05. В таблице приведены значения чувствительности и ложной тревоги при некотором наборе пороговых значений для методов распознавания A и B.

Использование ROC анализа для максимизации дохода банка

Таблица 1

Метод A		Метод B	
Чувствительность	Ложная тревога	Чувствительность	Ложная тревога
0.03	0.001	0.03	0.001
0.08	0.002	0.16	0.002
0.13	0.01	0.28	0.02
0.19	0.03	0.44	0.06
0.27	0.07	0.57	0.08
0.34	0.09	0.61	0.09
0.47	0.11	0.67	0.11
0.61	0.14	0.69	0.14
0.74	0.17	0.72	0.17
0.91	0.21	0.78	0.2
0.97	0.24	0.83	0.23
	0.28	0.88	0.27
1			
		0.92	0.32
		0.98	0.35
		1	0.37

Вопросы. Позволяют ли приведённые в таблице 1 данные сделать вывод о потенциальной возможности увеличении дохода банка при использовании метода A или метода B? Какой из двух методов позволяет получить более высокий доход?

Вопросы

Позволяют ли приведённые в таблице 1 данные сделать вывод о потенциальной возможности увеличении дохода банка при использовании метода A или метода B? Какой из двух методов позволяет получить более высокий доход?

Использование ROC анализа для максимизации дохода банка

Средний доход банка на одну поданную заявку на кредит в случае, когда методы распознавания не используются очевидно может быть найден по формуле
$D = d * (1 - p) - p * L = 10000 * 0.95 - 0.05 * 45000 = 7250$ При использовании метода распознавания с чувствительностью Sen и уровнем ложной тревоги Fa . Величина потерь, произошедших непосредственно из-за отказов от выплат по кредиту, которая без применения методов распознавания была равна $p * L$, становится равной $pL(1 - Sen)$.

Использование ROC анализа для максимизации дохода банка

Величина дохода, полученная на добросовестных заёмщиков, которая без применения методов распознавания была равна $d * (1 - p)$, в случае применения метода распознавания оказывается равной $d * (1 - p)(1 - Fa)$. Таким образом величина дохода в случае использование метода распознавания рассчитывается по формуле

$$D = d * (1 - p)(1 - Fa) - pL(1 - Sen)$$

Проверка строк таблицы 1 показывает, что наибольший доход по формуле достигается для метода B при чувствительности 0.57 и ложной тревоге 0.08.