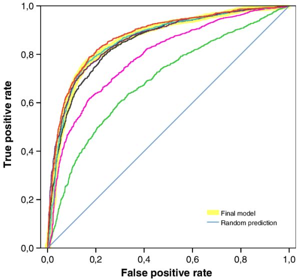
Оценка классификаторов

Виктор Китов

victorkitov.github.io



Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Содержание

Матрица ошибок

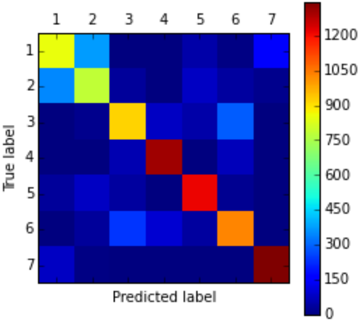
Матрица ошибок (confusion matrix) - таблица сопряженности между y и $\widehat{y}$:

$$\begin{array}{cc} & \text{Прогноз } \widehat{y} \\ & \begin{array}{cccc} 1 & 2 & \cdots & C \end{array} \\ \text{Истинный класс } y \quad \begin{array}{c} 1 \\ 2 \\ \vdots \\ C \end{array} & \begin{bmatrix} n_{11} & n_{12} & & \\ n_{21} & n_{22} & & \\ & & \ddots & \\ & & & n_{CC} \end{bmatrix} \end{array}$$

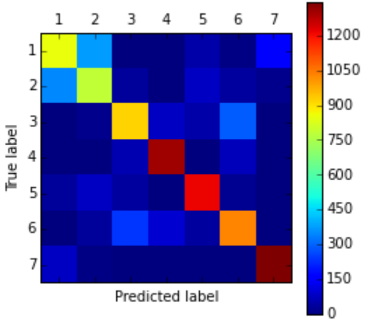
Корректные классификации - на диагонали.
Ошибки классификации - вне диагонали.

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} n_{cc}}{\sum_{i,j=1}^{C} n_{ij}}; \quad \text{ErrorRate} = 1 - \text{Accuracy} = \frac{\sum_{i,j=1;i\neq j}^{C} n_{ij}}{\sum_{i,j=1}^{C} n_{ij}}$$

Визуализация матрицы ошибок



Визуализация матрицы ошибок



- Видим, что ошибки сконцентрированы на разделении классов 1 и 2.
- Вариант решения:
  - объединим классы 1 и 2 в новый класс «1+2»
  - решим задачу классификации на {«1+2»,3,4,5,6,7}
  - разделим множество «1+2» отдельным классификатором.

Случай 2х классов

Матрица ошибок:

		Прогноз		
		+	-	всего
Факт	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
	всего	$\widehat{P}$	$\widehat{N}$	

Случай 2х классов

Матрица ошибок:

Прогноз

Факт

	+	-	всего
+	TP (true positives)	FN (false negatives)	P
-	FP (false positives)	TN (true negatives)	N
всего	$\widehat{P}$	$\widehat{N}$	

Точность:	$\frac{TP+TN}{P+N}$
Частота ошибок:	1-точность$=\frac{FP+FN}{P+N}$

Случай 2х классов

Матрица ошибок:

		Прогноз		
		+	-	всего
Факт	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
	всего	$\widehat{P}$	$\widehat{N}$	

Точность:	$\frac{TP+TN}{P+N}$
Частота ошибок:	1-точность$=\frac{FP+FN}{P+N}$

Точность и частота ошибок не информативны для неравномерного распределения классов.

Метрики качества для положительного класса

Точность (Precision)	$\frac{TP}{\widehat{P}}$
Полнота (Recall), TPR	$\frac{TP}{P}$
F-мера	$\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$
Взвешенная F-мера	$\frac{1}{\frac{\beta^2}{1+\beta^2}\frac{1}{Precision}+\frac{1}{1+\beta^2}\frac{1}{Recall}}$

TPR, recall	$\frac{TP}{P}$
FPR	$\frac{FP}{N}$

- Доля правильных положительных классификаций TPR
  - true positive rate, recognition rate
- Доля неправильных положительных классификаций FRP
  - false positive rate, false alarm.

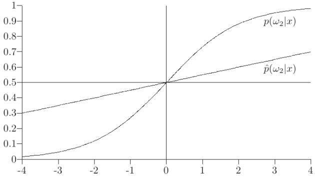
Содержание

Оценка прогнозов меток и вероятностей

- **Дискриминационные (discriminability) метрики качества** оценивают качество предсказания меток классов.
  - примеры: частота ошибок, точность, полнота, и т.д.

Оценка прогнозов меток и вероятностей

- **Дискриминационные (discriminability) метрики качества** оценивают качество предсказания меток классов.
  - примеры: частота ошибок, точность, полнота, и т.д.
- **Вероятностные (reliability) метрики качества** оценивают качество предсказания вероятностей классов.
  - условное правдоподобие выборки $\prod_{i=1}^{N} \widehat{p}(y_i|x_i)$
  - оценка Бриера (Brier score): $\frac{1}{N}\sum_{n=1}^{N} \left\|\mathbf{p}_n - \widehat{\mathbf{p}}_n\right\|^2$



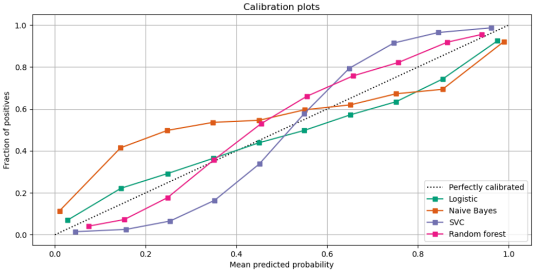
Пример: метки прогнозируются хорошо, а классы плохо.

Предсказание вероятностей

- Любой классификатор можно научить предсказывать вероятности классов подобрав $f(\cdot)$:

$$p(y=+1|x)=f(g(x))$$

- График калибровки (calibration plot) - соответствие между $p(y=+1|x)$ и $f(g(x_n))$ для объектов из ячеек $\{\beta_i \leq f(g(x)) \leq \beta_{i+1}\}_i$:



Основные подходы для предсказания вероятностей

- шкалирование Платта
  - A, B находятся методом максимального правдоподобия по отложенной выборке

$$p(y=+1|x)=\frac{1}{1+exp(Ag(x)+B)}$$

- напрямую по графику калибровки (гистограмме)
- с помощью изотонической регрессии (isotonic regression)

$$\begin{cases}\sum_{n=1}^{N}\left(\widehat{y}_n-y_n\right)^2\to\min_{\widehat{y}_1,\ldots\widehat{y}_N}\\ \widehat{y}_j\ge\widehat{y}_i\quad\forall(i,j):\;\; x_j\ge x_i\end{cases}$$

$$\begin{cases}\sum_{n=1}^{N}\left(\widehat{p}_n-p_n\right)^2\to\min_{\widehat{p}_1,\ldots\widehat{p}_N}\\ \widehat{p}_j\ge\widehat{p}_i\quad\forall(i,j):\;\; g_j\ge g_i\end{cases}$$

Содержание

Решающее правило бинарной классификации

- Используем относительный рейтинг $g(x) = g_{+1}(x) - g_{-1}(x)$.
- Классификация $\widehat{y}(x) = \text{sign}(g(x))$.
- Введем параметр $\alpha \in \mathbb{R}$, контролирующий предпочтения между классами:

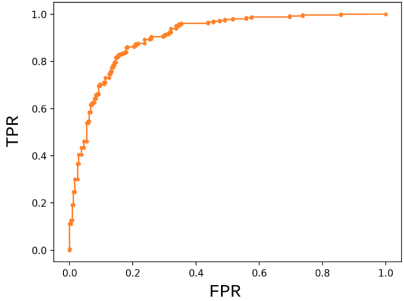
$$\widehat{y}(x) = \text{sign}(g(x) - \alpha)$$

  - $\downarrow \alpha$: больше $\widehat{y} = +1$; $\uparrow \alpha$: меньше $\widehat{y} = +1$.
- Можем обучить модель 1 раз, а потом использовать в разных режимах:
  - детекция самолетов в мирное/военное время
  - выдача кредитов в период экономического бума/спада
- В случае неравных потерь $\lambda_{+1} \neq \lambda_{-1}$ тоже нужно подбирать α:
  - $\lambda_{+1} = \text{cost}(\widehat{y} = -1 | \ y = +1)$; $\lambda_{-1} = \text{cost}(\widehat{y} = +1 | \ y = -1)$

ROC кривая (receiver operating characteristic)

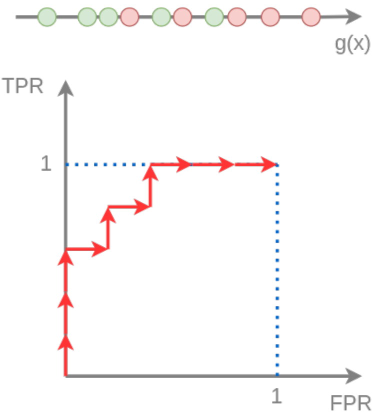
- $TPR = TPR(\alpha)$, $FPR = FPR(\alpha)$.
- ROC кривая- функция TPR(FPR).

$$TPR = \frac{TP}{P} \qquad FPR = \frac{FP}{N}$$



- Как TPR и FPR изменяются с α?

Построение ROC-кривой по выборке



- Сдвигаемся справа налево вдоль $g(x)$.
  - при пересечении положительного объекта: $\uparrow$ на $1/N_{+}$.
  - при пересечении отрицательного объекта: $\rightarrow$ на $1/N_{-}$.

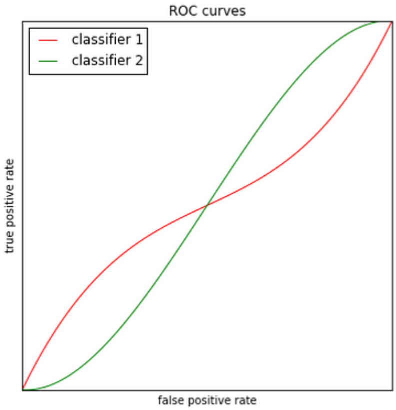
Вопросы

- Более высота ROC кривой связана с качеством классификации?
- Какова ROC-кривая для случайного угадывания $\widehat{y}(x) = \text{sign}(\xi - \alpha)$, $\xi \sim \textit{Uniform}[0, 1]$?
- Как улучшить классификатор для вогнутой ROC кривой?
- Как поменяется ROC кривая при инвертировании классификатора:

$$\text{sign}\left(g(x) - \alpha\right) \longrightarrow \text{sign}\left(\alpha - g(x)\right)$$

Композиции классификаторов

Как создать семейство классификаторов с максимально высокой ROC-кривой в таком случае?

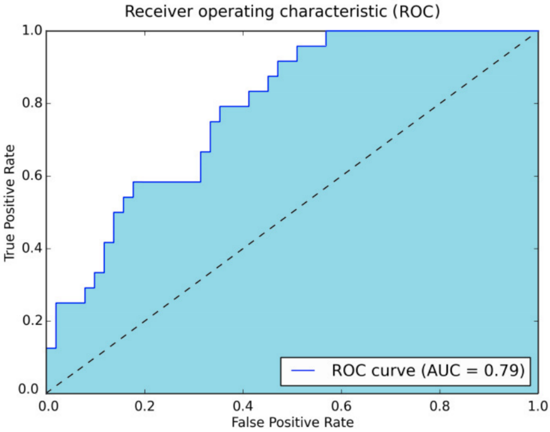


Преимущества ROC-кривой

- оценка семейства классификаторов, параметризованных α.
- инвариантность к монотонным преобразованиям $g(x) \to g'(x) = f(g(x)),\ \uparrow f(\cdot)$
    - $\text{sign}(g(x) - \alpha) \Longleftrightarrow \text{sign}(f(g(x)) - f(\alpha))$, поэтому точке $TPR(\alpha), FPR(\alpha)$ соответствует $TPR'(f(\alpha)), FPR'(f(\alpha))$.

Площадь под кривой

- Площадь под ROC-кривой (area under curve, AUC) - интегральная мера качества семейства классификаторов.



Эквивалентное определение AUC

- **AUC=доле корректно упорядоченных пар:**
- выберем случайно
    - отрицательный объект $(x_i, y_i = -1)$ и положительный объект $(x_j, y_j = +1)$
    - тогда AUC - вероятность верного упорядочивания таких объектов

$$AUC = p\left(g(x_i) < g(x_j)\right)$$

Эквивалентное определение AUC

- **AUC=доле корректно упорядоченных пар:**
- выберем случайно
  - отрицательный объект $(x_i, y_i = -1)$ и положительный объект $(x_j, y_j = +1)$
  - тогда AUC - вероятность верного упорядочивания таких объектов

$$AUC = p\left(g(x_i) < g(x_j)\right)$$

- Для конечной выборки:

$$AUC = \frac{\sum_{(i,j):y_i=-1,y_j=1} \mathbb{I}\left[g(x_j) > g(x_i)\right]}{\#[i : y_i = -1]\#[j : y_j = 1]}$$

AUC=доля верно упорядоченных пар объектов

$x_{(1)},...x_{(N)}$- упорядоченные объекты по рейтингу:

$$g\left(x_{(1)}\right) < g\left(x_{(2)}\right) < ... < g\left(x_{(N)}\right)$$

$$\widehat{y}_k(x) = \text{sign}\left(g(x) \ge g(x_{(k)})\right),$$

$$TPR_k = \frac{\sum_{n=k}^{N} \mathbb{I}[y_{(n)} = +1]}{N_+}, \; FPR_k = \frac{\sum_{n=k}^{N} \mathbb{I}[y_{(n)} = -1]}{N_-},$$

$$TPR_k, FRP_k \; \downarrow \text{по } k.$$

$k = N$: 1ая точка, а $k = 1$: $-$последняя точка на ROC

AUC=доля верно упорядоченных пар объектов

Интегруруем справа-налево по формуле трапеций:

$$AUC = \sum_{k=1}^{N-1} \frac{TPR_{k+1} + TPR_k}{2} (FPR_k - FPR_{k+1})$$

$$= \sum_{k=1}^{N-1} \frac{\sum_{n=k+1}^{N} \mathbb{I}[y_{(n)} = +1] + \sum_{n=k}^{N} \mathbb{I}[y_{(n)} = +1]}{2N_+} \times$$

$$\times \left(\sum_{n=k}^{N} \mathbb{I}[y_{(n)} = -1] - \sum_{n=k+1}^{N} \mathbb{I}[y_{(n)} = -1] \right) =$$

$$= \sum_{k=1}^{N-1} \frac{\sum_{n=k+1}^{N} \mathbb{I}[y_{(n)} = +1] + \frac{1}{2}\mathbb{I}[y_{(k)} = +1]}{N_+} \cdot \frac{\mathbb{I}[y_{(k)} = -1]}{N_-}$$

$$= \frac{1}{N_+ N_-} \sum_{k=1}^{N-1} \sum_{n=k+1}^{N} \mathbb{I}[y_{(n)} = +1]\mathbb{I}[y_{(k)} = -1] = \frac{1}{N_+ N_-} \sum_{k<n} \mathbb{I}[y_{(k)} < y_{(n)}]$$

Сглаживание AUC

AUC - кусочно-постоянна из-за $\mathbb{I}[\cdot]$:

$$AUC = \frac{\sum_{(i,j):y_i=-1,y_j=1} \mathbb{I}\left[g(x_j) > g(x_i)\right]}{\#[i : y_i = -1]\#[j : y_j = 1]}$$

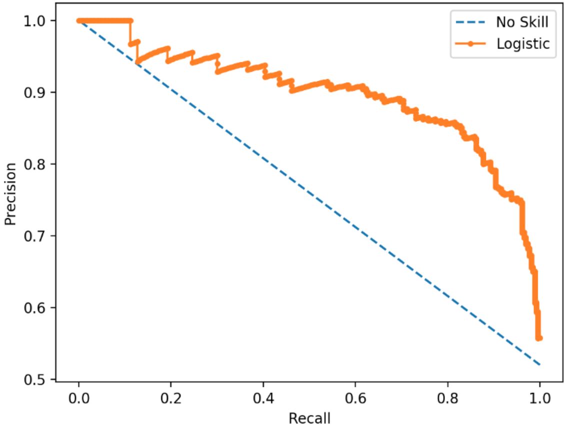
Сглаженная версия: $\mathbb{I}[u] \to \log\sigma(u) = \log(1/(1+e^{-u}))$

$$AUC' = \frac{\sum_{(i,j):y_i=-1,y_j=1} \log\sigma\left(g(x_j) > g(x_i)\right)}{\#[i : y_i = -1]\#[j : y_j = 1]}$$

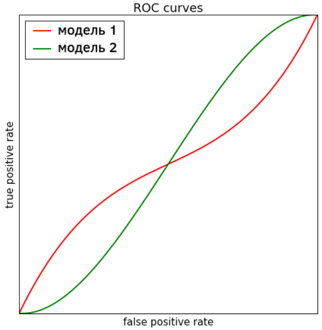
Можем оптимизировать по ней параметры модели!

Аналог ROC-кривой: точность(полнота)

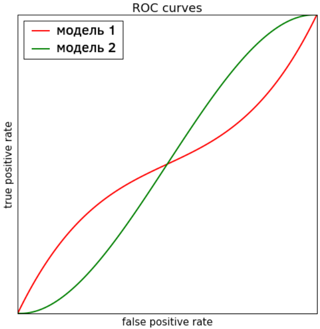
$$Precision = \frac{TP}{\widehat{P}} \qquad Recall = \frac{TP}{P}$$



Сравнение классификаторов по ROC-кривой



Сравнение классификаторов по ROC-кривой



Как сравнивать классификаторы?

- Фиксированные $\lambda_{+1}, \lambda_{-1}$: по точкам на ROC-кривой
- Неизвестные $\lambda_{+1}, \lambda_{-1}$: по AUC
- Частично известные $\lambda_{+1}, \lambda_{-1}$: по LC-индексу

Изолинии потерь

- Вероятности ошибок: $p(\widehat{y} = -1|y = +1) = 1 - TPR$, $p(\widehat{y} = +1|y = -1) = FPR$
- Изолиния потерь (потери$\equiv L$):

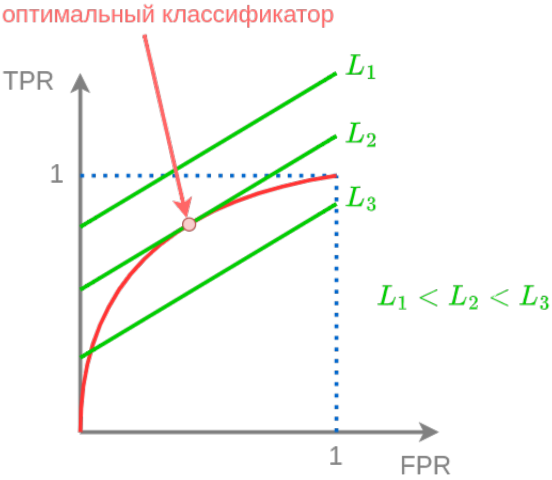
Изолинии потерь

- Вероятности ошибок: $p(\widehat{y}=-1|y=+1)=1-TPR$, $p(\widehat{y}=+1|y=-1)=FPR$
- Изолиния потерь (потери$\equiv L$):

$$L=p(y=+1)(1-TPR)\lambda_{+1}+p(y=-1)FPR\lambda_{-1}$$
$$(TPR-1)\,p(y=+1)\lambda_{+1}=-L+\lambda_{-1}p(y=-1)FPR$$
$$TPR=1+\frac{\lambda_{-1}p(y=-1)FPR-L}{\lambda_{+1}p(y=+1)}$$

- Оптимальный классификатор - точка касания изолинии и ROC-кривой
  - в ней тангенс угла наклона $\frac{\lambda_{-\mathbf{1}}p(y=-1)}{\lambda_{+\mathbf{1}}p(y=+1)}$

Лучший классификатор на ROC-кривой

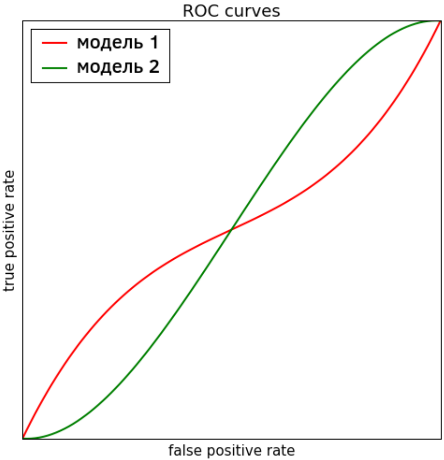


Площадь под кривой (AUC)

- Глобальная характеристика качества для различных α
- AUC$\in [0, 1]$
- AUC=0.5 - случайное угадывание
- AUC=1 - идеальное упорядочивание (безошибочная классификация при определенном α)

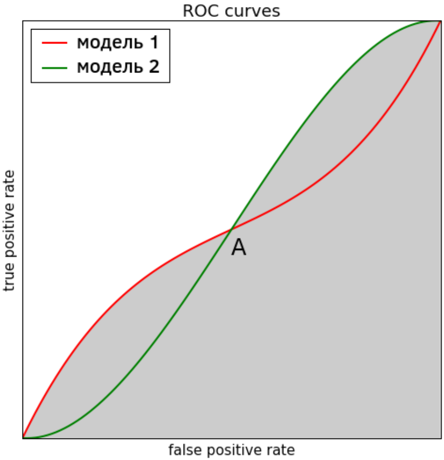
Композиция классификаторов

Рассмотрим 2 классификатора:



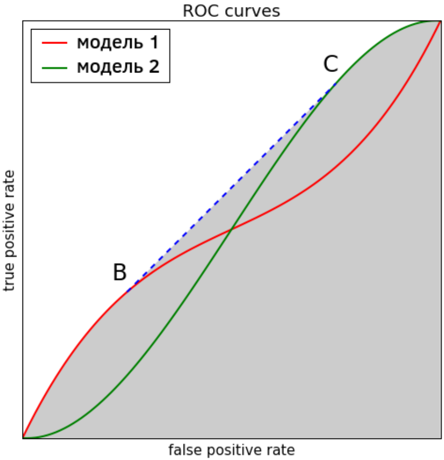
Композиция классификаторов

Как получить такую AUC?



Композиция классификаторов

Как получить такую AUC?



LC индекс

- Заданные $\lambda_{+1}, \lambda_{-1}$: слишком специфично.
- Неопределенные $\lambda_{+1}, \lambda_{-1}$: слишком общий случай.
- LC индекс вычисляют для промежуточного сценария.

1. отмасштабируем λ_{+1} и λ_{-1} так, что $\lambda_{+1} + \lambda_{-1} = 1$
2. определим $\lambda_1 = \lambda$, $\lambda_{-1} = 1 - \lambda$
3. для каждого $\lambda \in [0, 1]$ вычислим
$$S(\lambda) = \begin{cases} +1 & \text{если 1й классификатор лучше} \\ -1 & \text{если 2й классификатор лучше} \end{cases}$$
4. прикинем плотность распределения λ: $p(\lambda)$ (например, "треугольный" случай)
5. выберем классификатор 1, если $\int_0^1 S(\lambda)p(\lambda)d\lambda > 0$ иначе классификатор 2.

Точность и полнота - многоклассовый случай

	бинарный случай	макроусреднение	микроусреднение
Точность	$\frac{TP}{\widehat{P}}$	$\frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{\widehat{P}_c}$	$\frac{\sum_{c=1}^{C}TP_c}{\sum_{c=1}^{C}\widehat{P}_c}$
Полнота	$\frac{TP}{P}$	$\frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{P_c}$	$\frac{\sum_{c=1}^{C}TP_c}{\sum_{c=1}^{C}P_c}$

Обозначения:

- TP_c - # верно предсказанных объектов класса c.
- P_c - # объектов класса c.
- $\widehat{P}_c$ - # объектов, предсказанных как класс c.

Макроусреднение учитывает метрики равномерно по классам, а микроусреднение - по объектам.

Заключение

- Матрица ошибок дает больше информации, чем частота ошибок.
- Precision, recall и (TPR, FPR) используются для несбалансированных классов.
    - многоклассовый случай: микро или макроусреднение
- Можно оценивать качество
    - предсказания меток классов (accuracy, precision, recall)
    - упорядочивания по рейтингу (ROC-кривая, AUC)
    - вероятностей классов (оценка Бриера, условное правдоподобие)
- Площадь под ROC-кривой-вероятность верного упорядочивания всех пар объектов по рейтингу.