Прикладные задачи анализа данных

Функции ошибки / функционалы качества

Дьяконов А.Г.

Московский государственный университет имени М.В. Ломоносова (Москва, Россия)

Функции ошибки / функционалы качества

Пожалуй, самое главное, при решении задачи…

а что такое решение!

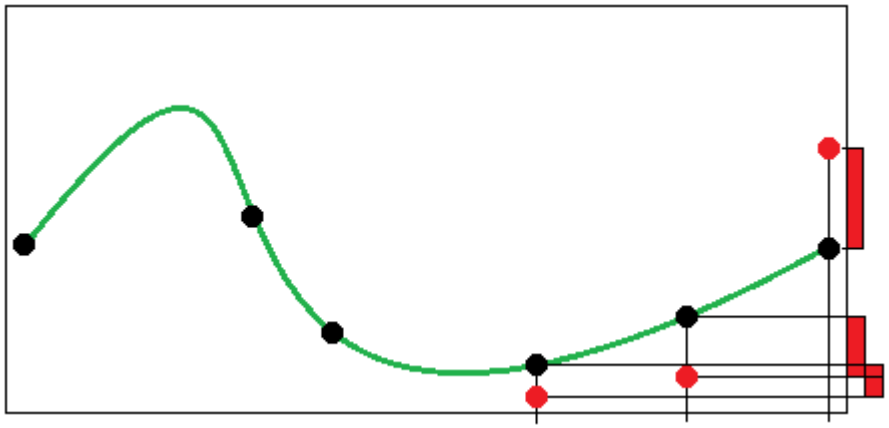
В анализе данных:

- **формализация ответа (формат)**
- **как ответ оценивается (критерий качества)**

Случай из практики: задача про траектории зрачка

(задача с 3 классами, а не с двумя)

Задача регрессии

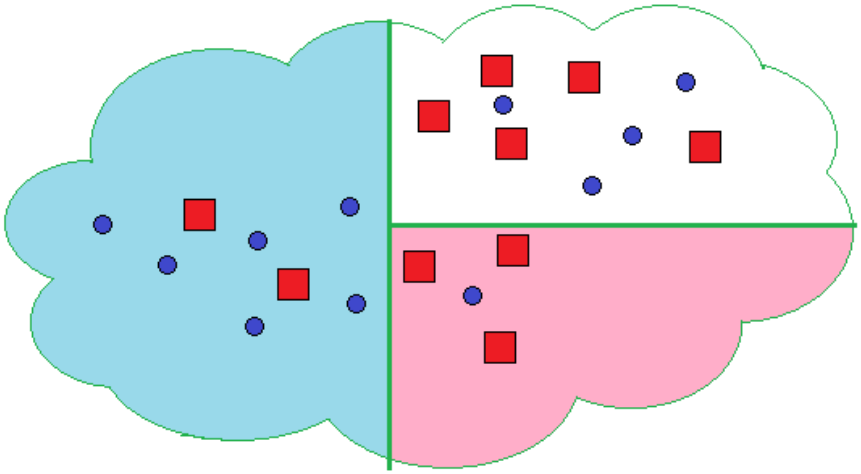


Будем дальше пытаться всё решать в классе констант

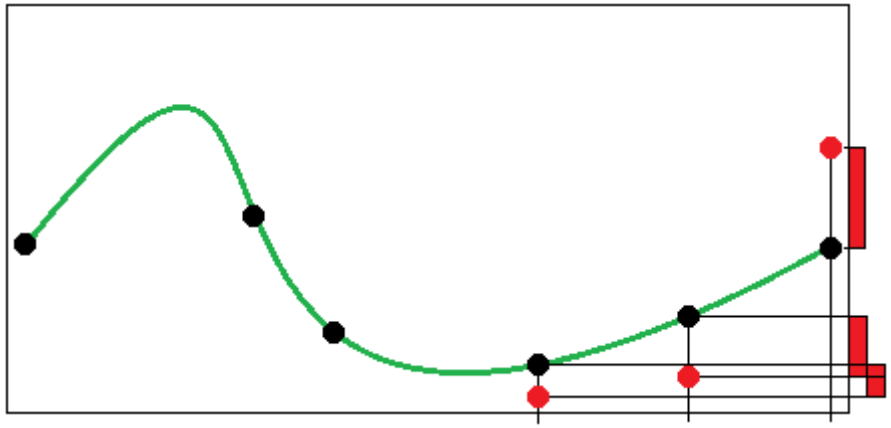
1. Простейшее решение

2. Примерно это и происходит в листьях обобщённых деревьев

3. Раскрывает природу функционалов

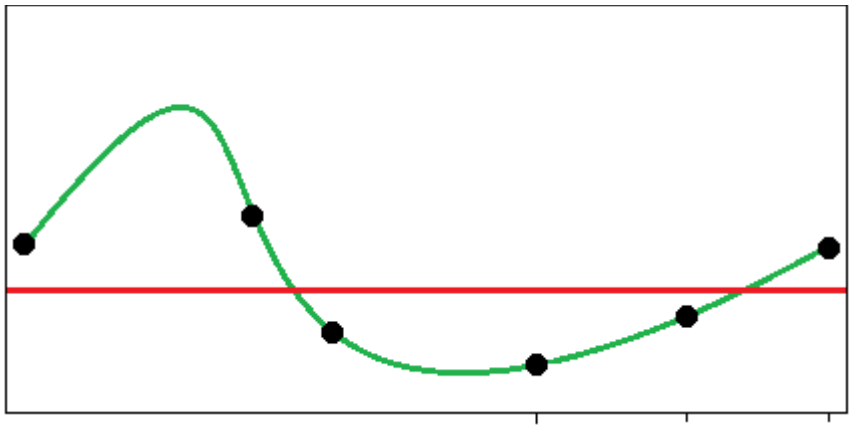


Средний модуль отклонения – Mean Absolute Error (MAE), Mean Absolute Deviation (MAD)



$$MAE = \frac{1}{q}\sum_{i=1}^{q} | a_i - y_i |$$

Напоминание:



$$\frac{1}{q}\sum_{i=1}^{q} | a - y_i | \to \min$$

$$a = \text{median}(\{y_i\}_{i=1}^{q})$$

Это открывает смысл решений!

Средний модуль отклонения

Способы использования тайных знаний:

- медиана, вместо усреднения, в ансамбле
- округление ответа (если целевой вектор целочисленный)

Средний квадрат отклонения ~ Mean Squared Error (MSE)

$$RMSE = \frac{1}{q}\sum_{i=1}^{q} | a_i - y_i |^2$$

$$\frac{1}{q}\sum_{i=1}^{q} | a - y_i |^2 \to \min$$

$$a = \frac{1}{q}\sum_{i=1}^{q} y_i$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{q}\sum_{i=1}^{q} | a_i - y_i |^2}$$

- **Способы использования тайных знаний**
- **ничего не делать (в RF, GBM и т.д. всё равно усредняют)**
- **метод НСКО – классическая регрессия!**

Обобщения

$$RMSE = \sqrt[p]{\frac{1}{q}\sum_{i=1}^{q} w_i \,|\, \varphi(a_i) - \varphi(y_i) \,|^p}$$

Рецепты

1. Преобразование целевого вектора $\varphi(y)$

2. Веса ~ вероятности появления объектов в сэмплировании

3. В случае нетривиальных p – прямая настройка

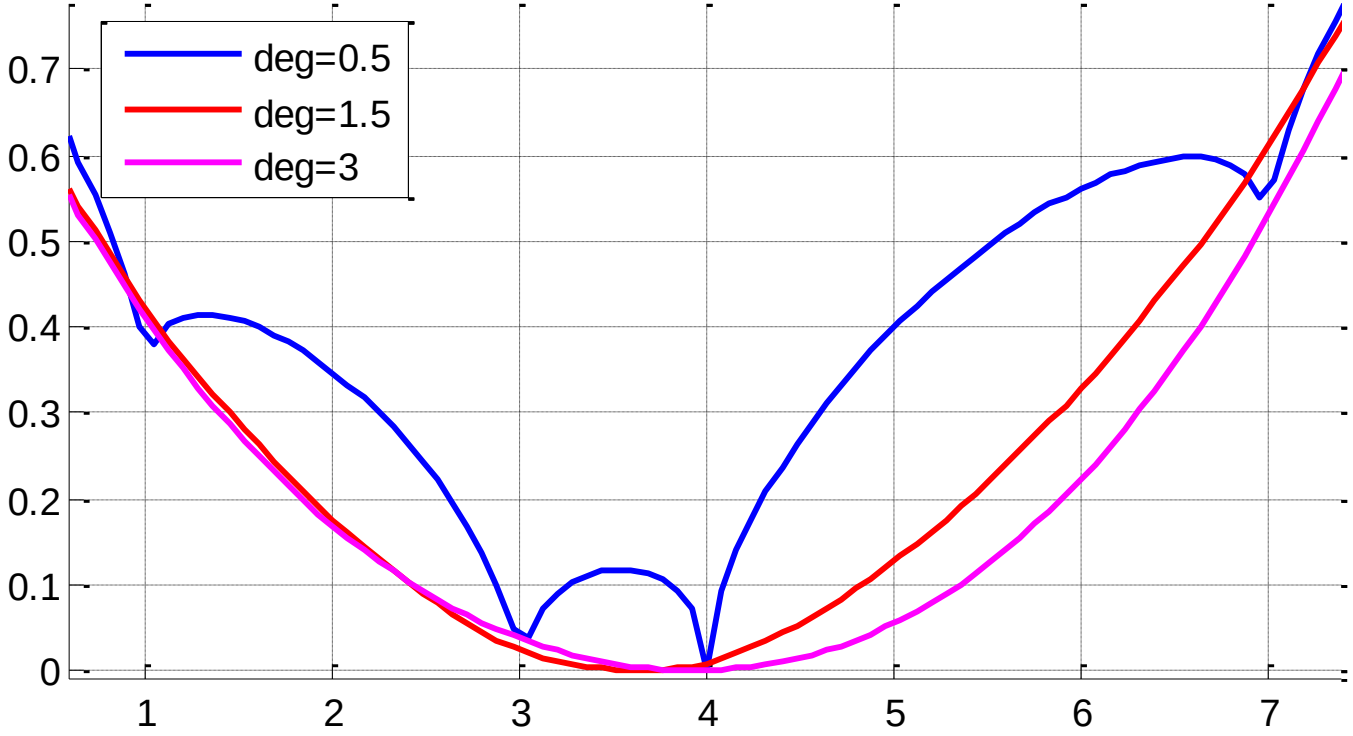
Прямая настройка – пример

$$F(\underbrace{B}_{\text{стандартная оптимизация}} \cdot \underbrace{C_c}_{\text{простое РР}}) \to \min_c$$

Как в задаче CrowdFlower (выбор порогов)

Есть обобщения, где берётся медиана, а не усреднение!

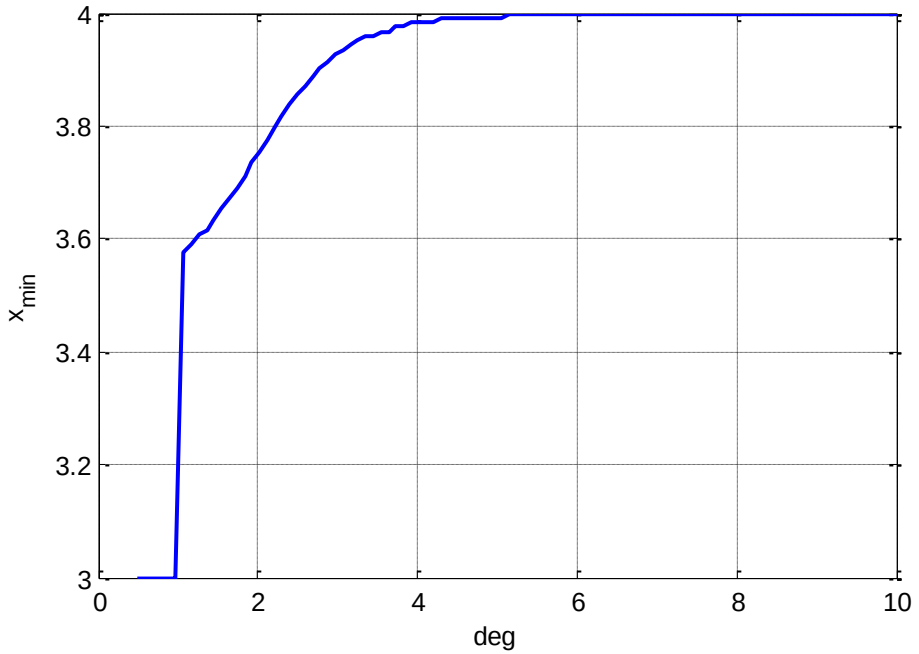
Про нетривиальные p




```
x = linspace(0,8,101);
f = @(a,b) abs(a-b).^0.5;
y = f(x,1) + f(x,3)+ f(x,4) + f(x,7);
y = (y-min(y));
y = y/max(y);
f = @(a,b) abs(a-b).^1.5;
y2 = f(x,1) + f(x,3)+ f(x,4) + f(x,7);
y2 = (y2-min(y2));
y2 = y2/max(y2);
f = @(a,b) abs(a-b).^3;
```

```
y3 = f(x,1) + f(x,3)+ f(x,4) + f(x,7);
y3 = (y3-min(y3));
y3 = y3/max(y3);
hold on; grid on
plot(x,y,'LineWidth',2)
plot(x,y2,'r','LineWidth',2)
plot(x,y3,'m','LineWidth',2)
legend('deg=0.5','deg=1.5','deg=3')
```

Как точка минимума зависит от степени



Symmetric mean absolute percentage error (SMAPE or sMAPE)

$$\mu = \frac{2}{q}\sum_{i=1}^{q}\frac{|y_i - a_i|}{y_i + a_i} = 100\% \cdot \frac{1}{q}\sum_{i=1}^{q}\frac{|y_i - a_i|}{(y_i + a_i)/2}$$

Когда надо интерпретировать погрешность как проценты

- плохо, если есть нули (и отрицательные значения)

Mean Absolute Percent Error (MAPE)

$$100\% \cdot \frac{1}{q}\sum_{i=1}^{q}\frac{|y_i - a_i|}{|y_i|}$$

PMAD

$$PMAD = \frac{\frac{1}{q}\sum_{i=1}^{q} | y_i - a_i |}{\sum_{i=1}^{q} | y_i |}$$

Меры на сравнении с бенчмарком

Классная идея:
сделать простой алгоритм и смотреть ошибку относительно его.

$$r_i = e_i / e_i'$$

Mean Relative Absolute Error (MRAE)

$$MRAE = \frac{1}{q}\sum_{i=1}^{q} | r_i |$$

Меры на сравнении с бенчмарком

а можно так...

$$REL\_MAE = \frac{\sum_{i=1}^{q} | y_i - a_i |}{\sum_{i=1}^{q} | y_i - a'_i |}$$

или Percent Better

$$PB(MAE) = \frac{1}{q} \sum_{i=1}^{q} I[| y_i - a_i | < | y_i - a'_i |]$$

Как выбрать бенчмарк в задачах прогнозирования?

Нормированные ошибки

Не зависят от шкалы…

$$q_t = \frac{e_t}{\frac{1}{q-1}\sum_{i=2}^{q} | y_i - y_{i-1} |}$$

Mean Absolute Scaled Error

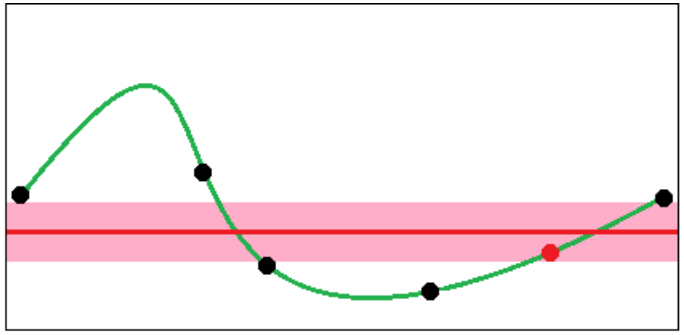
$$MASE = \frac{1}{q}\sum_{i=1}^{q} | q_i |$$

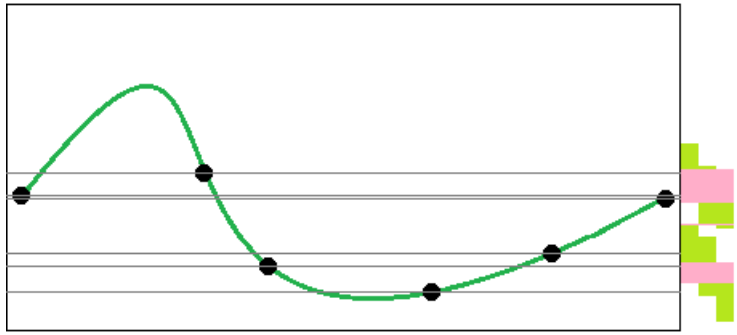
Какие ещё бывают функционалы в регрессии?

С точностью до порога

$$\frac{1}{q}\sum_{i=1}^{q} I[|\, y_i - a_i \,| < \varepsilon]$$

(это функционал качества) был в задаче Dunnhumby



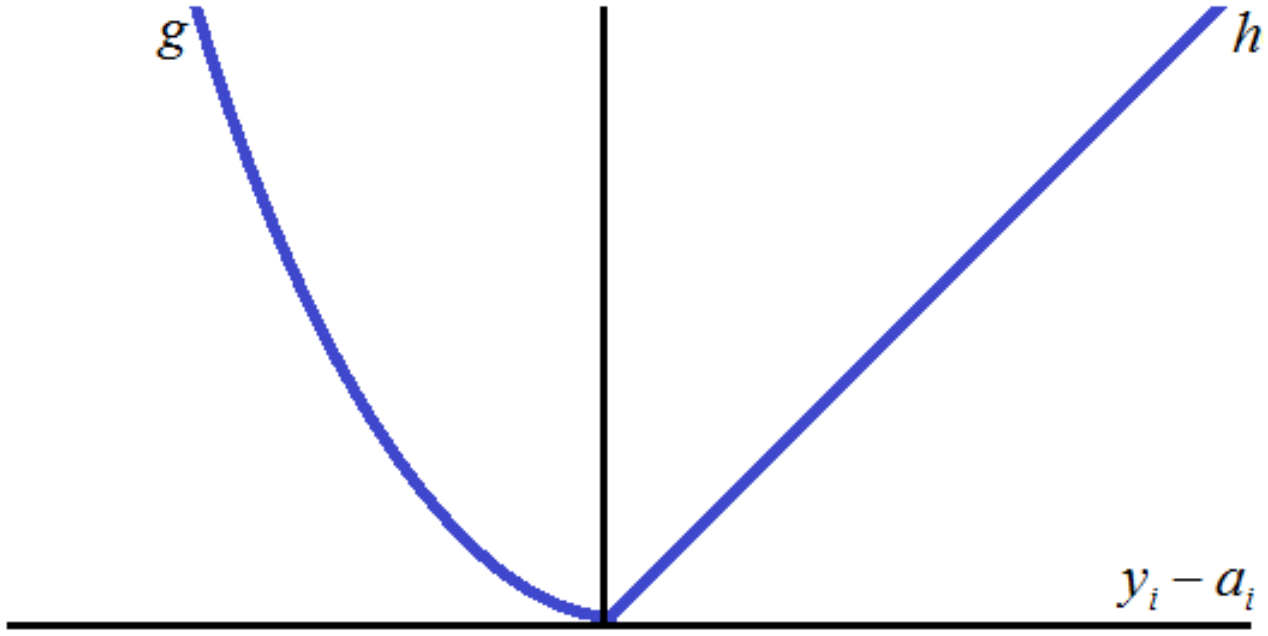


Минимизация графика ошибки (на вертикали)

Мы всегда его минимизируем – не только здесь!

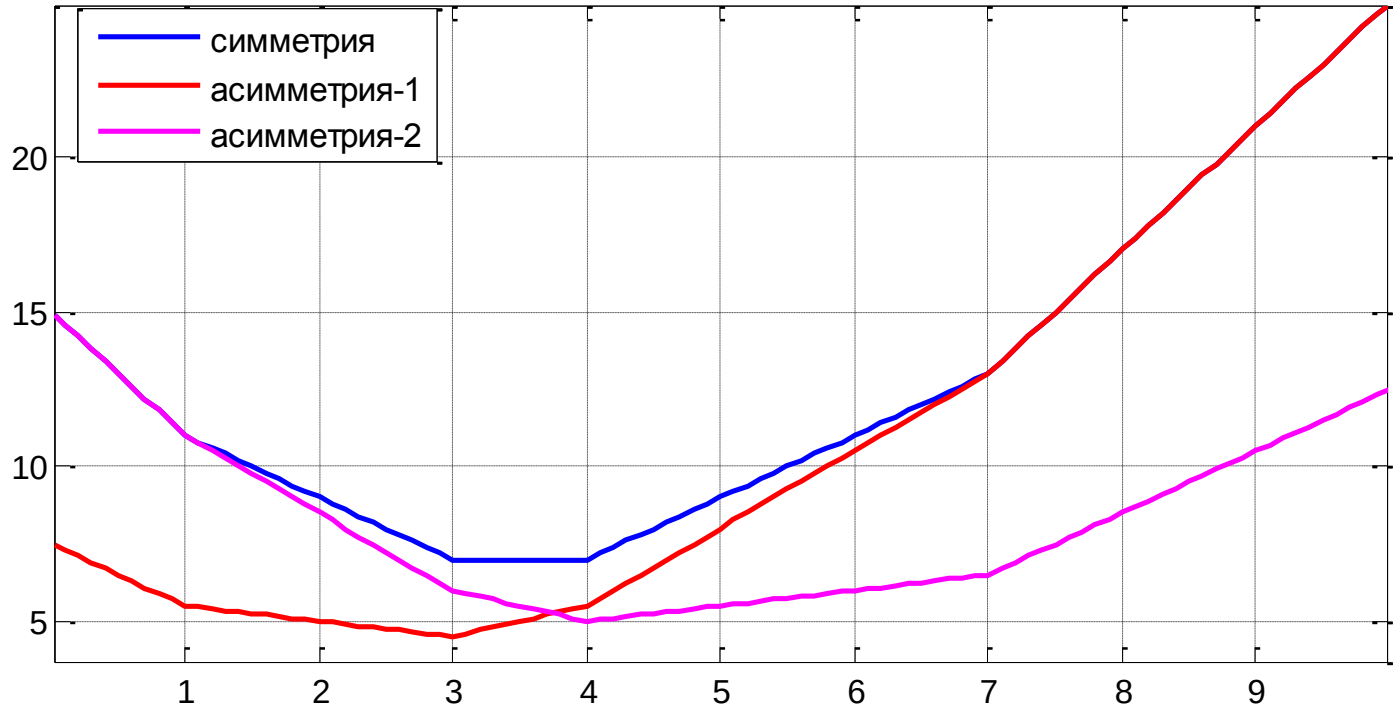
Несимметричные функции потерь

$$\frac{1}{q}\sum_{i=1}^{q}\begin{cases} g(|\,y_i - a_i\,|), & y_i < a_i, \\ h(|\,y_i - a_i\,|), & y_i \ge a_i, \end{cases}$$



Зачем нужны такие функции?

Несимметричные функции потерь




```
x = linspace(0,10,101);
y = abs(x-1) + abs(x-3)+ abs(x-4) + abs(x-7);
f = @(a,b) (a>b).*abs(a-b) + 0.5*(a<=b).*abs(a-b);
y2 = f(x,1) + f(x,3)+ f(x,4) + f(x,7);
f = @(a,b) 0.5*(a>b).*abs(a-b) + (a<=b).*abs(a-b);
y3 = f(x,1) + f(x,3)+ f(x,4) + f(x,7);
hold on; grid on
plot(x,y,'LineWidth',2)
plot(x,y2,'r','LineWidth',2)
plot(x,y3,'m','LineWidth',2)
legend('симметрия', 'асимметрия-1', 'асимметрия-2')
```

Совет

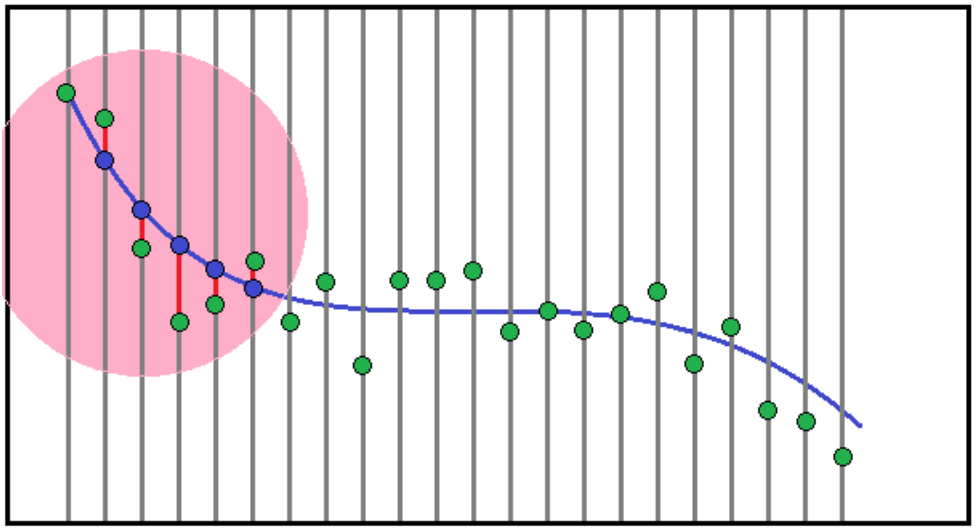
Функции ошибок иногда и классные признаки…

Пример: **в Casuality придумываем бенчмарки (восстановление одной переменной по другой), признаки – их относительные ошибки, т.к. абсолютные брать нельзя**

Почему?

Совет

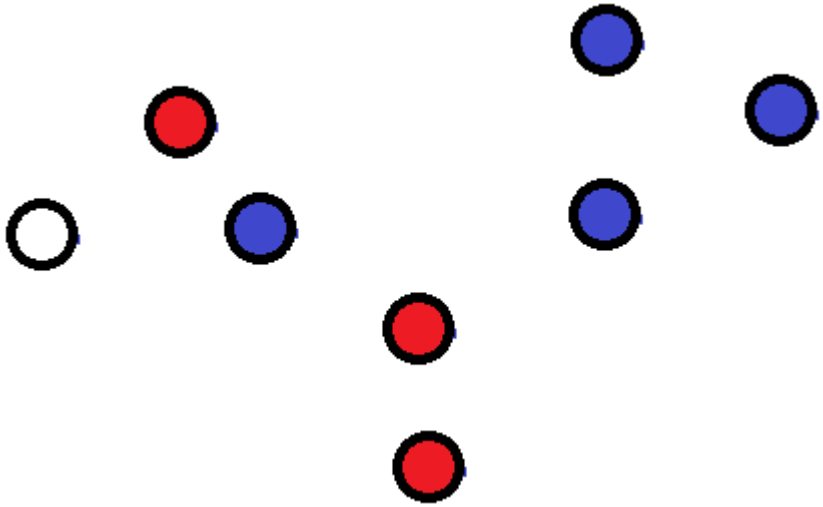
Аналогично во многих задачах с сигналами…



**Признак – не только коэффициенты в приближении,
но и ошибка приближения!**

~ отклонение от типичного поведения

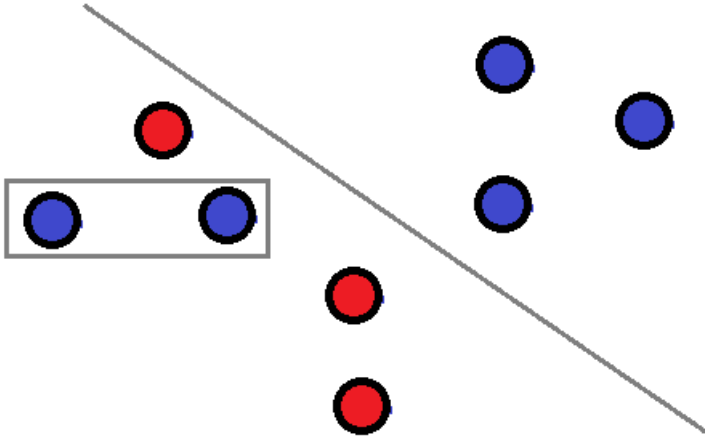
Задача классификации



Обычная точность – Mean Consequential Error

$$MCE = \frac{1}{q}\sum_{i=1}^{q} I[a_i = y_i]$$

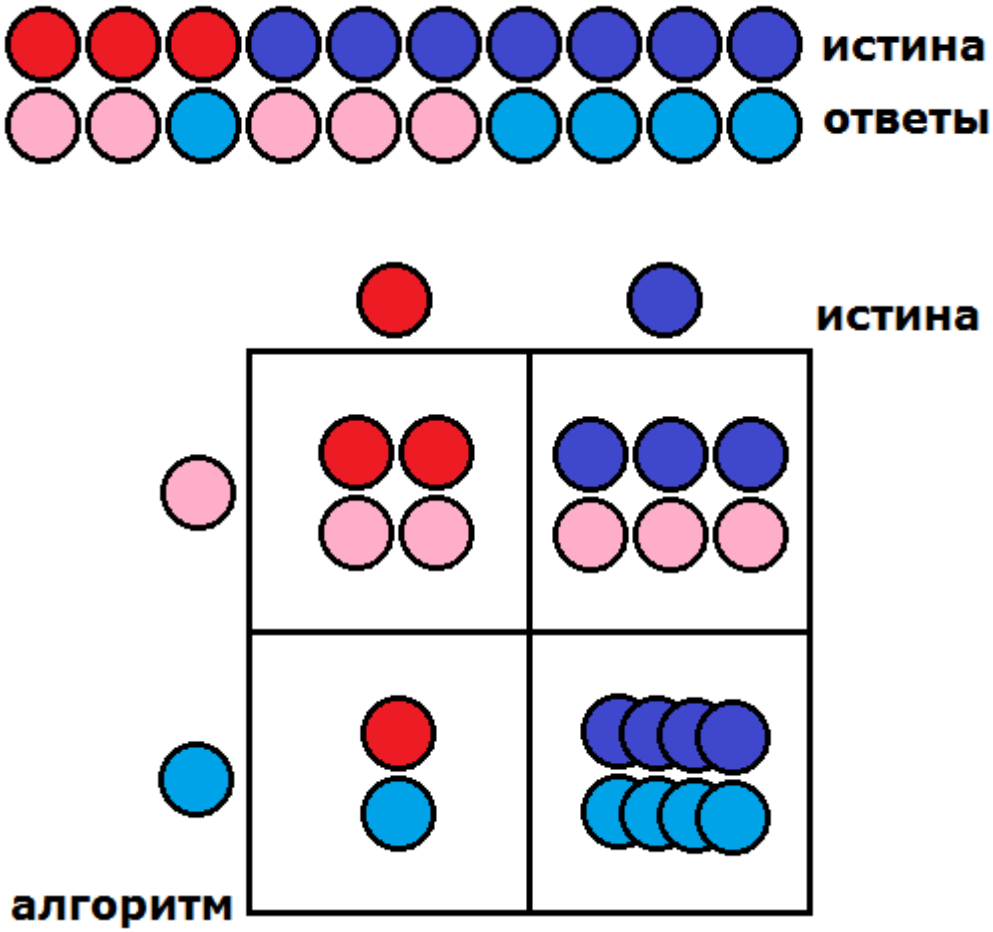
- **первое, что приходит в голову**
- **не учитывает разную мощность классов**



```
y = [0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0]
```

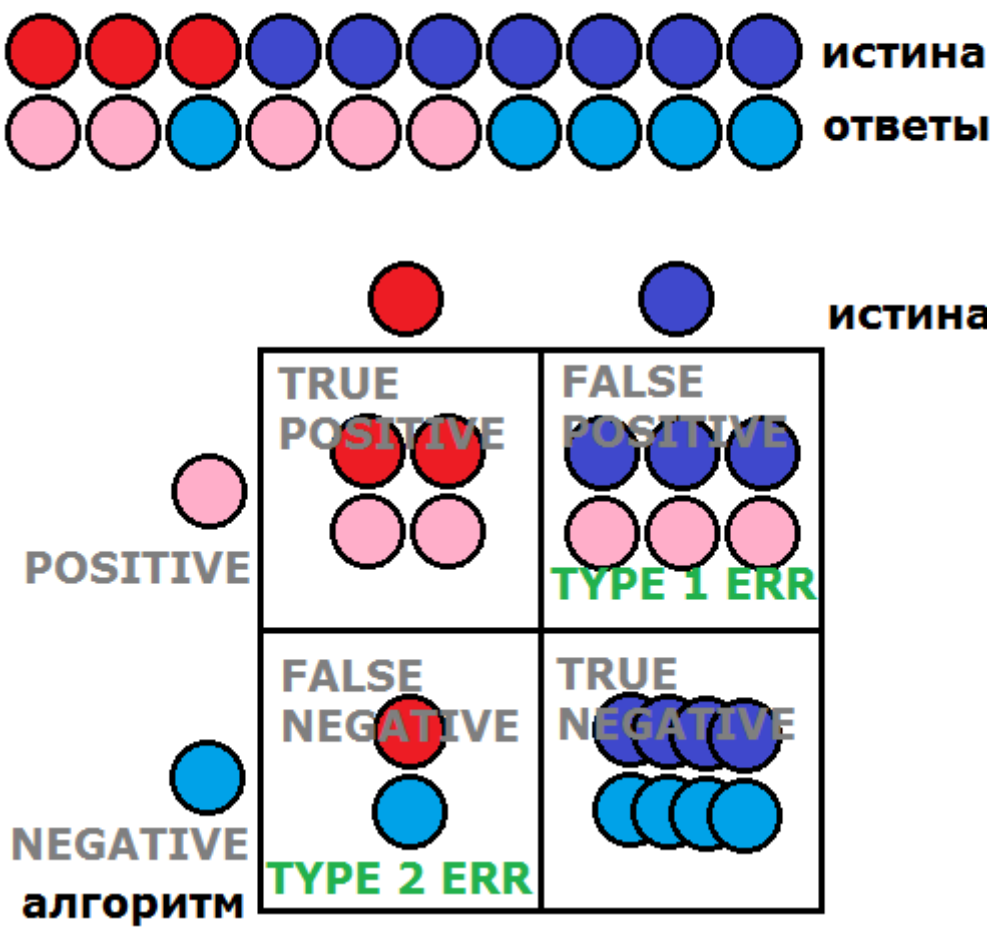
Выгодно выдавать решение – константу 0!

Задача классификации с двумя классами



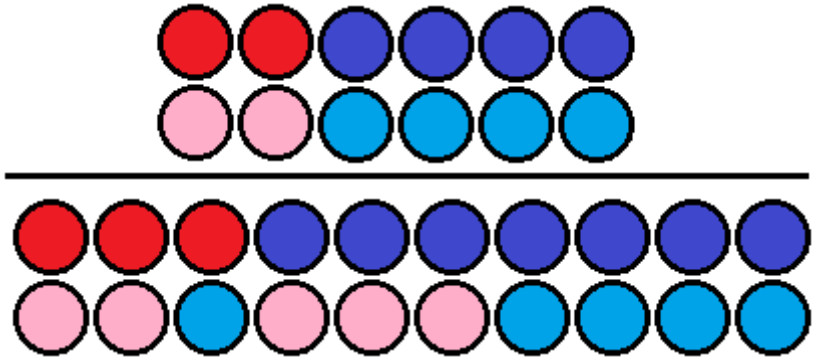
Confusion Matrix

Задача классификации с двумя классами



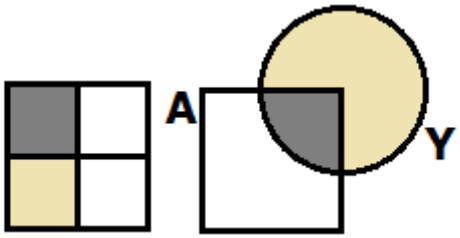
Точность Accuracy

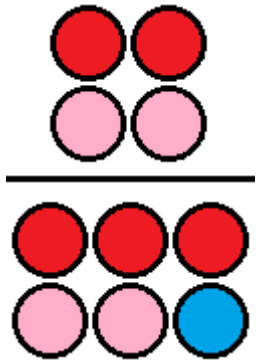
$$ACC = \frac{TP+TN}{ALL}$$



Полнота (Sensitivity, True Positive Rate, Recall, Hit Rate)

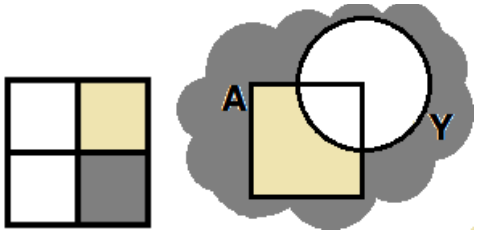
$$TPR = \frac{TP}{TP+FN}$$

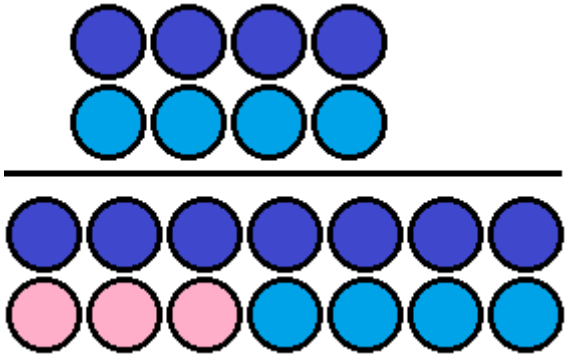




Specificity (True Negative Rate)

$$SPC = \frac{TN}{FP+TN}$$

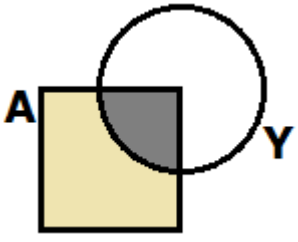


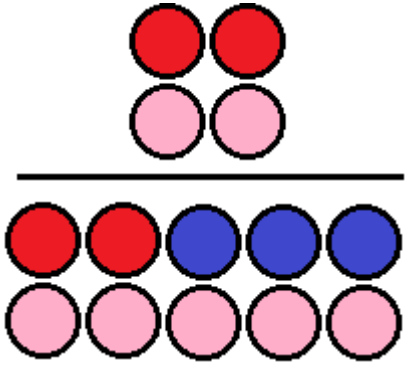


FPR = 1 – Specificity

Точность (Precision, Positive Predictive Value)

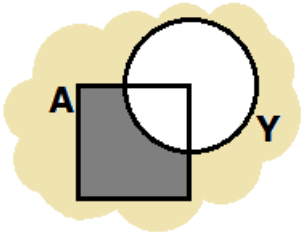
$$PPV = \frac{TP}{TP + FP}$$

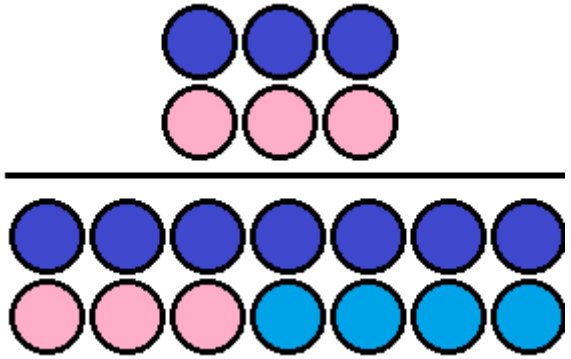




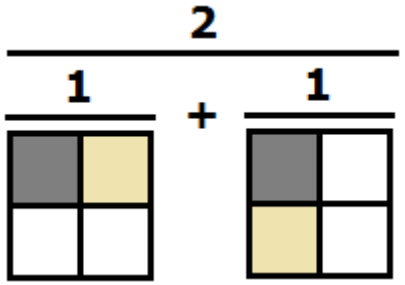
False Positive Rate (FPR, fall-out, false alarm rate)

$$FPR = \frac{FP}{FP + TN}$$





F_1 score



$$\frac{2}{\frac{1}{TP/(TP+FP)}+\frac{1}{TP/(TP+FN)}}=$$

$$=\frac{2TP}{2TP+FP+FN}$$

F_β score

$$\frac{1}{\frac{\alpha}{P}+\frac{1-\alpha}{R}}=\frac{PR}{\alpha R+(1-\alpha)P}=$$

$$=\frac{1}{\alpha}\frac{PR}{R+\left(\frac{1}{\alpha}-1\right)P}$$

$$\beta^2=\left(\frac{1}{\alpha}-1\right)$$

$$F_\beta=(1+\beta^2)\frac{PR}{R+\beta^2 P}$$

AUC ROC

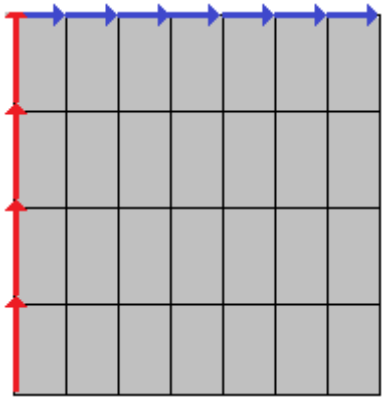
Функционал зависит не от конкретных значений, а от их порядка

ответы	истина
0.13	1
0.10	0
0	0
0.22	0
0.45	1
0.9	0
0.5	0
0.55	0
0.77	1
0.6	0
0.92	1

ответы	истина
0	0
0.10	0
013	1
0.22	0
0.45	1
0.5	0
0.55	0
0.6	0
0.77	1
0.9	0
0.92	1

отсортировать

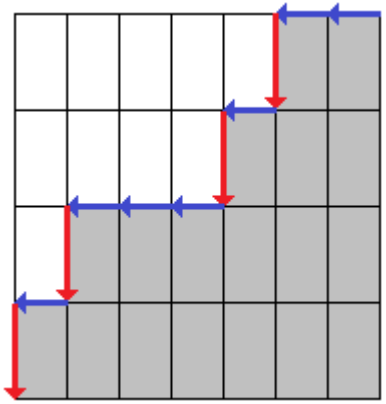
путешествие черепашки



0 0 0 0 0 0 0 1 1 1 1

Если справа налево идти

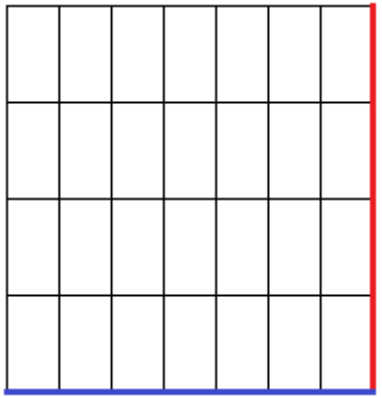
AUC = 28/28 = 1



0 0 1 0 1 0 0 0 1 0 1

Если слева направо идти

AUC = 18/28



1 1 1 1 0 0 0 0

AUC = 0/28 = 0

Смысл AUC

AUC ~ число правильно отсортированных пар (на рис. «кирпичики»)

Это сложно объяснить заказчику!

$$AUC = \frac{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j] I[a_i < a_j]}{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j]}$$

Чем хороша эта запись?

Смысл AUC

AUC ~ число правильно отсортированных пар (на рис. «кирпичики»)

Это сложно объяснить заказчику!

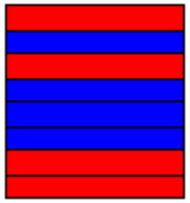
$$AUC = \frac{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j] I[a_i < a_j]}{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j]}$$

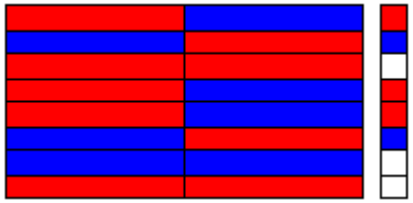
Чем хороша эта запись?

Можно обобщить, например, на регрессию.

Настройка RF/GBM на AUC ROC

Случай из жизни (Интернет-математика)



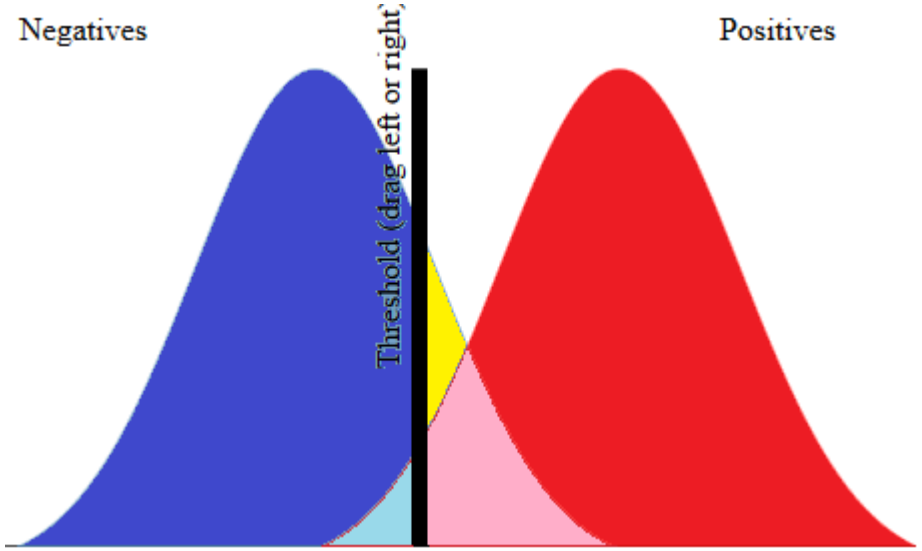


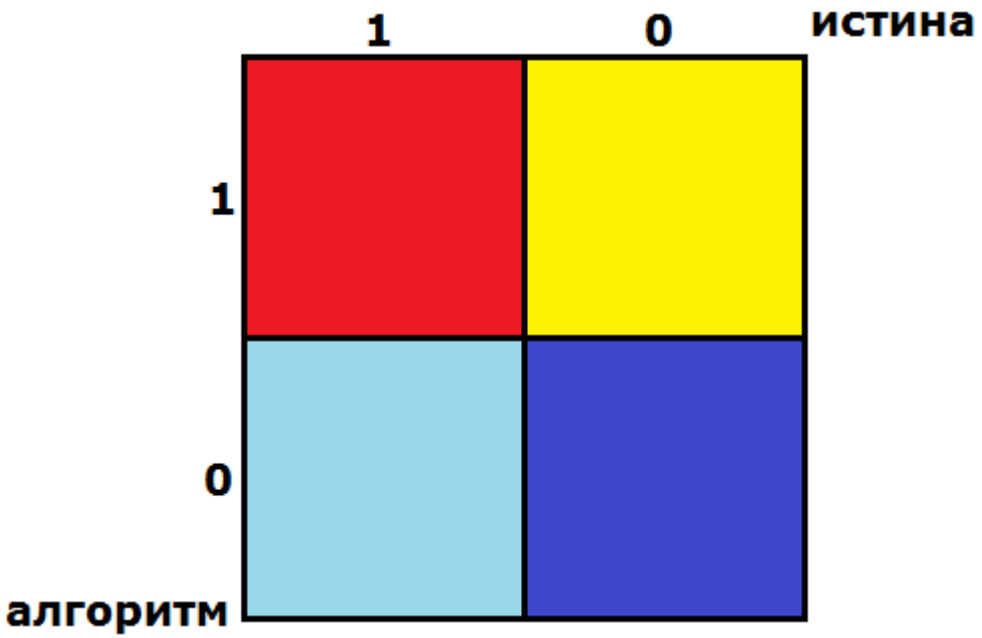
классификация → классификация пар

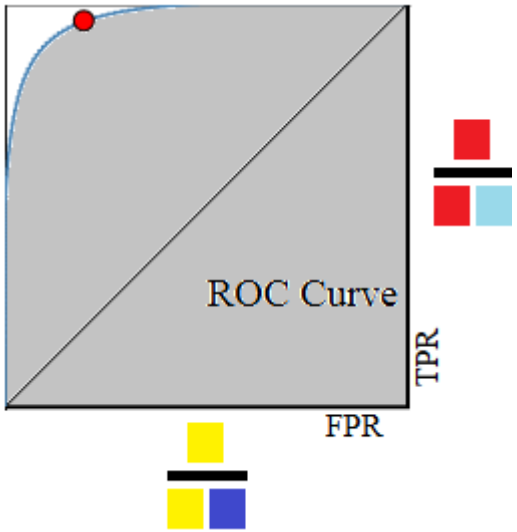
Можно дублировать,

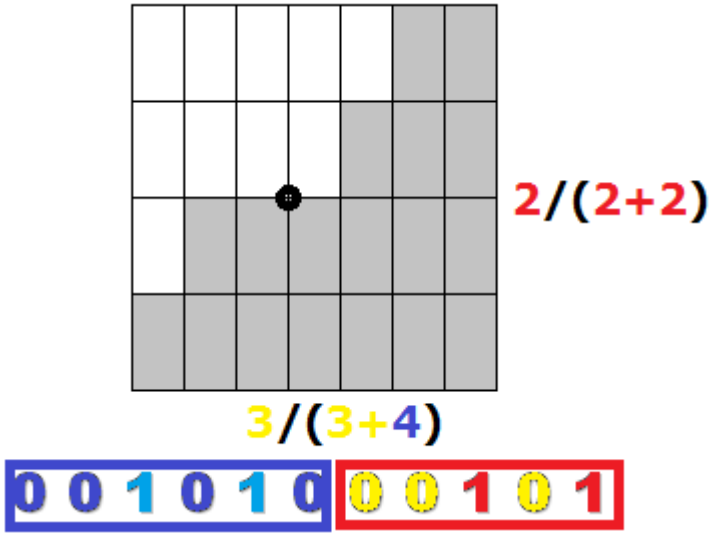
Можно брать разности/отношения.

AUC ROC







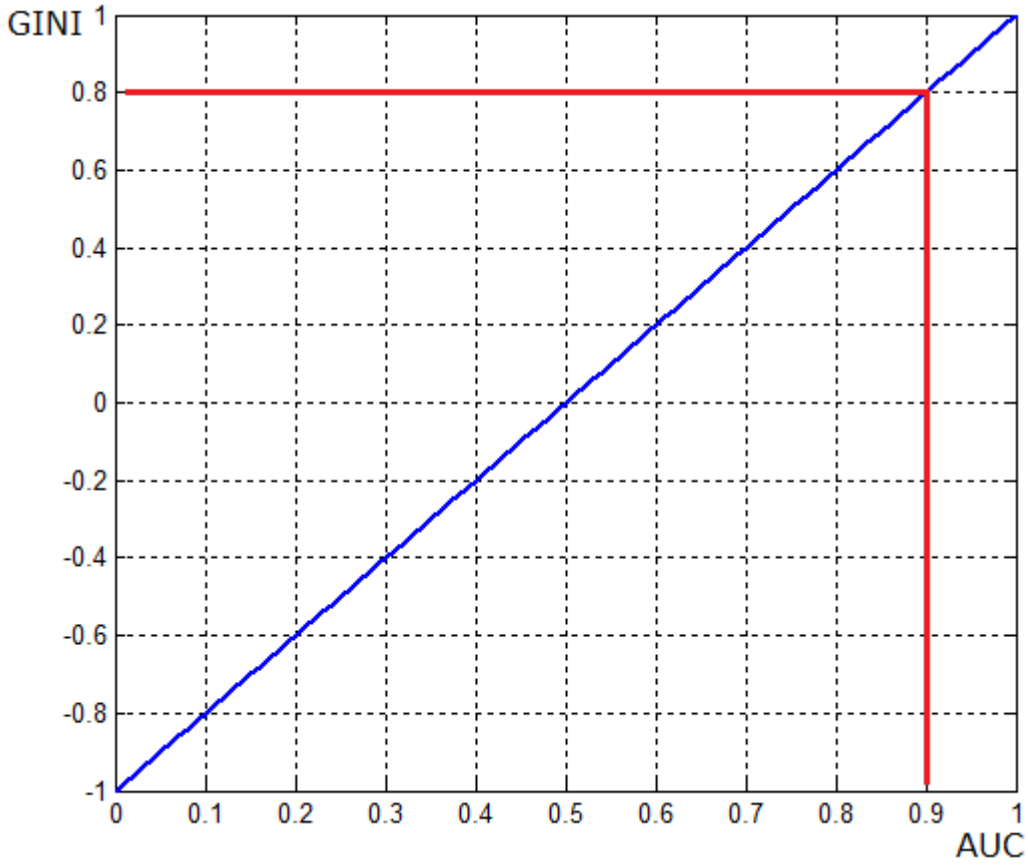


AUC – не всегда ступеньки!

GINI

В простейшем случае

$$GINI = 2 \cdot AUC - 1$$



Немного сбивает с толку разница масштабов 0.9 → 0.8

AUC ROC

+ в задачах, где важен порядок

+ учитывает разную мощность классов

+ не важны значения, важен порядок

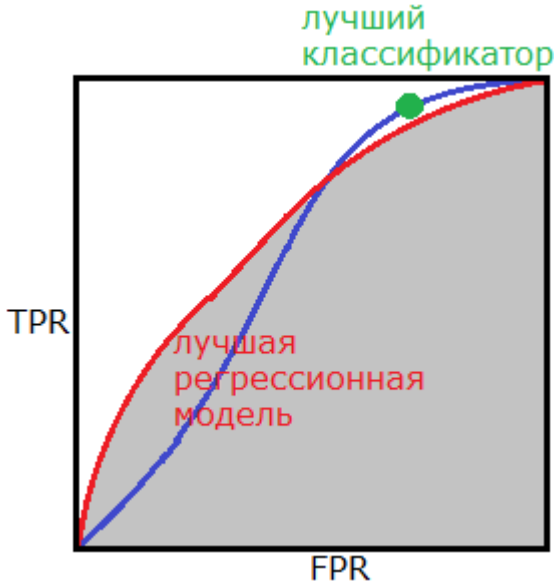
+ можно использовать для оценки признаков

- «завышает» качество

- оценивает не конкретный классификатор, а регрессию

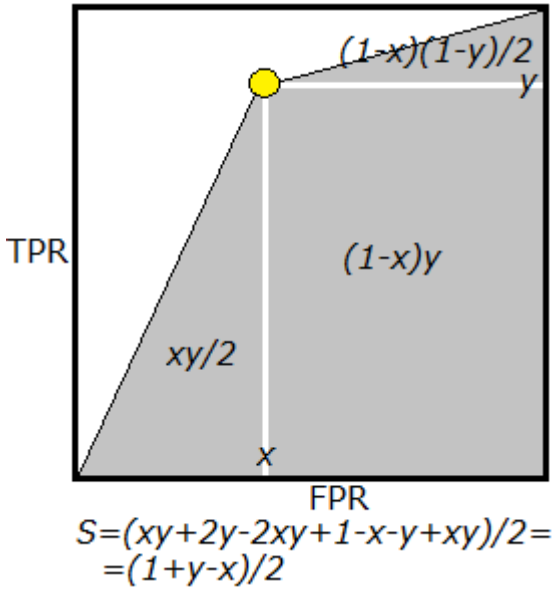
- сложно объяснить заказчику

- не путать классификацию и регрессию



Полностью бинарный случай

Если «вдруг» $a \in \{0,1\}^q$, $y \in \{0,1\}^q$, $F = AUC$**?! (Сбербанк)**



$S=(xy+2y-2xy+1-x-y+xy)/2=$
$=(1+y-x)/2$

Из рисунка

$$AUC = (1 + TPR - FPR)/2 =$$
$$= \frac{1}{2}\left[1 + \frac{TP}{TP+FN} - \frac{FP}{FP+TN}\right]$$

	1	0
1	TP	FP
0	FN	TN

A \ Y

$$= \frac{1}{2}\left[1 + \frac{\|a \cdot y\|}{\|y\|} - \frac{\|a \cdot \bar{y}\|}{\|\bar{y}\|}\right] =$$
$$= \frac{1}{2}\left[\frac{\|a \cdot y\|}{\|y\|} + \frac{\|\bar{a} \cdot \bar{y}\|}{\|\bar{y}\|}\right]$$

т.е. это точность с оглядкой на мощности классов…

$$TPR - FPR \to \max$$

Полностью бинарный случай

$$\frac{1}{2}\left[\frac{\|a \cdot y\|}{\|y\|} + \frac{\|\bar{a} \cdot \bar{y}\|}{\|\bar{y}\|}\right]$$

– примитивная точность (Accuracy, не Precision), если мощности классов совпадают.

А если выровнять мощности (как?), то можно смотреть на точность…

Log Loss

В задаче классификации с двумя непересекающимися классами (0, 1), когда ответ вероятность (?) принадлежности к классу 1

$$LOGLOSS = -\frac{1}{q}\sum_{i=1}^{q}(y_i \log a_i + (1-y_i)\log(1-a_i))$$

На что похоже?

Log Loss

В задаче классификации с двумя непересекающимися классами (0, 1), когда ответ вероятность (?) принадлежности к классу 1

$$LOGLOSS = -\frac{1}{q}\sum_{i=1}^{q}(y_i \log a_i + (1-y_i)\log(1-a_i))$$

На что похоже?

Вспоминаем…

$$\Pi = \prod_{i=1}^{n}\pi_p(x_i) = p^m(1-p)^{n-m} \sim$$

$$\frac{1}{n}\left(m\log p + (n-m)\log(1-p)\right)$$

Log Loss

Так понятнее…

$$-\begin{cases} \log a_i, & y_i = 1, \\ \log(1-a_i), & y_i = 0. \end{cases}$$

Нельзя ошибаться!

Изменяется от 0 до бесконечности

($\log 0.5$**, на самом деле нет – см. дальше)**

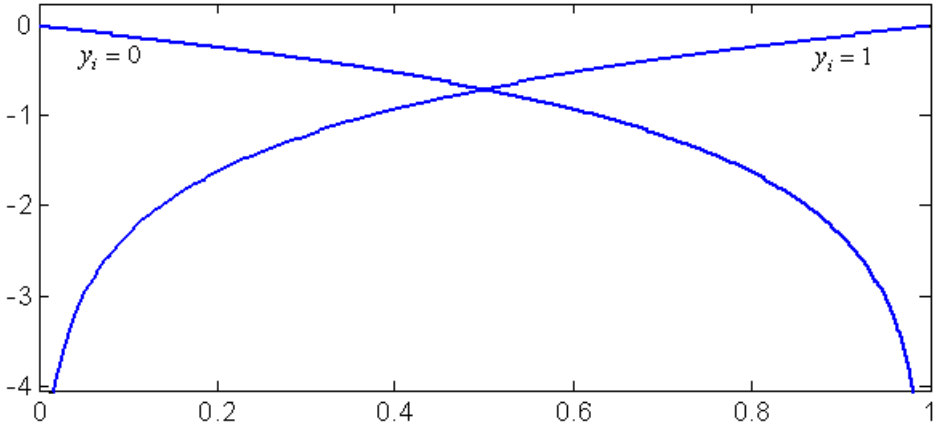


График функций $\log(a_i)$ **и** $\log(1-a_i)$ **от** a_i**.**

Посчитаем матожидание ошибки –

у нас один (i-й) объект, который с вероятностью p принадлежит классу 1.

$$-p\log(a_i)-(1-p)\log(1-a_i)$$

Минимизируем это выражение:

$$\frac{p}{a_i}-\frac{1-p}{1-a_i}=0$$

$$a_i = p$$

О чудо!

Но так не всегда…

Задача классификации {0,1} с ответами на [0,1]

Реальный случай

Пусть ошибка:

$$|y_i - a_i| \cdot \begin{cases} 0.8, & y_i = 1, \\ 0.2, & y_i = 0, \end{cases} \quad (*)$$

где $y_i \in \{0,1\}$ **– верная классификация** i**-го объекта,**

$a_i \in [0,1]$ **– ответ нашего алгоритма.**

Заказчик: важно получать значения из отрезка [0,1]
и интерпретировать как вероятности принадлежности к классу 1

Вычисление матожидания ошибки

Пусть i-й объект принадлежит к классу 1 с вероятностью p

Посчитаем матожидание нашей ошибки:

$$0.8\,|1-a_i\,|\,p+0.2\,|\,a_i\,|(1-p)=$$
$$=0.8p-0.8pa_i+0.2a_i-0.2pa_i=$$
$$=0.8p-(p-0.2)a_i$$

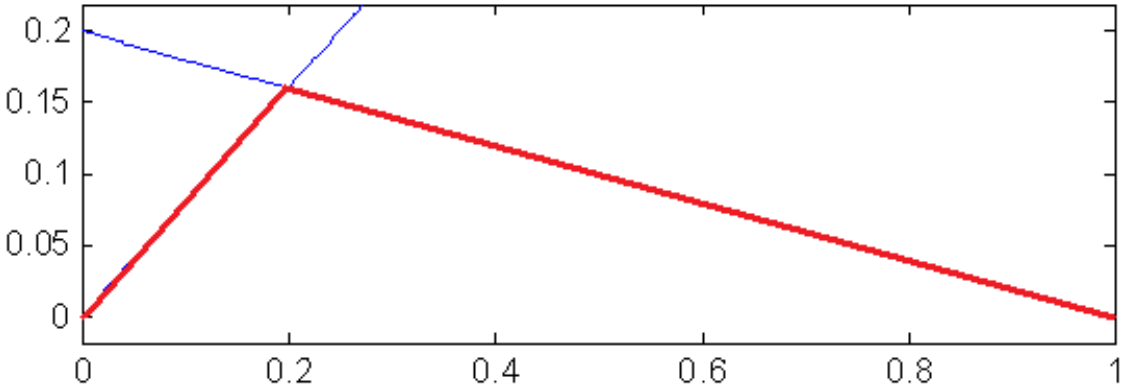
Оптимальное решение

(которое минимизирует матожидание ошибки)

$$a_i=\begin{cases}0, & p<0.2,\\ 1, & p\geq 0.2.\end{cases}$$

Неправильный выбор функционала

Функционал (*) вынуждает нас выдавать значения из множества {0,1}.



Матожидание ошибки в зависимости от p.

В чём ошибка заказчика, как исправить?

Совет

Ищите матожидание!

Пробуйте константные решения.

Многоклассовая задача

Hamming Loss

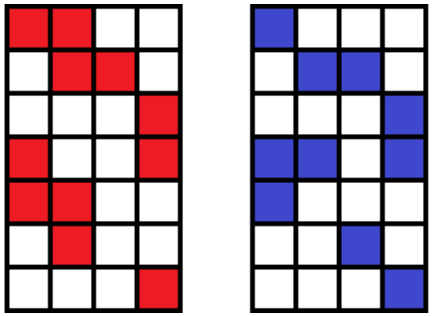
Число ошибок в векторе классификаций

$$HL(\tilde{a}, \tilde{y}) = \frac{\| \tilde{a} \oplus \tilde{y} \|}{l}$$

Log Loss

$$LOGLOSS = -\frac{1}{q} \sum_{i=1}^{q} \sum_{j=1}^{l} y_j \log a_j$$

Полнота и т.п. – всё что придумывается со строками матрицы



Полнота и т.п. – по строкам или столбцам

Это множества – и можно усреднять функции сходства множеств

Как использовать на практике (LSHTC)

- **Решающее правило с отсечкой:** $\alpha_{ij} = \begin{cases} 1, & \gamma_{ij} > \min(c, \max(\{\gamma_{ij}\}_{j=1}^{l}), \\ 0, & \text{иначе.} \end{cases}$
- **Решать задачу по вертикали / по горизонтали**

Функционал в LSHTC

$$\tilde{F} = \frac{2\tilde{P}\tilde{R}}{\tilde{P} + \tilde{R}}$$

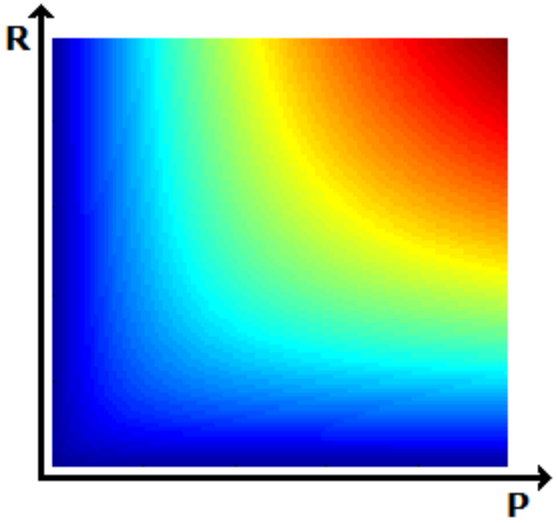
```
Id,Predicted
1,12 35 200
2,54 55
3,11
4,1 7 101
...
```

$$\tilde{P} = \frac{1}{l}\sum_{j=1}^{l}\frac{TP_j}{TP_j + FP_j}$$

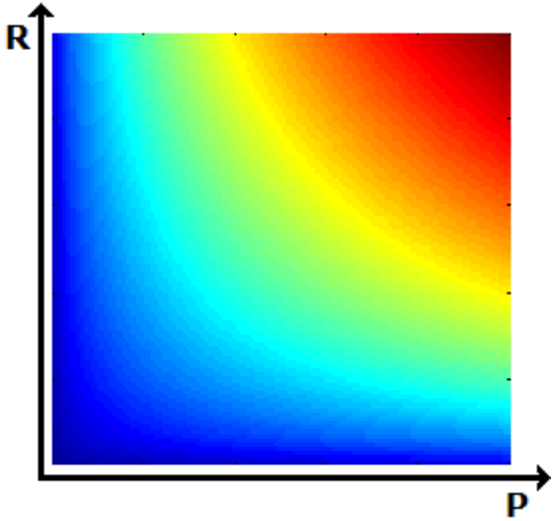
$$\tilde{R} = \frac{1}{l}\sum_{j=1}^{l}\frac{TP_j}{TP_j + FN_j}$$

Почему используется F-мера

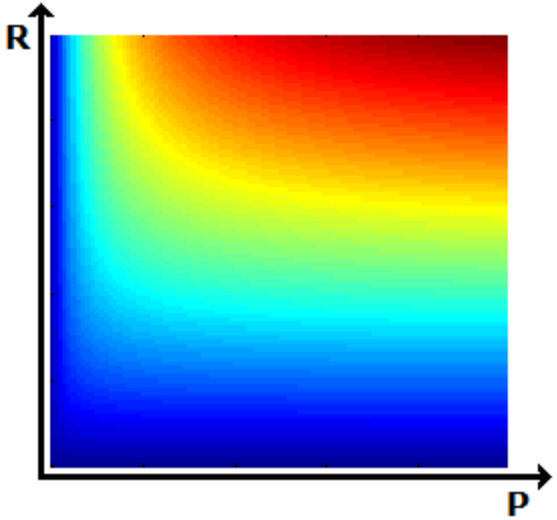
$2/(1/P+1/R)$



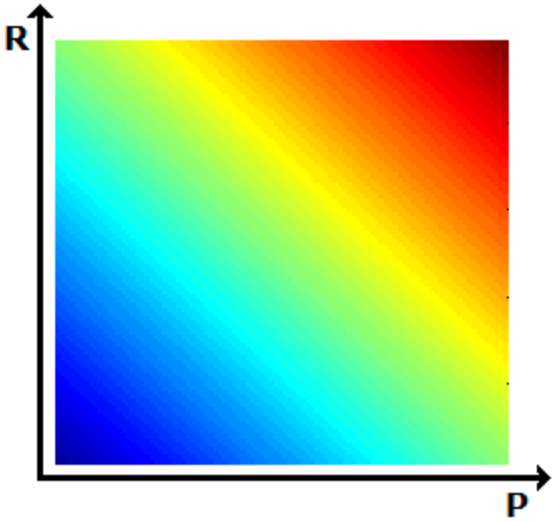
$\sqrt{P \cdot R}$



$2/(0.9/P+0.1/R)$



$(P+R)/2$



Очень полезно «чувствовать функции»

Пример из жизни: лайки

L	D	
+100	0	Совсем хорошо
+10	0	Хорошо
+1	–0	Мало статистики, но нет минусов
+2	–1	Есть минусы
+10	–9	Много минусов
+100	–100	Неоднозначно
+1	–1	Мало статистики
+9	–10	Много плюсов
+1	–2	Мало плюсов
0	–1	Нет плюсов

Как придумать один признак на базе двух?

Очень полезно «чувствовать функции»

Пример из жизни: лайки

L	D	$(\lvert L\rvert-\lvert D\rvert)/\sqrt{\lvert L\rvert+\lvert D\rvert}$
+100	0	10.0000
+10	0	3.1623
+1	–0	1.0000
+2	–1	0.5774
+10	–9	0.2294
+100	–100	0
+1	–1	0
+9	–10	-0.2294
+1	–2	-0.5774
0	–1	-1.0000

Оценка результатов поиска/рекомендаций

Average Precision at n

Для оценки адекватности множества рекомендаций

$$ap@n = \sum_{k=1}^{n} \frac{P(k)}{\min(n,m)}$$

$$P(k) = \begin{cases} y_1 + \ldots + y_k, & y_k = 1, \\ 0, & y_k = 0, \end{cases}$$

y_i **– бинарное значение релевантности**

рекомендации (алгоритма):

– правильные

$$ap@10 = \frac{1}{6}\left[\frac{1}{1} + \frac{2}{3} + \frac{3}{6}\right]$$

Mean Average Precision

– усреднение ap@n по всем пользователям

Классические функционалы в поиске

Выдали id документов/товаров/..., а их ценность (релевантность):

$$y_1,\ldots,y_q$$

Cumulative Gain

$$y_1+\ldots+y_q$$

Discounted Cumulative Gain

$$DCG = y_1 + \sum_{i=2}^{q} \frac{y_i}{\log_2(i)} = y_1 + y_2 + \frac{y_3}{\log_2 3} + \ldots + \frac{y_q}{\log_2 q}$$

т.е. выгодно сначала выдавать более ценные

Ещё вариант:

$$\sum_{i=1}^{q} \frac{2^{y_i}-1}{\log_2(i+1)}$$

Normalized DCG

$$nDCG = \frac{DCG}{IDCG}$$

IDCG = ideal DCG

для того, чтобы не было зависимости от длины выдачи

Что ещё может встретиться…

$$\frac{1}{|Z|}\sum_{z\in Z}\frac{|\{r_1,\ldots,r_{\min(S,R,z)}\}\cap\{s_1,\ldots,s_{\min(S,R,z)}\}|}{\min(S,R,z)}$$

$r_1,\ldots,r_R$ – **рекомендации**

$s_1,\ldots,s_S$ – **правильные ответы**

$Z=\{5,10,15,20,25,30\}$

Quadratic Weighted Kappa

показывает согласованность порядков,
когда ответы "мера релевантности"

y = 1 1 1 2 2 3 3 3 # правильный ответ
a = 1 1 2 1 3 2 3 3 # наш ответ
	0.6666667

a = 1 1 1 2 2 3 3 3 # наш ответ	1
a = 3 3 3 2 2 1 1 1 # наш ответ	-1

Quadratic Weighted Kappa

```
  E = table(y) %*% t(table(a))
  O = table(y,a)
  E = E/sum(E)*sum(O)
  n = length(unique(y))
  W = (matrix(1:n,nr=n,nc=n) -
matrix(1:n,nr=n,nc=n,byrow = TRUE))**2/(n-1)**2
  kappa = 1-sum(W*O)/sum(W*E)
```

```
a = 1 1 2 1 3 2 3 3
y = 1 1 1 2 2 3 3 3

O =
   a
y   1 2 3
  1 2 1 0
  2 1 0 1
  3 0 1 2
W =
     [,1] [,2] [,3]
[1,] 0.00 0.25 1.00
[2,] 0.25 0.00 0.25
[3,] 1.00 0.25 0.00
```

```
E =
   a
y   1 2 3
  1 9 6 9
  2 6 4 6
  3 9 6 9
E = нормализованная
   a
y        1    2     3
  1 1.125 0.75 1.125
  2 0.750 0.50 0.750
  3 1.125 0.75 1.125

Kappa =
0.6666667
```

Редакторское расстояние

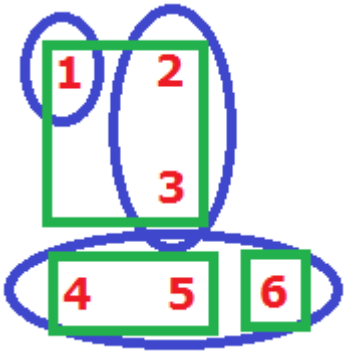
Операции

добавление к кластеру

создание кластера с одним объектом

удаление из кластера

удаление кластера с одним объектом




```
1 2 3;4 5;6
1 2 3; 4 5   [delC]
2 3; 4 5   [del]
2 3; 4 5; 1   [insC]
2 3; 4 5 6; 1   [ins]
```

	2 3	4 5 6	1
1 2 3	1	6	2
4 5	4	1	3
6	3	2	2

Редакторское расстояние

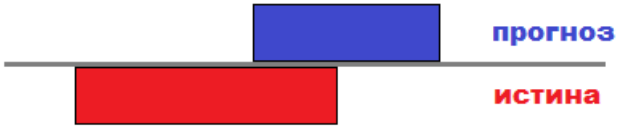
- **Плохо заносить не в тот кластер (целых две операции на перенос)**
- **Плохо создавать неправильный кластер**

⇒ осторожный алгоритм

- **Многое зависит от операций…**

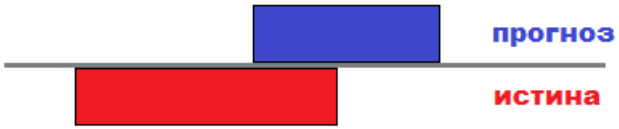
Задача с «неклассическим целевым вектором»

Надо предсказывать не значение,
а интервал [a, b]



Как измерить качество?

Задача с интервальным целевым вектором



Интервал – это множество!

Коэффициент Жаккара (Jaccard)

$$\frac{|A \cap B|}{|A \cup B|}$$

коэффициент Шимкевича-Симпсона (Szymkiewicz, Simpson)

$$\frac{|A \cap B|}{\min(|A|,|B|)}$$

коэффициент Браун-Бланке (Braun-Blanquet)

$$\frac{|A \cap B|}{\max(|A|,|B|)}$$

См. википедию «Коэффициент сходства» для переноса идеи Колмогорова об обобщённом среднем…

Вариации на тему усреднения…

коэффициент Сёренсена (Sörensen)

$$\frac{2|A\cap B|}{|A|+|B|}$$

коэффициент Кульчинского (Kulczinsky)

$$\frac{|A\cap B|}{2}\,\frac{1}{1/|A|+1/|B|}$$

коэффициент Отиаи (Ochiai)

$$\frac{|A\cap B|}{\sqrt{|A|\cdot|B|}}$$

Меры включения

$$\frac{|A\cap B|}{|A|}$$

$$\frac{|A\cap B|}{2|A|-|A\cap B|}$$

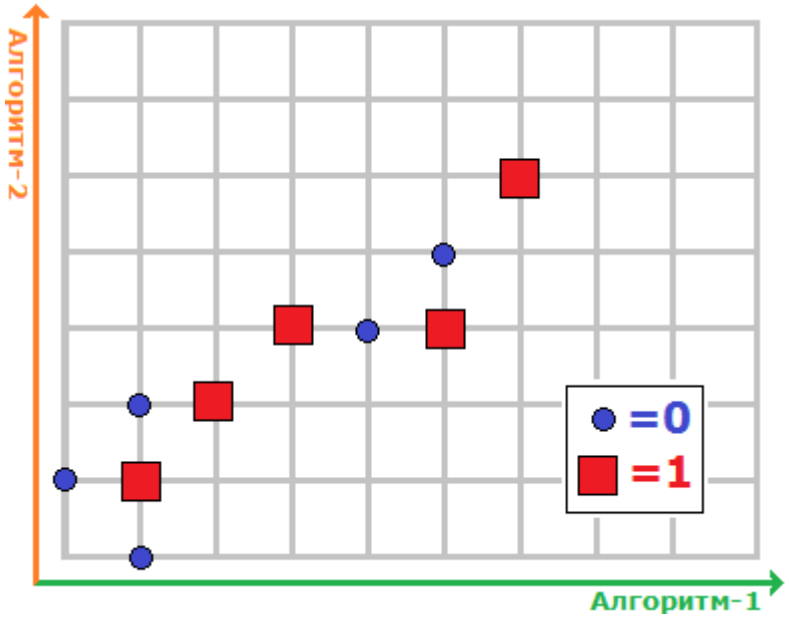
$$\frac{|A\cap B|}{|B|}$$

$$\frac{|A\cap B|}{2|B|-|A\cap B|}$$

Как решать задачи с интервалами? Потом вернёмся…

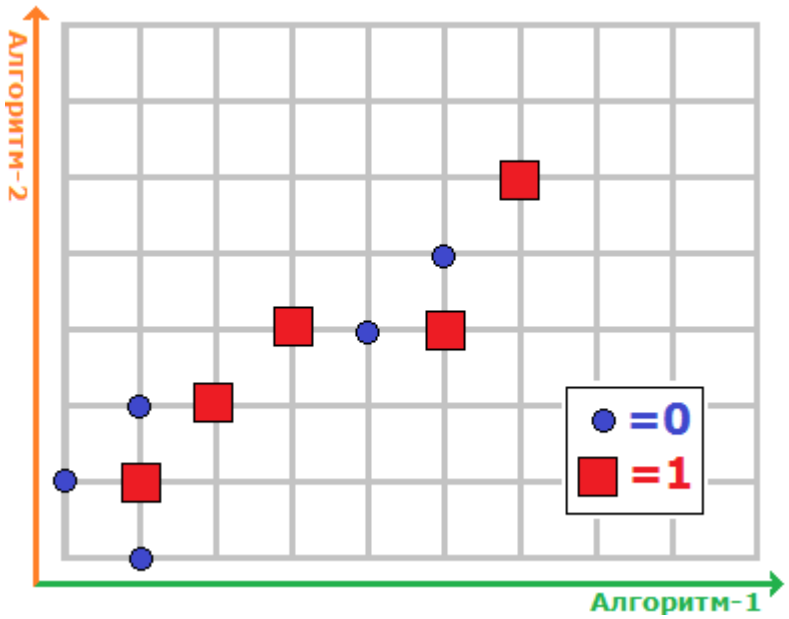
Упражнение №1.

Рассматривается задача классификации на два класса. На рисунке показаны объекты в пространстве ответов двух алгоритмов. Вычислить AUC ROC для алгоритмов.



Упражнение №1 - Решение

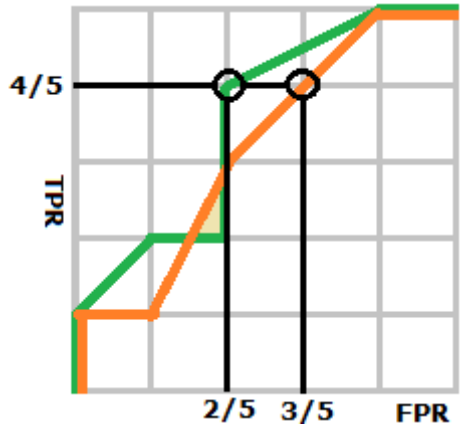
1. Смотрим проекции на оси – ответы алгоритмов

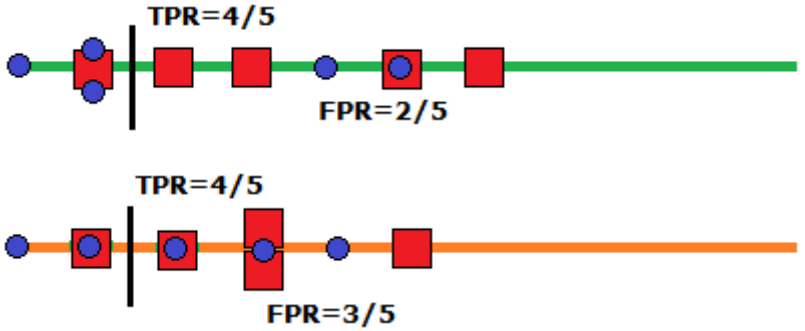


Первый алгоритм:

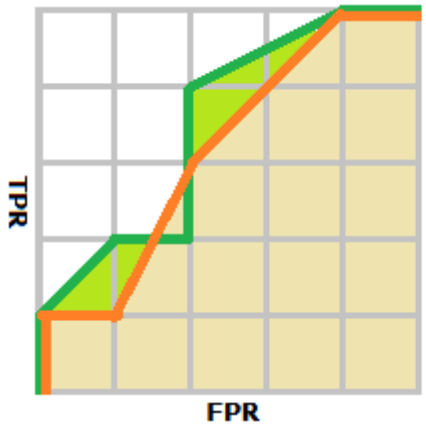
Второй алгоритм:

2. По проекциям строим ROC - кривые:





3. Вычисляем площади под ROC - кривыми:



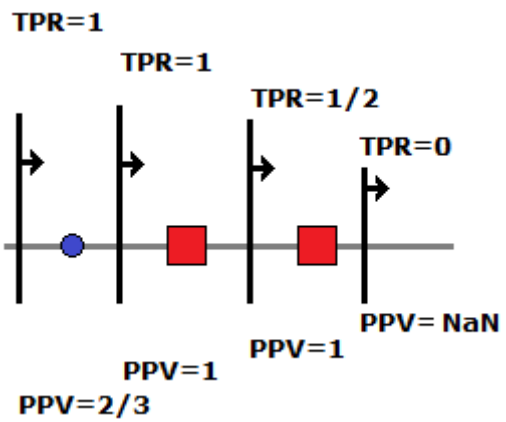
16/25
17.5/25

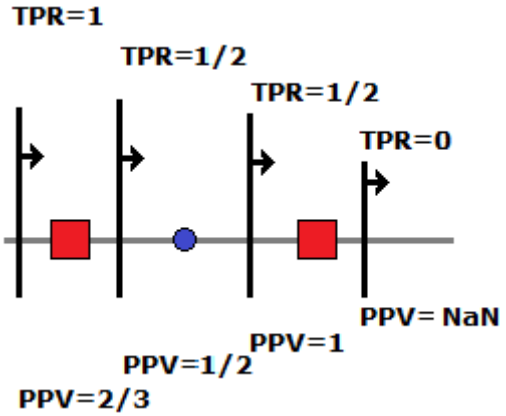
Упражнение №2.

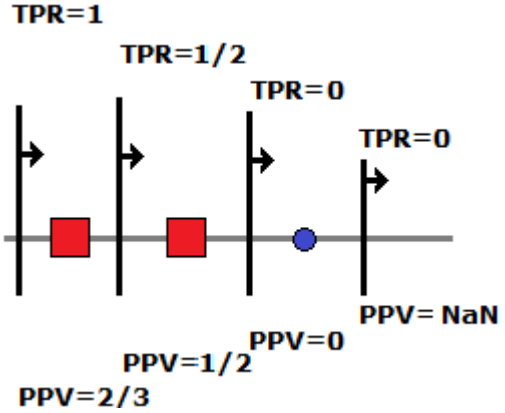
Какие значения F_1-меры могут быть у классификатора в задаче с двумя непересекающимися классами и тремя объектами?

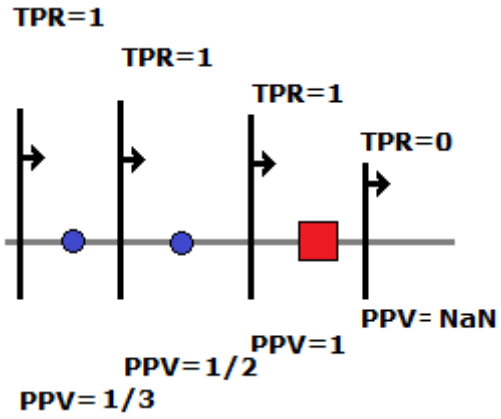
Упражнение №2 – Решение.

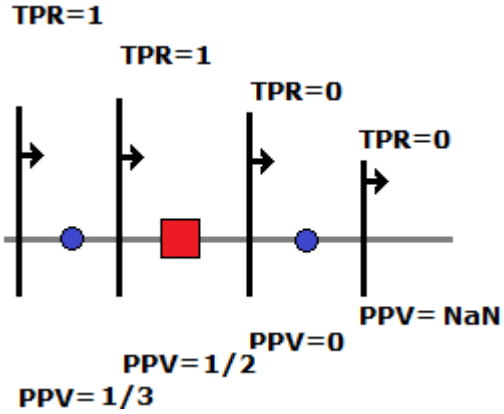
Можно честно рассмотреть все возможные случаи:

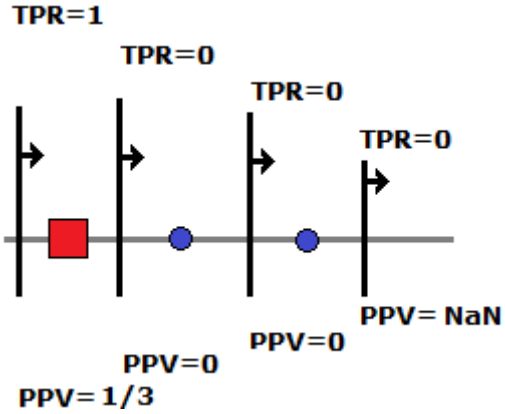












Упражнение №2 - Решение.

Получаем, что F1-мера – среднее гармоническое чисел из пар
(1, 1), (1/2, 1), (2/3, 1), (1/3, 1), (1/2, 1/2), (0, 0)

Все возможные значения F1-меры:
1, 0.8, 2/3, 0.5, 0

Но можно быстрее догадаться до ответа…

Литература

Tom Fawcett An introduction to ROC analysis // Pattern Recognition Letters Volume 27 Issue 8, 2006, P. 861-874.
https://ccrma.stanford.edu/workshops/mir2009/references/ROCintro.pdf

Стрижов В.В. Функция ошибки в задачах восстановления регрессии // Заводская лаборатория, 2013, 79(5): 65-73.
http://strijov.com/papers/Strijov2012ErrorFn.pdf

К.Д. Маннинг, П. Рагхаван, Х. Шютце «Введение в информационный поиск» // . — Вильямс, 2011.