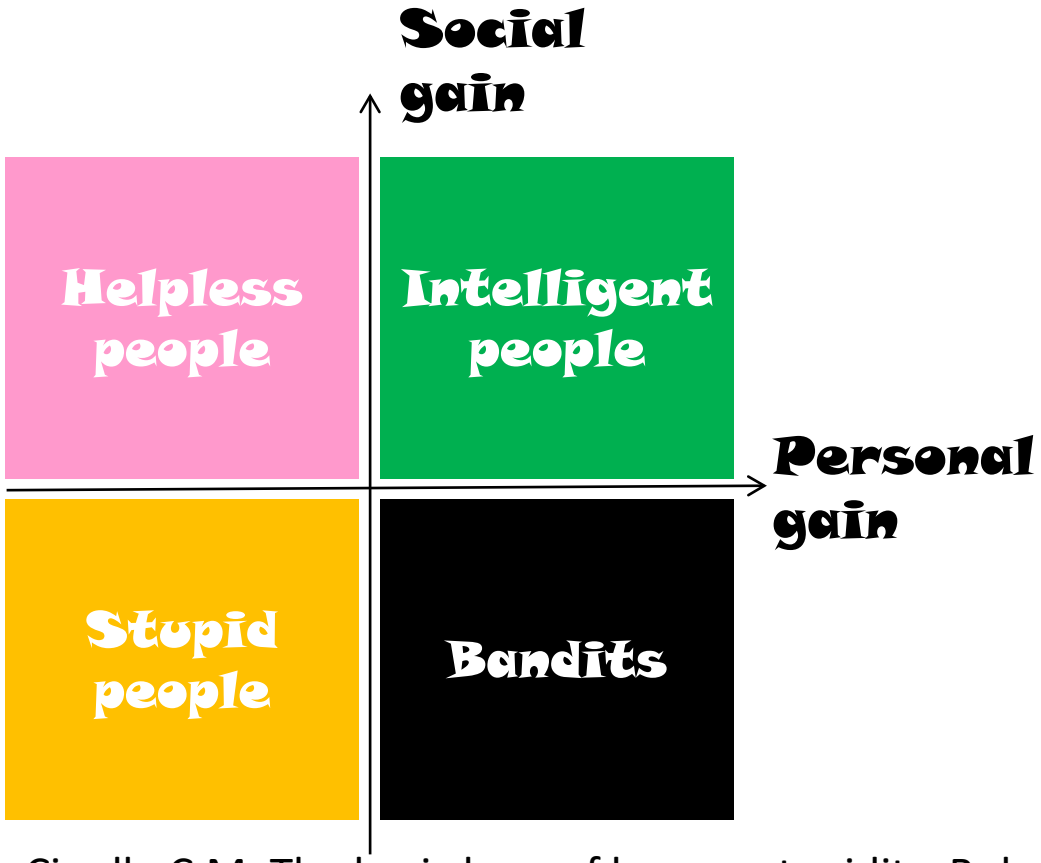
Задача классификации данных

Классификация – нить Ариадны в лабиринте природы.

Жорж Санд

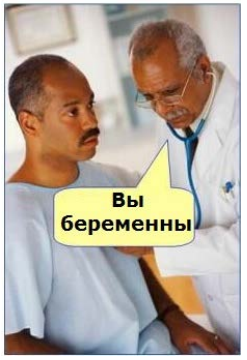


Cipolla C.M. The basic laws of human stupidity. Bologna: il Muli, 2011

Содержание

- Основные понятия
- Деревья решений
- Байесовская классификация
- Классификация по ближайшим соседям
- **Оценка качества классификации**
  - **Меры качества**
  - **Подготовка тестовой выборки**
- Ансамблевая классификация

Матрица ошибок (Confusion matrix)

Прогноз \ Класс	C (беременность)	$\neg C$ (нет беременности)
C (беременность)	***True Positives*** 	***False Positives*** 
$\neg C$ (нет беременности)	***False Negatives*** 	***True Negatives*** 

Матрица ошибок (Confusion matrix)

Прогноз \ Класс	$\boldsymbol{C}$	$\boldsymbol{\neg C}$
$\boldsymbol{C}$	$True\ Positives\ (\boldsymbol{TP})$	$False\ Positives\ (\boldsymbol{FP})$
$\boldsymbol{\neg C}$	$False\ Negatives\ (\boldsymbol{FN})$	$True\ Negatives\ (\boldsymbol{TN})$

- TP: верно распознанные объекты класса C
- TN: верно распознанные объекты класса $\neg C$
- FN: объекты класса C, неверно распознанные как объекты класса $\neg C$
- FP: объекты класса $\neg C$, неверно распознанные как объекты класса C
- $P = TP + FN$: объекты класса C
- $N = FP + TN$: объекты класса $\neg C$

Accuracy, recognition rate (Аккуратность)
Error/misclassification rate (Доля ошибок)

- $\boldsymbol{Accuracy = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+FN+FP+TN}}$
- $\boldsymbol{Error = 1 - Accuracy = \frac{FP+FN}{P+N} = \frac{FP+FN}{TP+FN+FP+TN}}$
- Наиболее простой способ оценки качества: доля верных/неверных ответов
- Неадекватен при дисбалансе классов ($P \ll N$ или $P \gg N$)

П\К	$\boldsymbol{C}$	$\boldsymbol{\neg C}$
$\boldsymbol{C}$	TP	FP
$\boldsymbol{\neg C}$	FN	TN

Неадекватность Accuracy и Error при дисбалансе классов в обучающей выборке

П\К	Рак = да	Рак = нет
Рак = да	$TP = 1$	$FP = 9$
Рак = нет	$FN = 2$	$TN = 988$
Всего	$P = 3$	$N = 997$

- $Accuracy = \frac{1+988}{1+2+9+988} = 0.989$, $Error = 0.011$
- Классификатор идеально распознает отсутствие рака и плохо распознает рак
- При дисбалансе классов в выборке **нужны отдельные меры качества распознавания объектов из классов C и $\neg C$**

Sensitivity, True Positive rate (Чувствительность), Specificity, True Negative rate (Специфичность)

- *Чувствительность* – доля верно распознанных объектов класса C

П\К	C	$\neg C$
C	TP	FP
$\neg C$	FN	TN

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{P}} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

- *Специфичность* – доля верно распознанных объектов класса $\neg C$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{N}} = \frac{\boldsymbol{TN}}{\boldsymbol{TN + FP}}$$

- Аккуратность выражается через чувствительность и специфичность

$$Accuracy = \frac{TP + TN}{P + N} = Sensitivity \cdot \frac{P}{P + N} + Specificity \cdot \frac{F}{P + N}$$

Пример

Прогноз\Класс	Рак = Да	Рак = Нет
Рак = Да	**90** TP	**140** FP
Рак = Нет	**210** FN	**9560** TN
Всего	300 P	9700 N

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

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

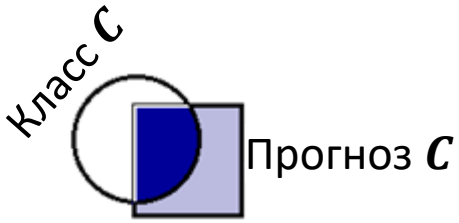
- $Accuracy = \frac{90+9560}{300+9700} = 97.7\%, Error = 1 - Accuracy = 2.3\%$
- $Sensitivity = \frac{90}{90+210} = 30\%$
- $Specificity = \frac{9560}{9560+140} = 98.6\%$

$Sensitivity$ и $Specificity$ не вполне адекватны при дисбалансе классов

Precision (точность), Recall (полнота)

- *Точность* показывает долю объектов, верно распознанных моделью как объекты класса C, от всех объектов, распознанных моделью как объекты класса C

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

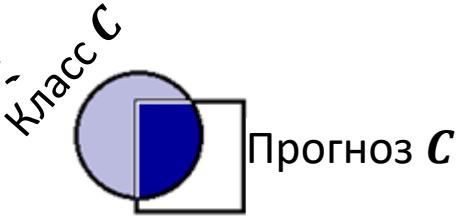


П\К	C	$\neg C$
C	TP	FP
$\neg C$	FN	TN

- *Полнота* показывает долю объектов, верно распознанных моделью как объекты класса C, от всех объектов класса C

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} =$$

$$= \boldsymbol{Sensiti}$$



П\К	C	$\neg C$
C	TP	FP
$\neg C$	FN	TN

Пример

Прогноз\Класс	Рак = Да	Рак = Нет
Рак = Да	**90** TP	**140** FP
Рак = Нет	**210** FN	**9560** TN
Всего	300 P	9700 N

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

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

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

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

- $Accuracy = \frac{90+9560}{300+9700} = 96.5\%, Error = 3.5\%$
- $Sensitivity = \frac{90}{90+210} = 30\%, Specificity = \frac{9560}{9560+140} = 98.6\%$
- $Precision = \frac{90}{90+140} = 39.13\%, Recall = Sensitivity = 30\%$

Пример

Прогноз\Класс	Рак = Да	Рак = Нет
Рак = Да	**140** TP	**90** FP
Рак = Нет	**160** FN	**9610** TN
Всего	300 P	9700 N

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

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

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

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

- $Accuracy = \frac{140+9610}{300+9700} = 97.5\%, Error = 2.5\%$
- $Sensitivity = \frac{140}{140+160} = 46.67\%, Specificity = \frac{9610}{9610+160} = 99.07\%$
- $Precision = \frac{140}{140+90} = 60.87\%, Recall = Sensitivity = 46.67\%$

Пример

Прогноз\Класс	Рак = Да	Рак = Нет
Рак = Да	**200** TP	**30** FP
Рак = Нет	**100** FN	**9670** TN
Всего	300 P	9700 N

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

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

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

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

- $Accuracy = \frac{200+9670}{300+9700} = 98.7\%, Error = 1.3\%$
- $Sensitivity = \frac{200}{200+100} = 66.7\%, Specificity = \frac{9670}{9670+200} = 99.69\%$
- $Precision = \frac{200}{200+30} = 86.9\%, Recall = Sensitivity = 66.7\%$

Мера F

- Мера F (F_1 или F-score) – среднее гармоническое точности и полноты

$$\boldsymbol{F = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}}$$

- Среднее гармоническое двух чисел ближе минимальному из них, чем среднее арифметическое и геометрическое

$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{\frac{a+b}{2}}$	$\boldsymbol{\sqrt[2]{ab}}$	$\boldsymbol{\frac{2}{\frac{1}{a}+\frac{1}{b}}}$
1	5	3	2.236	1.667

Пример

Прогноз\Класс	Рак = Да	Рак = Нет
Рак = Да	**90** TP	**140** FP
Рак = Нет	**210** FN	**9560** TN
Всего	300 P	9700 N

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

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

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

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

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- $Accuracy = \frac{90+9560}{300+9700} = 96.5\%, Error = 3.5\%$
- $Sensitivity = \frac{90}{90+210} = 30\%, Specificity = \frac{9560}{9560+140} = 98.6\%$
- $Precision = \frac{90}{90+140} = 39.13\%, Recall = \frac{90}{90+210} = 30\%$
- $F = \frac{2 \cdot 0.3913 \cdot 0.3}{0.3913+0.3} = 0.34$

Пример

Прогноз\Класс	Рак = Да	Рак = Нет
Рак = Да	**140** TP	**90** FP
Рак = Нет	**160** FN	**9610** TN
Всего	300 P	9700 N

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

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

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

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

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- $Accuracy = \frac{140+9610}{300+9700} = 97.5\%, Error = 2.5\%$
- $Sensitivity = \frac{140}{140+160} = 46.67\%, Specificity = \frac{9610}{9610+160} = 99.07\%$
- $Precision = \frac{140}{140+90} = 60.87\%, Recall = \frac{140}{140+160} = 46.67\%$
- $F = \frac{2 \cdot 0.6087 \cdot 0.4667}{0.6087+0.4667} = 0.53$

Пример

Прогноз\Класс	Рак = Да	Рак = Нет
Рак = Да	**200** TP	**30** FP
Рак = Нет	**100** FN	**9670** TN
Всего	300 P	9700 N

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

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

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

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

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- $Accuracy = \frac{200+9670}{300+9700} = 98.7\%, Error = 1.3\%$
- $Sensitivity = \frac{200}{200+100} = 66.7\%, Specificity = \frac{9670}{9670+200} = 99.69\%$
- $Precision = \frac{200}{200+30} = 86.9\%, Recall = \frac{200}{200+100} = 66.7\%$
- $F = \frac{2 \cdot 0.869 \cdot 0.667}{0.869+0.667} = 0.75$

Мера F_β

- F_β – взвешенная мера точности и полноты

$$\boldsymbol{F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 Precision + Recall}}$$

  - $F_\beta \equiv Precision$ при $\beta = 0$
  - $F_\beta \equiv Recall$ при $\beta = \infty$
  - $F_\beta \equiv F_1 \equiv F$ при $\beta = 1$

Обобщающая мера: weighted accuracy

- $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$
- $\boldsymbol{WeightedAccuracy = \frac{w_1TP+w_4TN}{w_1TP+w_2FP+w_3FN+w_4TN}}$
- Связь с остальными мерами

Мера	w_1	w_2	w_3	w_4
$Accuracy$	1	1	1	1
$Precision$	1	0	1	0
$Recall$	1	1	0	0
F_β	$1+\beta^2$	β^2	1	0

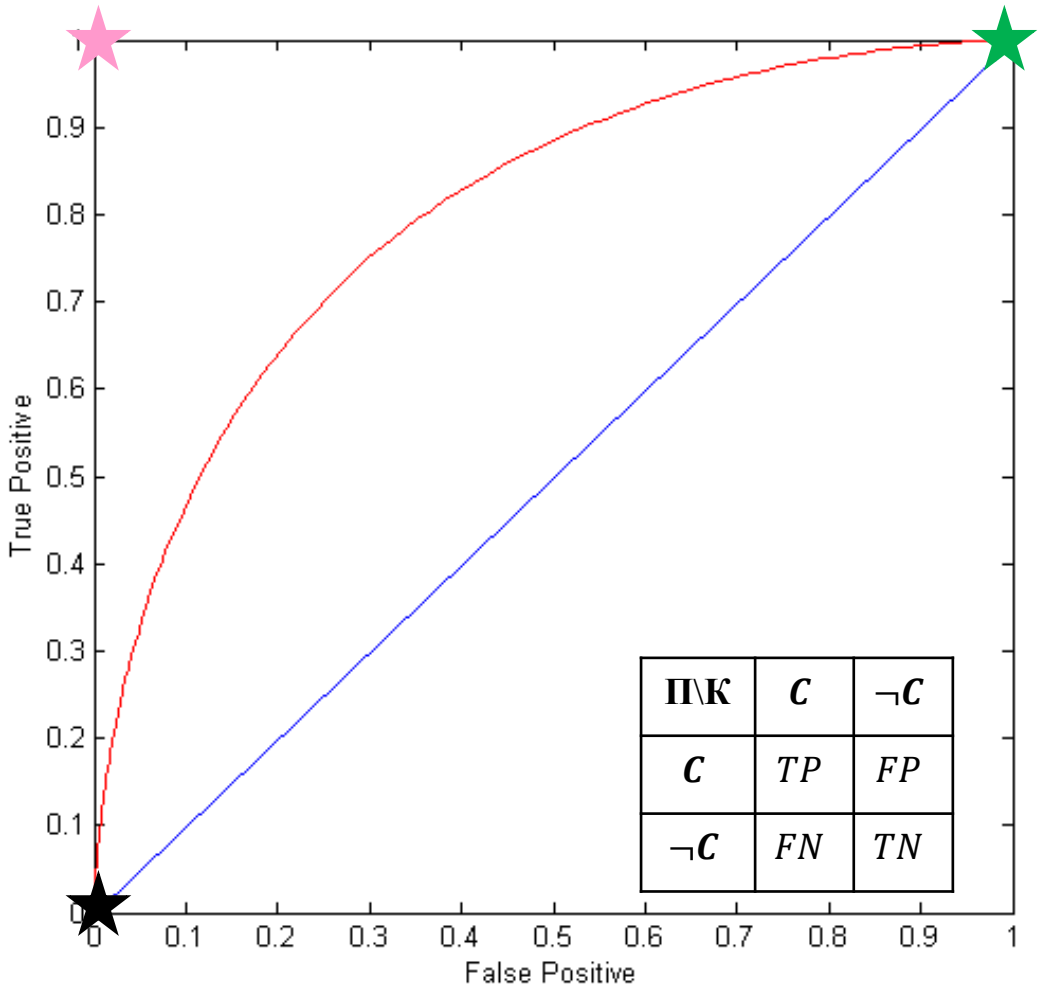
Кривые ROC (Receiver Operating Characteristics), оценка AUC (Area Under Curve)

- Показывает компромисс между точностью распознавания объектов классов C и $\neg C$

$$FPrate = \frac{FP}{N} = \frac{FP}{FP + TN}$$

$$TPrate = \frac{TP}{P} = \frac{TP}{TP+FN}$$

- Плоскость (FPR, TPR)
  - (0,0): $\forall x \in \neg C$
  - (1,1): $\forall x \in C$
  - (0,1): идеал
- Диагональ – угадывание
- Под диагональю: результат противоположен истине
- Показатель качества – площадь под кривой

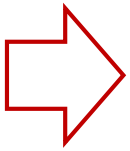


П\К	C	$\neg C$
C	TP	FP
$\neg C$	FN	TN

Построение ROC кривых

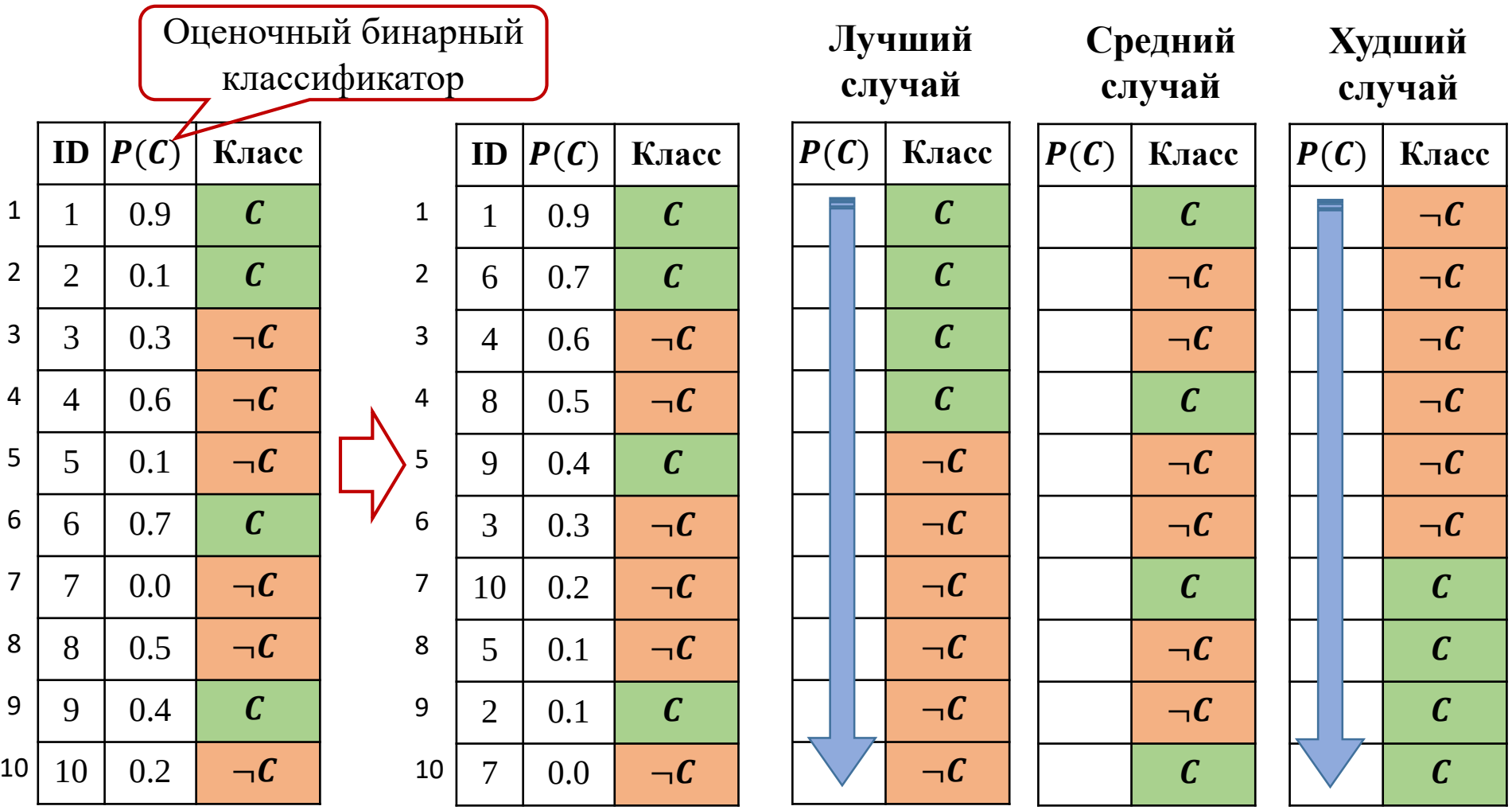
Оценочный бинарный классификатор

	ID	$P(C)$	Класс
1	1	0.9	C
2	2	0.1	C
3	3	0.3	$\neg C$
4	4	0.6	$\neg C$
5	5	0.1	$\neg C$
6	6	0.7	C
7	7	0.0	$\neg C$
8	8	0.5	$\neg C$
9	9	0.4	C
10	10	0.2	$\neg C$



	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$

Построение ROC кривых



	ID	$P(C)$	Класс
1	1	0.9	C
2	2	0.1	C
3	3	0.3	$\neg C$
4	4	0.6	$\neg C$
5	5	0.1	$\neg C$
6	6	0.7	C
7	7	0.0	$\neg C$
8	8	0.5	$\neg C$
9	9	0.4	C
10	10	0.2	$\neg C$

	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$

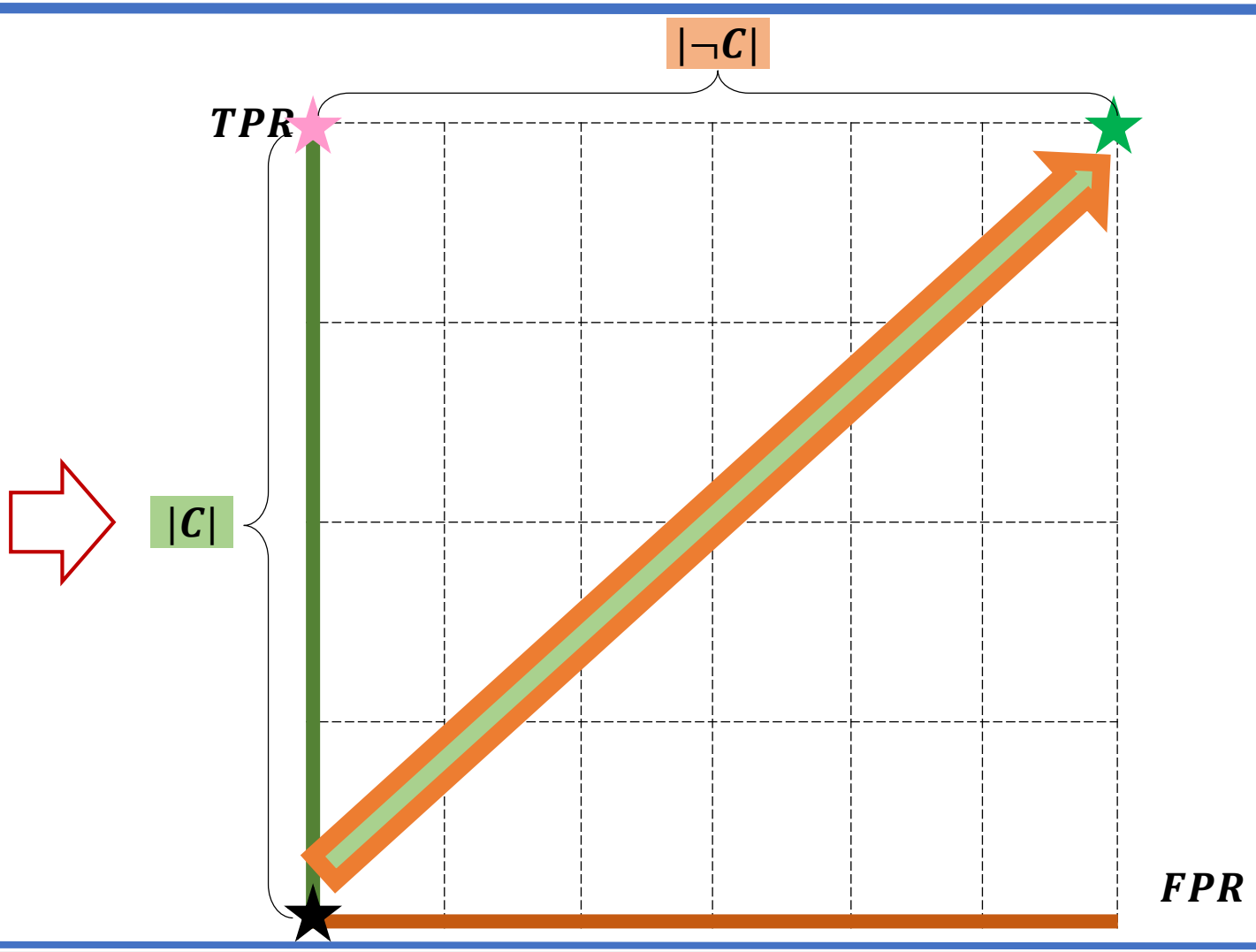
$P(C)$	Класс
	C
	C
	C
	C
	$\neg C$
	$\neg C$
	$\neg C$
	$\neg C$
	$\neg C$
	$\neg C$

$P(C)$	Класс
	C
	$\neg C$
	$\neg C$
	C
	$\neg C$
	$\neg C$
	C
	$\neg C$
	$\neg C$
	C

$P(C)$	Класс
	$\neg C$
	$\neg C$
	$\neg C$
	$\neg C$
	$\neg C$
	$\neg C$
	C
	C
	C
	C

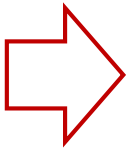
Построение ROC кривых

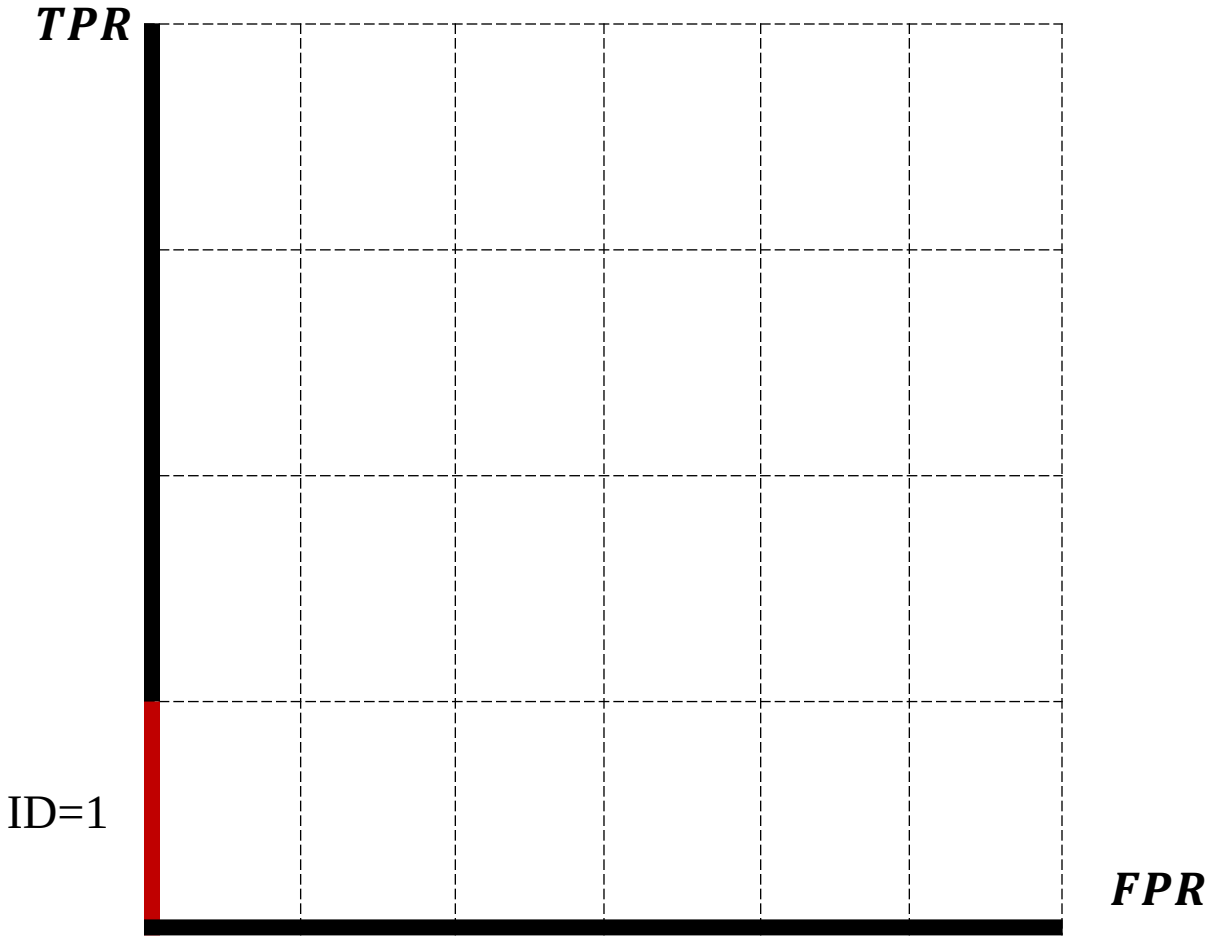
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$



Построение ROC кривых

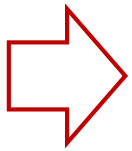
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$

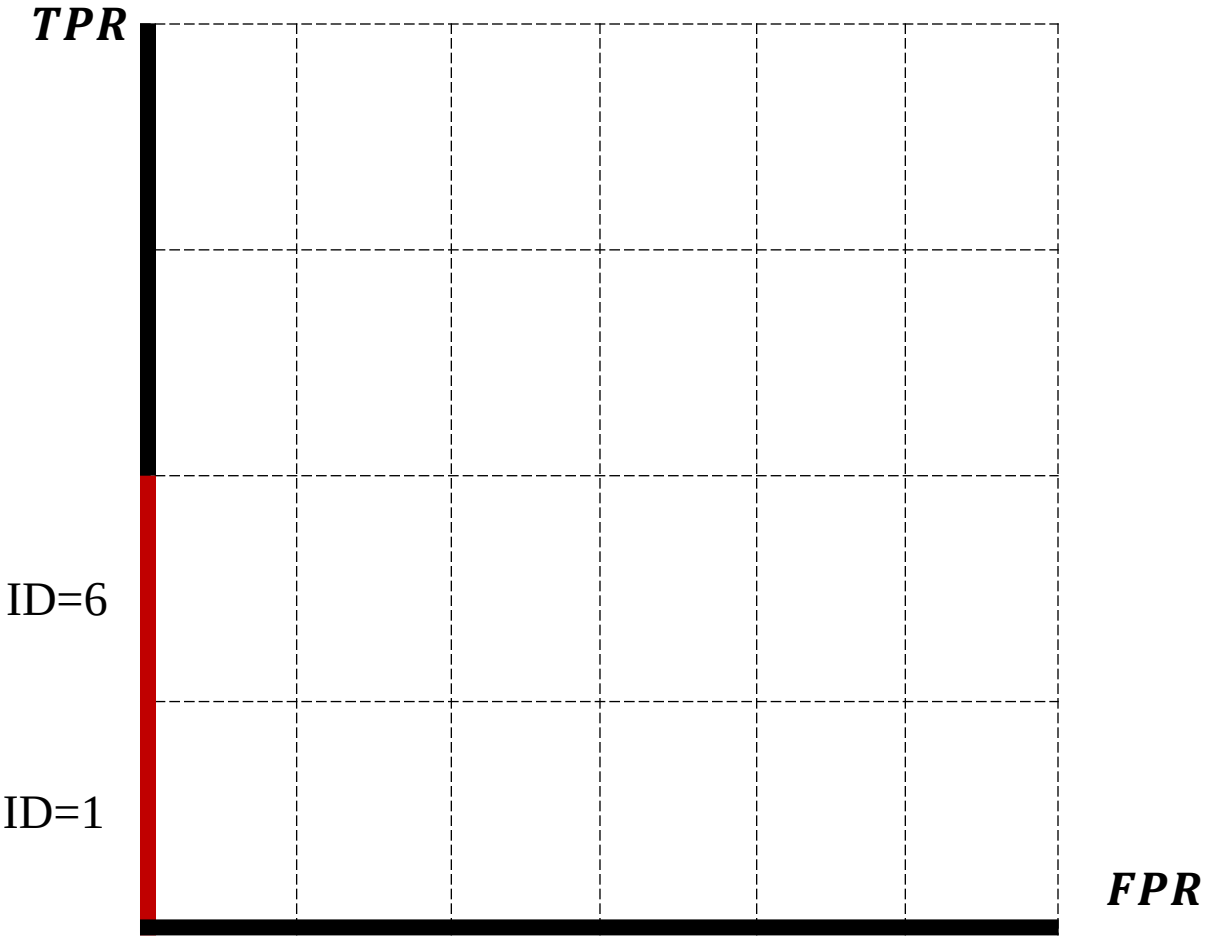




Построение ROC кривых

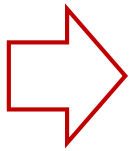
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$

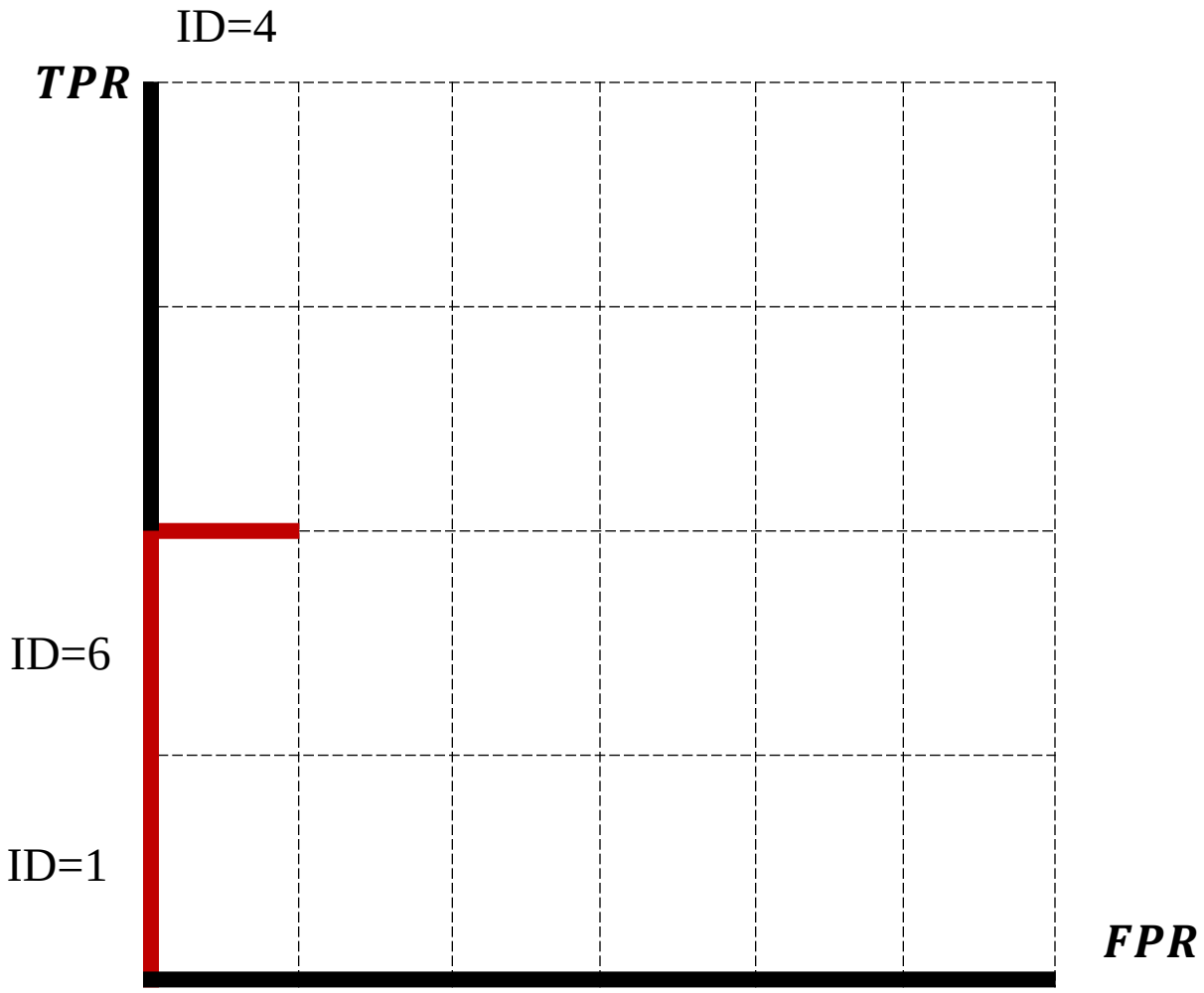




Построение ROC кривых

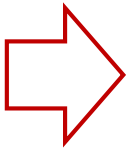
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$

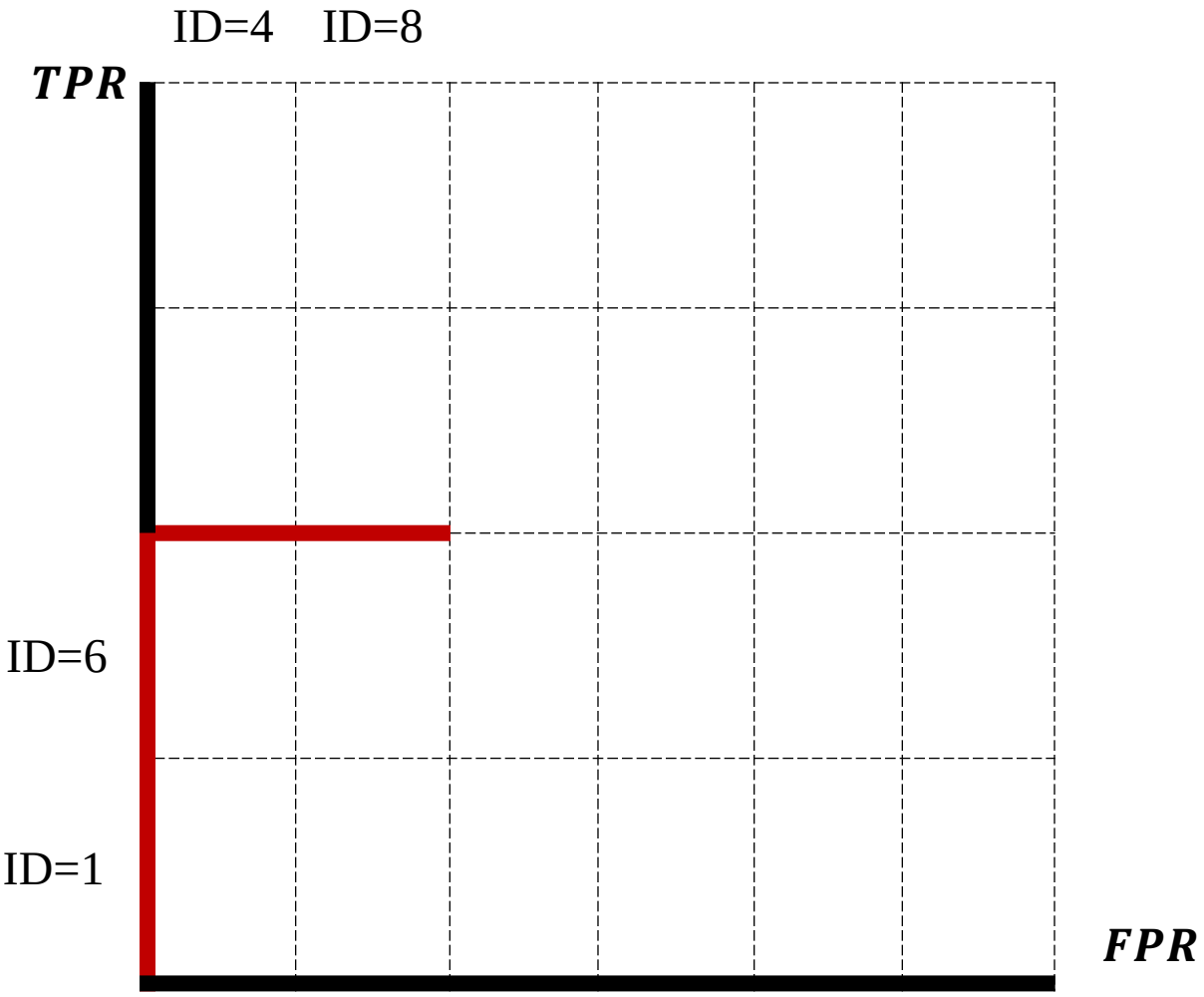




Построение ROC кривых

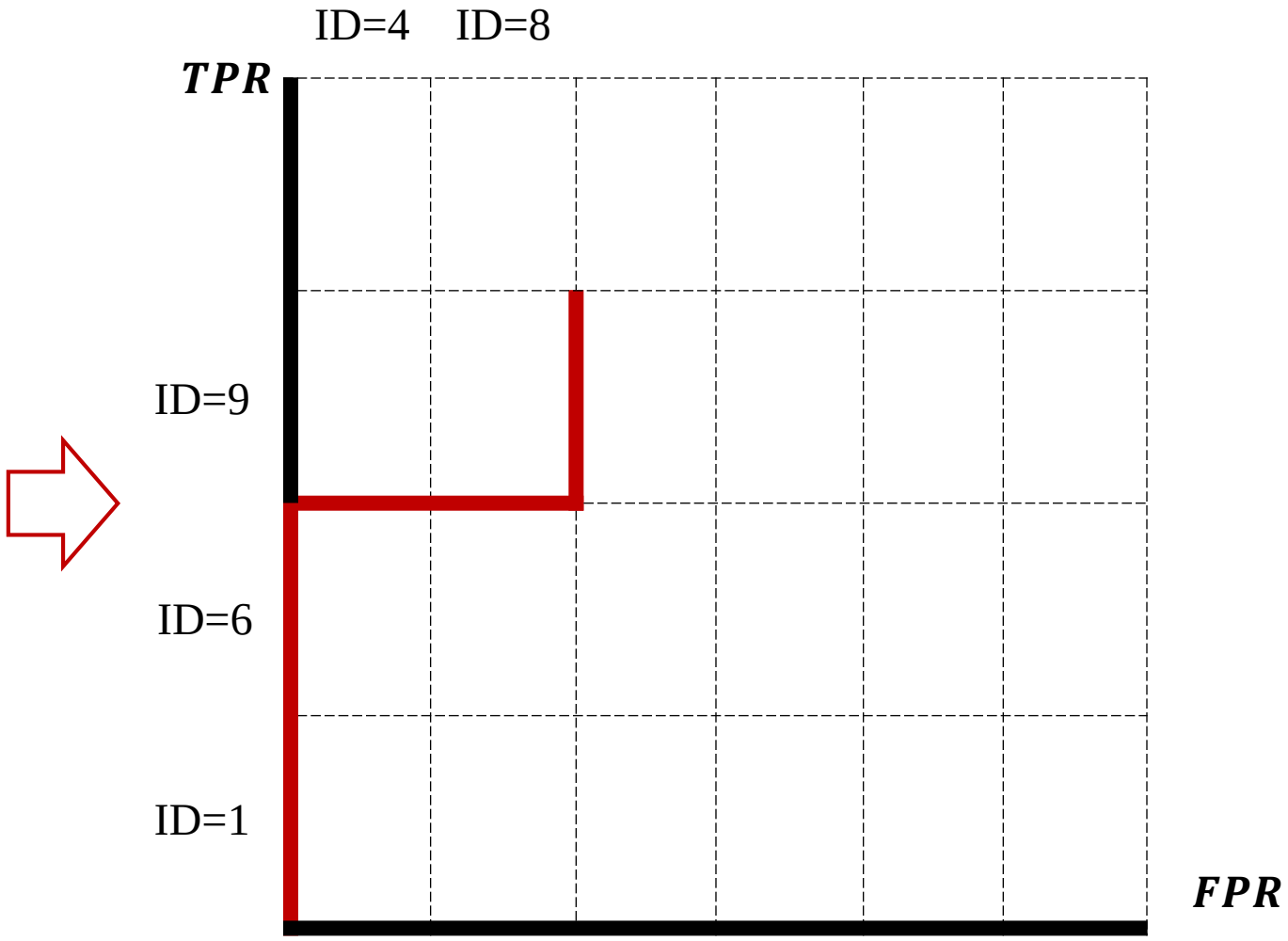
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$





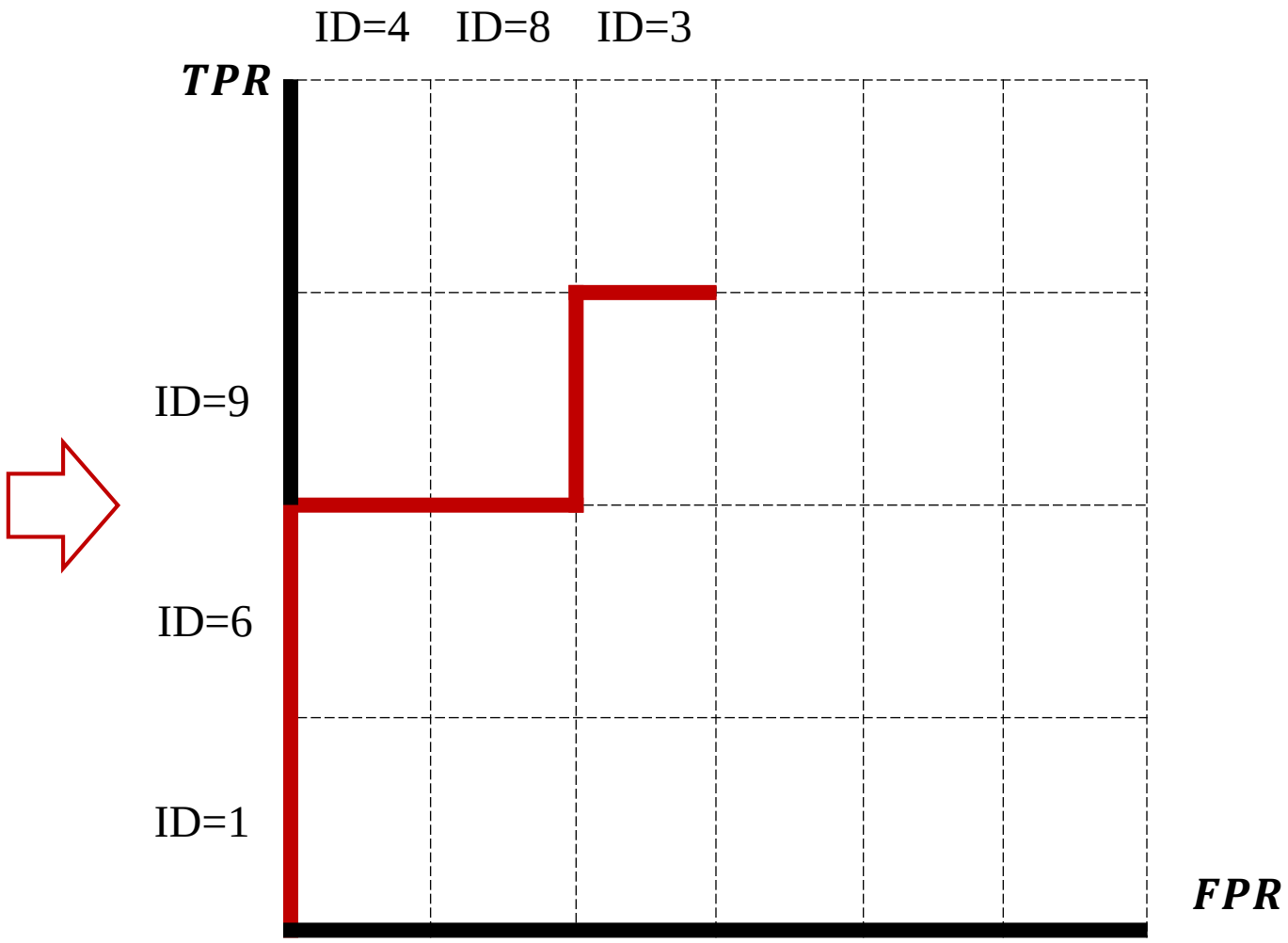
Построение ROC кривых

	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$



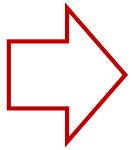
Построение ROC кривых

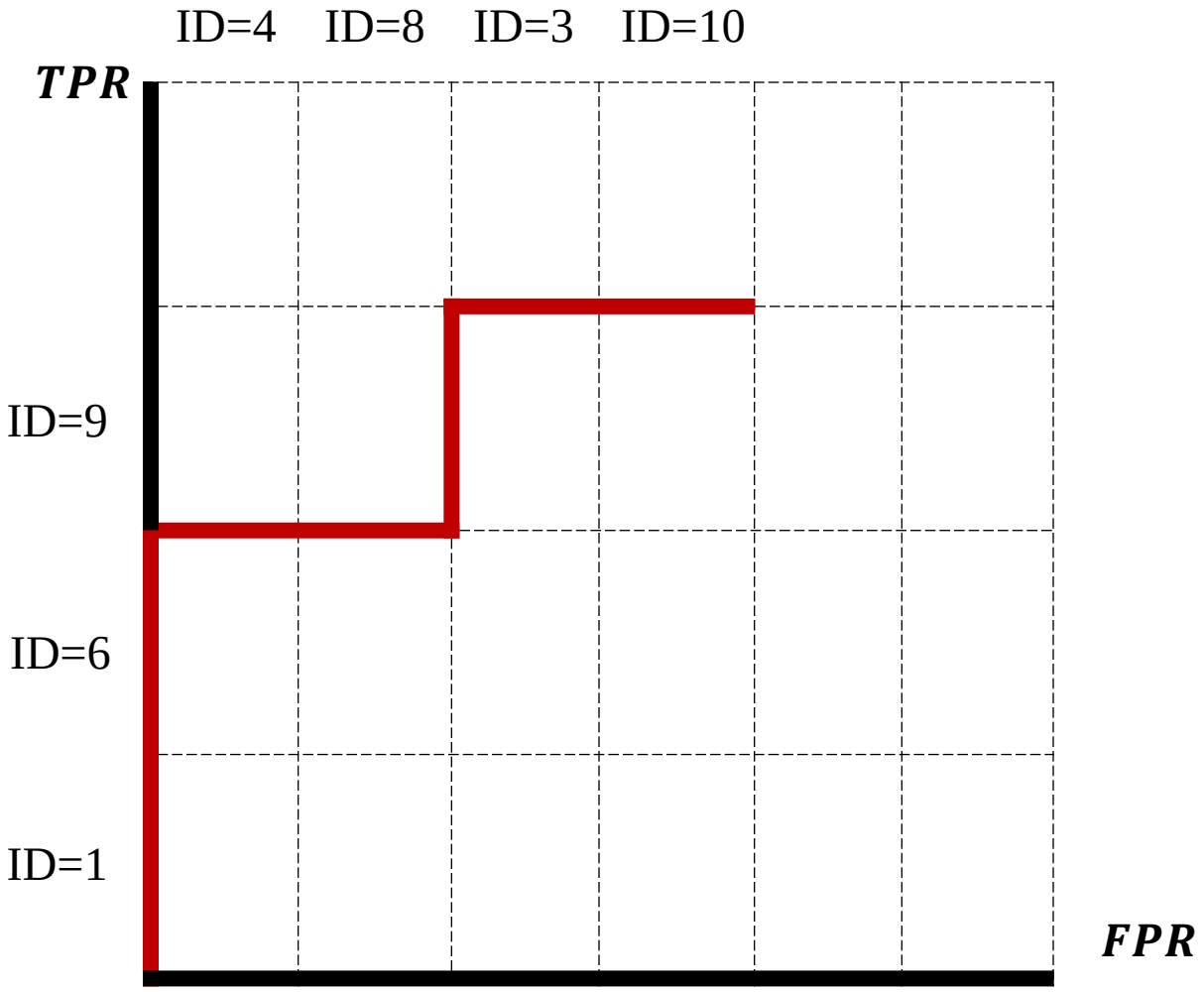
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$



Построение ROC кривых

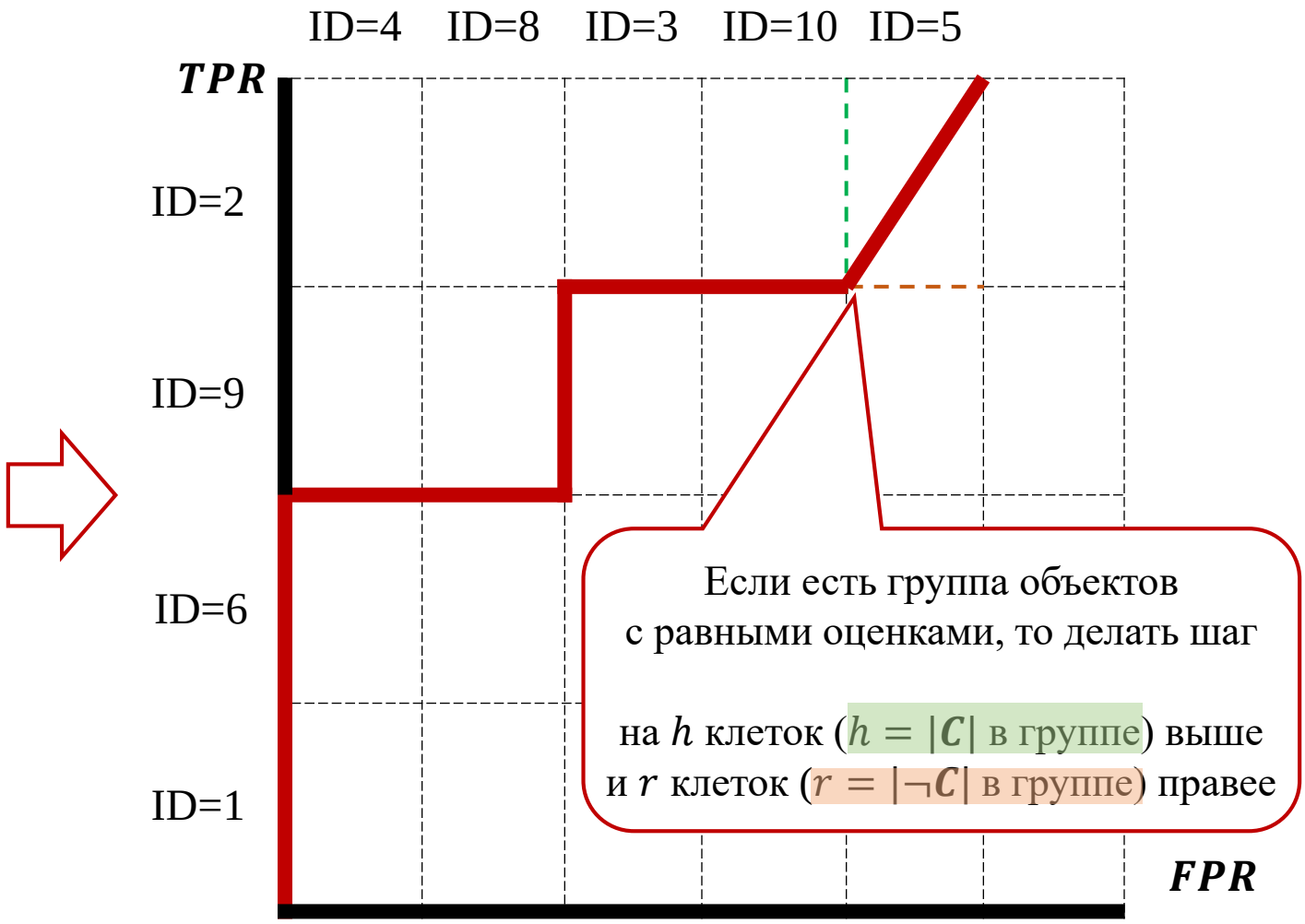
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$





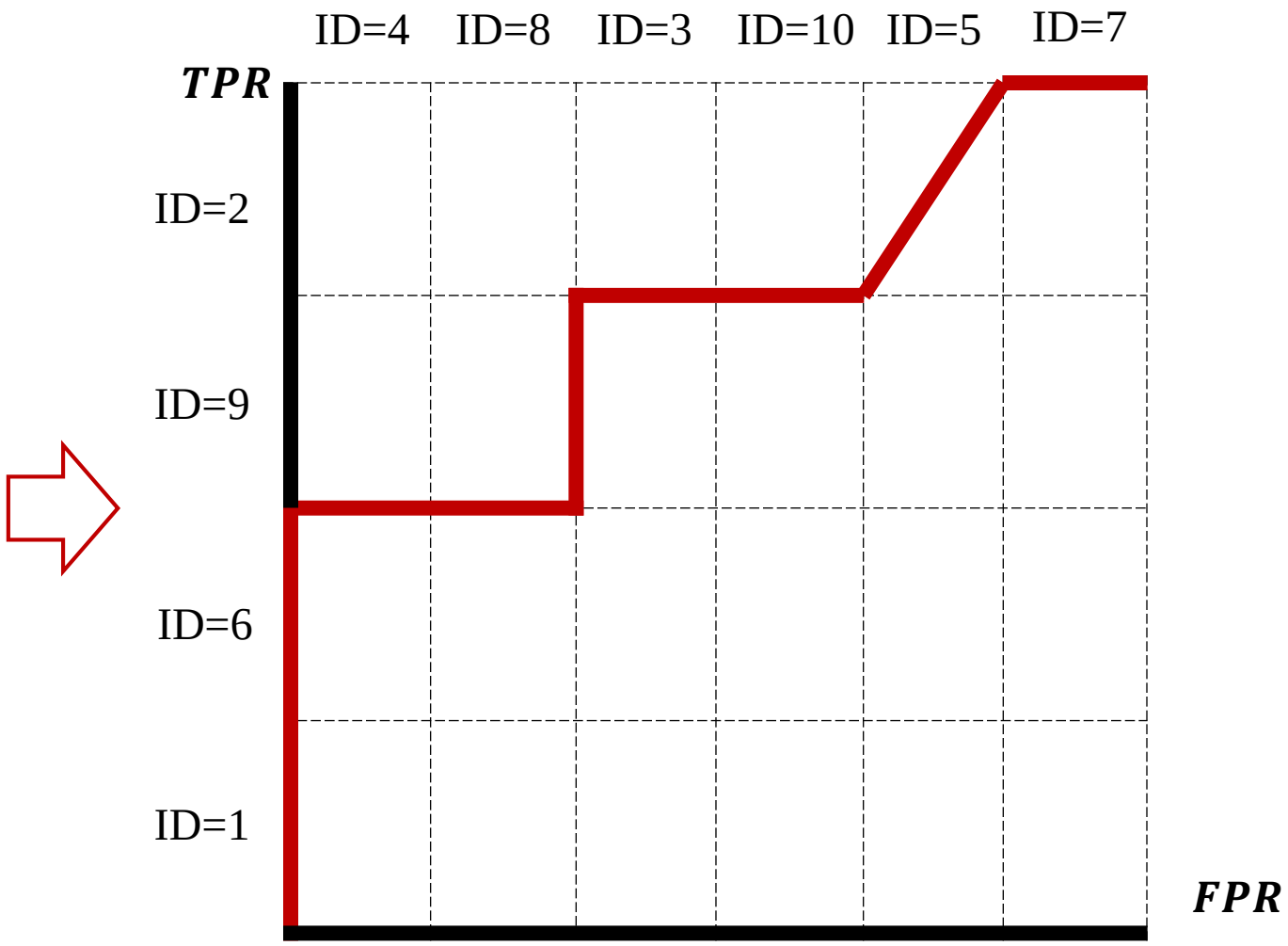
Построение ROC кривых

	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$



Построение ROC кривых

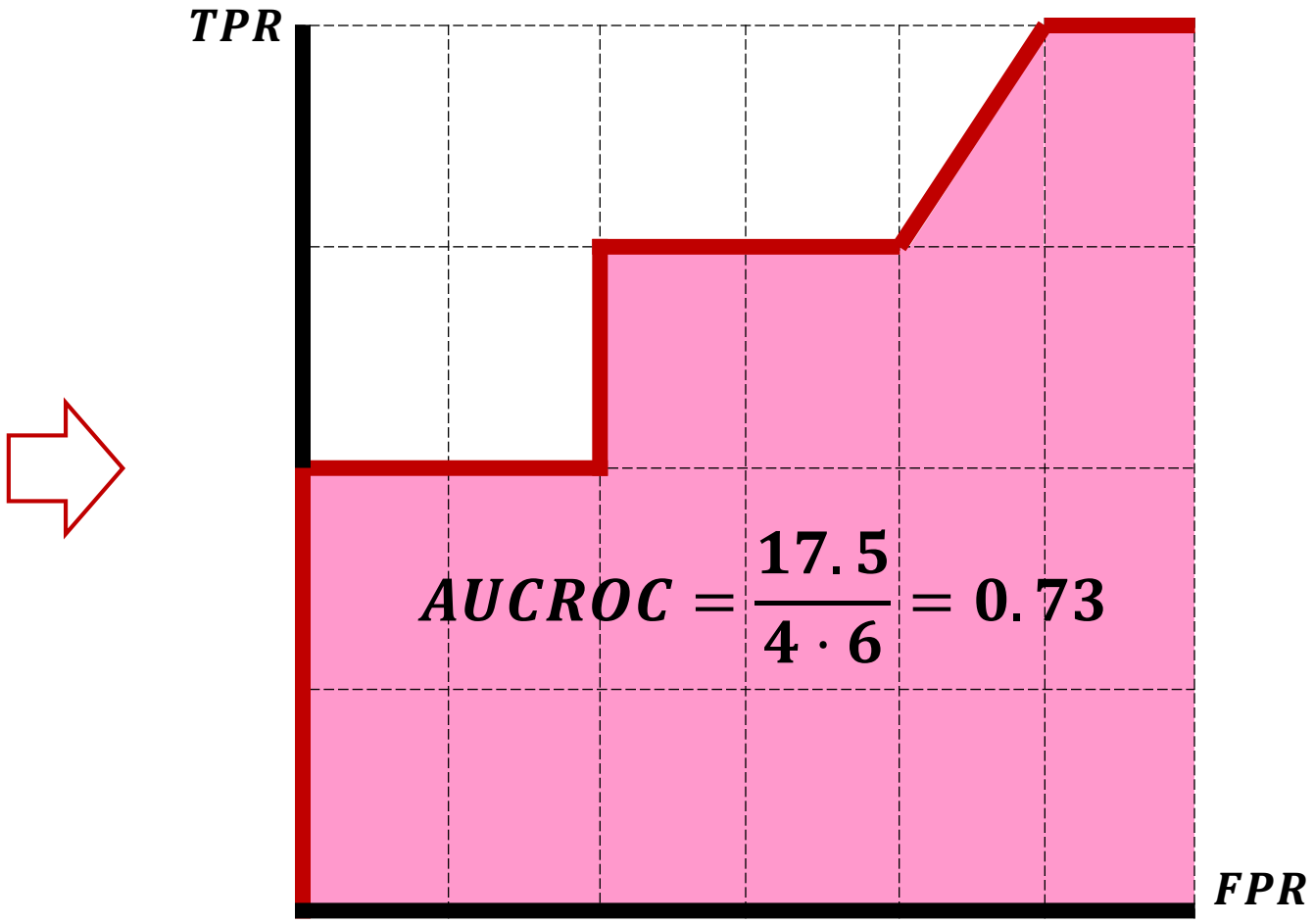
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$



Оценка AUC ROC

$$AUCROC = \frac{\{(C, \neg C)\}}{|C| \cdot |\neg C|}$$

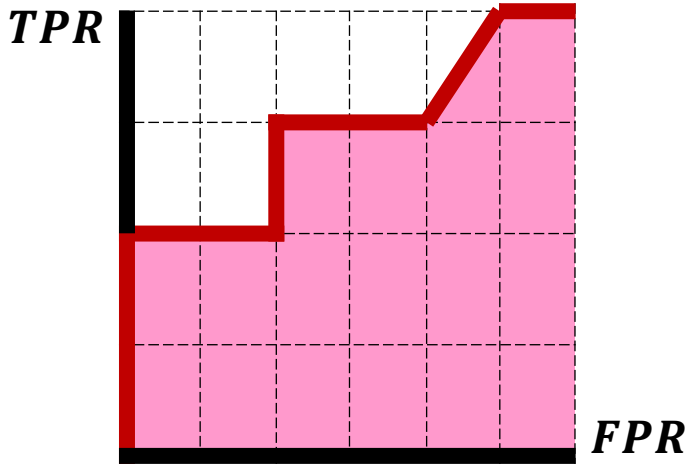
	ID	$P(C)$	Класс
1	1	0.9	C
2	6	0.7	C
3	4	0.6	$\neg C$
4	8	0.5	$\neg C$
5	9	0.4	C
6	3	0.3	$\neg C$
7	10	0.2	$\neg C$
8	5	0.1	$\neg C$
9	2	0.1	C
10	7	0.0	$\neg C$



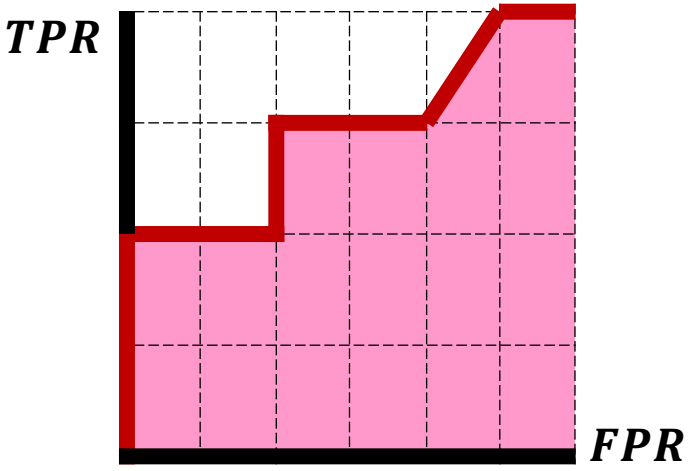
Выбор порога бинаризации на основе AUC ROC

	ID	$P(C)$	Класс	Прогноз при $P(C) \geq t$	
				$t = 0.4$	$t = 0.7$
1	1	0.9	C		
2	6	0.7	C		
3	4	0.6	$\neg C$		
4	8	0.5	$\neg C$		
5	9	0.4	C		
6	3	0.3	$\neg C$		
7	10	0.2	$\neg C$		
8	5	0.1	$\neg C$		
9	2	0.1	C		
10	7	0.0	$\neg C$		

$t = 0.4$	
FPR	?/6
TPR	?/4



$t = 0.7$	
FPR	?/6
TPR	?/4

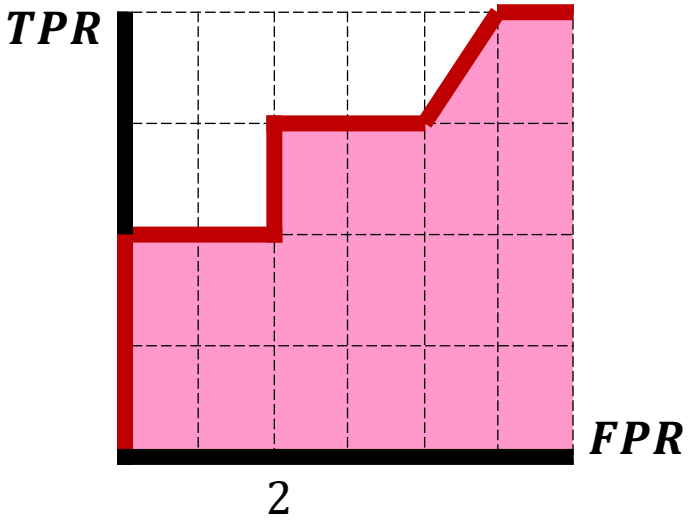


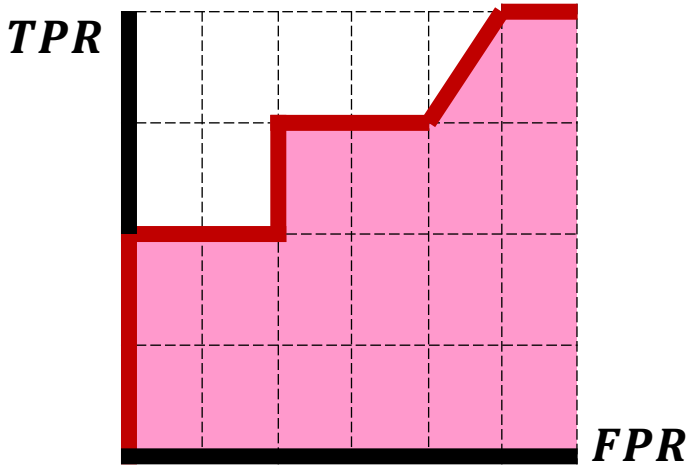
Выбор порога бинаризации на основе AUC ROC

	ID	$P(C)$	Класс	Прогноз при $P(C) \geq t$	
				$t = 0.4$	$t = 0.7$
1	1	0.9	C	C	
2	6	0.7	C	C	
3	4	0.6	$\neg C$	C	
4	8	0.5	$\neg C$	C	
5	9	0.4	C	C	
6	3	0.3	$\neg C$	$\neg C$	
7	10	0.2	$\neg C$	$\neg C$	
8	5	0.1	$\neg C$	$\neg C$	
9	2	0.1	C	$\neg C$	
10	7	0.0	$\neg C$	$\neg C$	

$t = 0.4$	
FPR	2/6
TPR	?/4

$t = 0.7$	
FPR	?/6
TPR	?/4



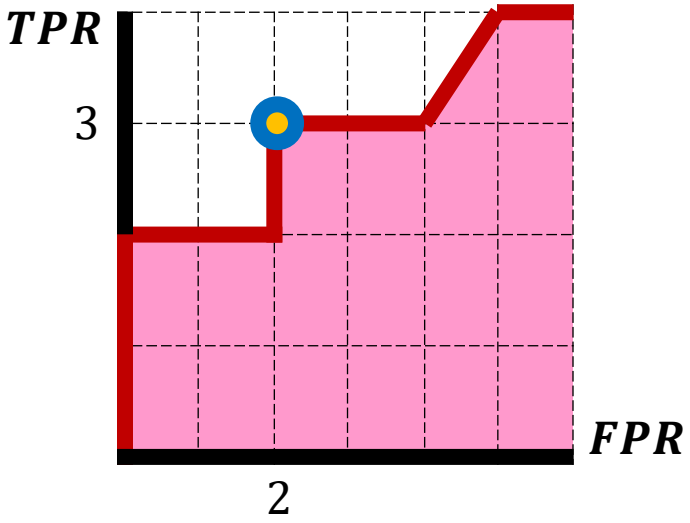


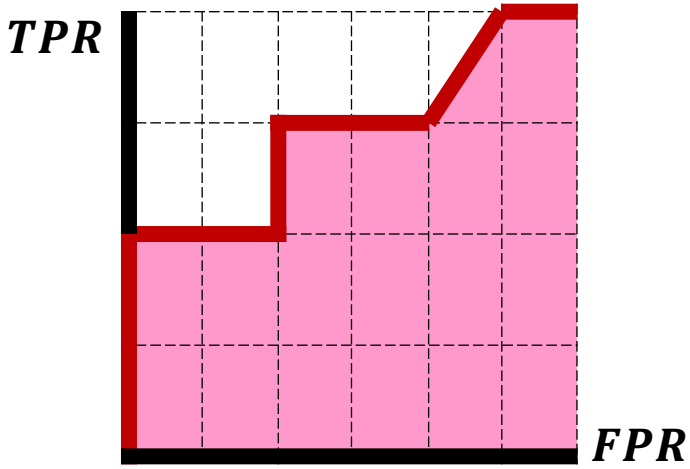
Выбор порога бинаризации на основе AUC ROC

	ID	$P(C)$	Класс	Прогноз при $P(C) \geq t$	
				$t = 0.4$	$t = 0.7$
1	1	0.9	C	C	
2	6	0.7	C	C	
3	4	0.6	$\neg C$	C	
4	8	0.5	$\neg C$	C	
5	9	0.4	C	C	
6	3	0.3	$\neg C$	$\neg C$	
7	10	0.2	$\neg C$	$\neg C$	
8	5	0.1	$\neg C$	$\neg C$	
9	2	0.1	C	$\neg C$	
10	7	0.0	$\neg C$	$\neg C$	

$t = 0.4$	
FPR	$2/6$
TPR	$3/4$

$t = 0.7$	
FPR	$?/6$
TPR	$?/4$

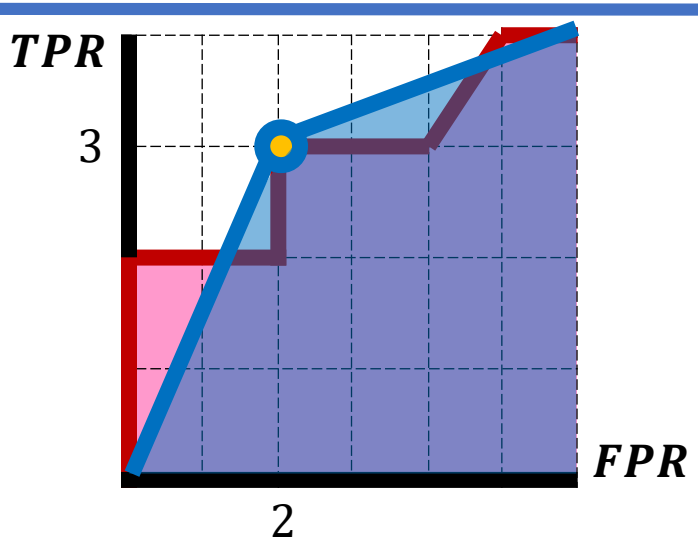




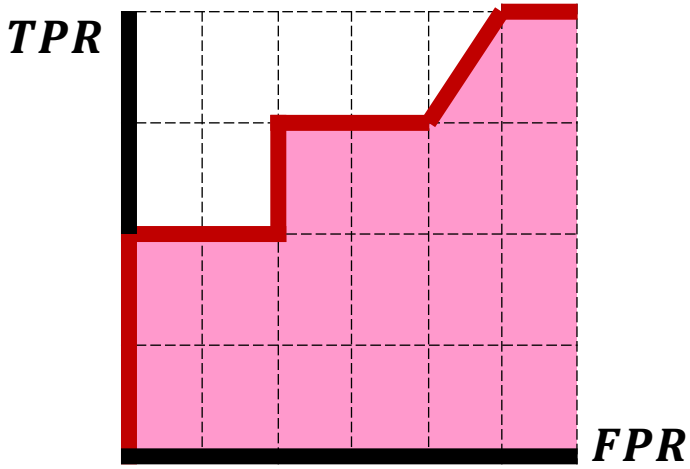
Выбор порога бинаризации на основе AUC ROC

	ID	$P(C)$	Класс	Прогноз при $P(C) \geq t$	
				$t = 0.4$	$t = 0.7$
1	1	0.9	C	C	
2	6	0.7	C	C	
3	4	0.6	$\neg C$	C	
4	8	0.5	$\neg C$	C	
5	9	0.4	C	C	
6	3	0.3	$\neg C$	$\neg C$	
7	10	0.2	$\neg C$	$\neg C$	
8	5	0.1	$\neg C$	$\neg C$	
9	2	0.1	C	$\neg C$	
10	7	0.0	$\neg C$	$\neg C$	

$t = 0.4$	
FPR	$\mathbf{2}/6$
TPR	$\mathbf{3}/4$



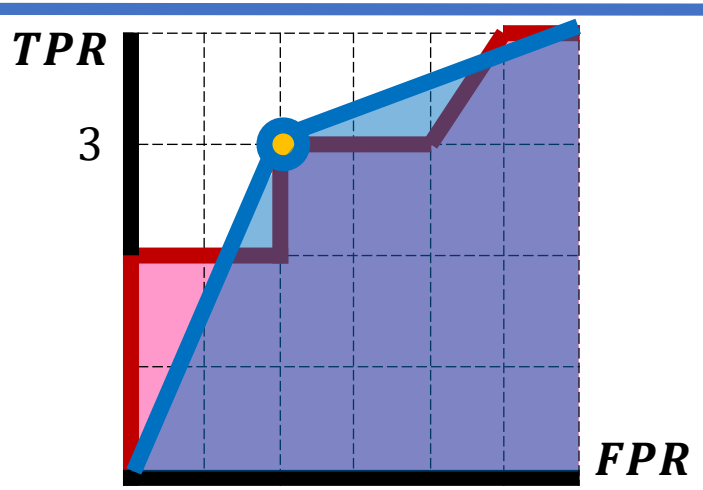
$t = 0.7$	
FPR	?/6
TPR	?/4



Выбор порога бинаризации на основе AUC ROC

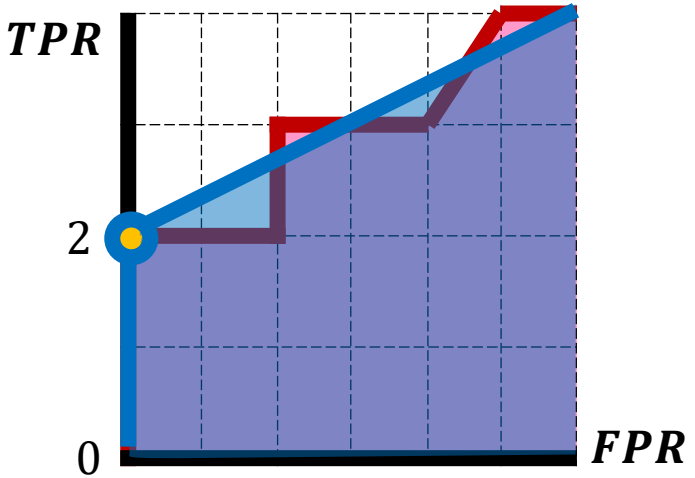
	ID	$P(C)$	Класс	Прогноз при $P(C) \geq t$	
				$t = 0.4$	$t = 0.7$
1	1	0.9	C	C	C
2	6	0.7	C	C	C
3	4	0.6	$\neg C$	C	$\neg C$
4	8	0.5	$\neg C$	C	$\neg C$
5	9	0.4	C	C	$\neg C$
6	3	0.3	$\neg C$	$\neg C$	$\neg C$
7	10	0.2	$\neg C$	$\neg C$	$\neg C$
8	5	0.1	$\neg C$	$\neg C$	$\neg C$
9	2	0.1	C	$\neg C$	$\neg C$
10	7	0.0	$\neg C$	$\neg C$	$\neg C$

$t = 0.4$	
FPR	$2/6$
TPR	$3/4$

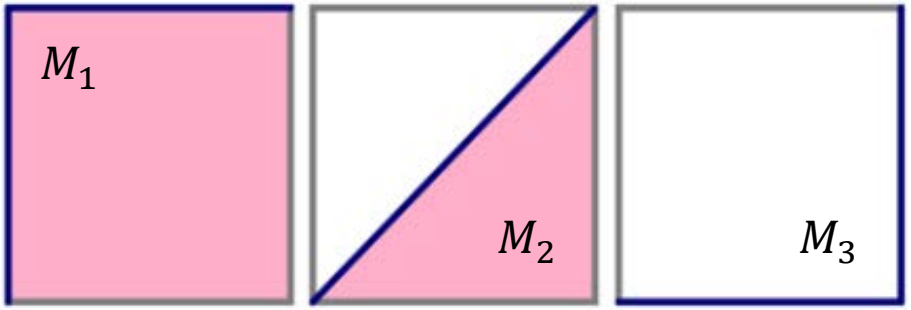


$AUC_{t=0.7} > AUC_{t=0.4}$

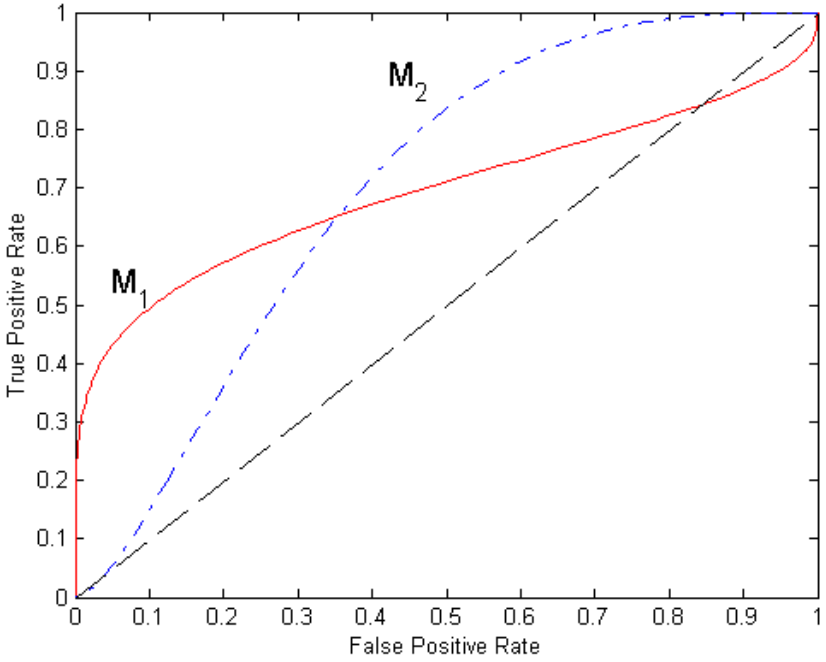
$t = 0.7$	
FPR	$0/6$
TPR	$2/4$



Выбор модели на основе AUC ROC



Классификатор	AUC ROC	Качество
M_1	1	Наилучший
M_2	0.5	Случайный
M_3	0	Наихудший



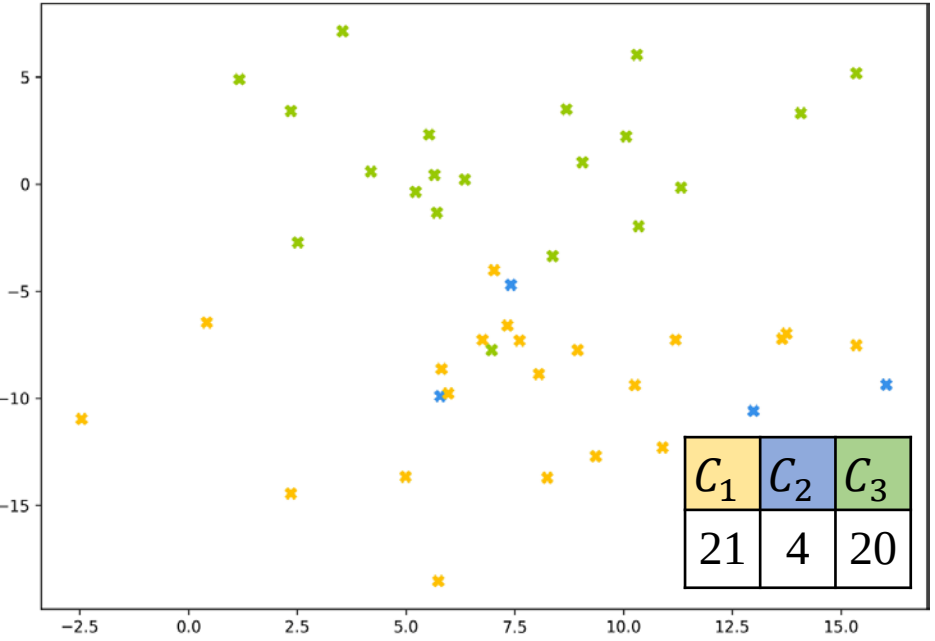
- Классификатор M_1 лучше при малых значениях FPR
- Классификатор M_2 лучше при больших значениях FPR

Качество многоклассовой классификации

- Матрица ошибок класса $\forall C_i \in \boldsymbol{C} = \{C_1, \dots, C_K\}$
- Микроусреднение
  - $\overline{TP} = \frac{1}{K}\sum_{i=1}^{K} TP_i$, $\overline{FP} = \dots$
  - $Precision = \frac{\overline{TP}}{\overline{TP}+\overline{FP}}$, $Recall = \dots$
  - Вклад маломощного класса в метрику мал
- Макроусреднение
  - $Precision_{C_i} = \frac{TP_i}{TP_i+FP_i}$, $Recall_{C_i} = \dots$
  - $Precision = \frac{1}{K}\sum_{i=1}^{K} Precision_{C_i}$, $Recall = \dots$
  - Вклад всех классов в метрику одинаков

П\К	C_i	$\boldsymbol{C}\backslash C_i$
C_i	TP	FP
$\boldsymbol{C}\backslash C_i$	FN	TN

Микроусреднение *vs.* макроусреднение



C_1	C_2	C_3
21	4	20

П\К	C_1	C_2	C_3
C_1	20	4	1
C_2	1	0	0
C_3	0	0	19

- Микроусреднение

$$Precision = \frac{\frac{1}{3}(20+0+19)}{\frac{1}{3}(20+0+19)+\frac{1}{3}(5+1+0)} = 0.87$$

- Макроусреднение

$$Precision = \frac{1}{3}\left(\frac{20}{20+5}+\frac{0}{0+1}+\frac{19}{19+0}\right) = 0.6$$

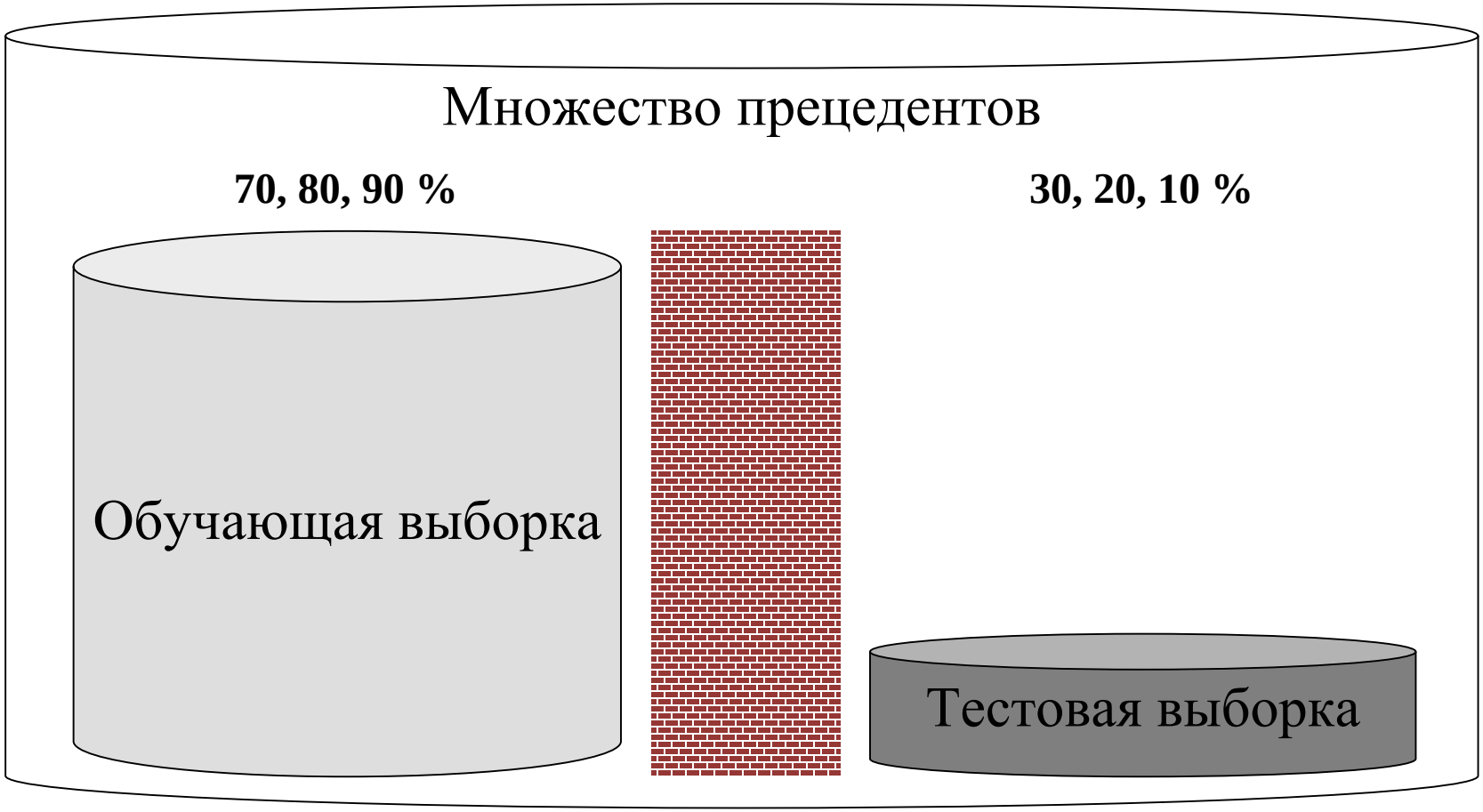
Прочие аспекты оценки качества классификации

- Скорость
  - обучение
  - применение
- Робастность (устойчивость к шумам и пропускам в данных)
- Интерпретируемость результатов
- Прочее (высота/ширина дерева решений и т.п.)

Содержание

- Основные понятия
- Деревья решений
- Байесовская классификация
- Классификация по ближайшим соседям
- **Оценка качества классификации**
  - **Меры качества**
  - **Подготовка тестовой выборки**
- Ансамблевая классификация

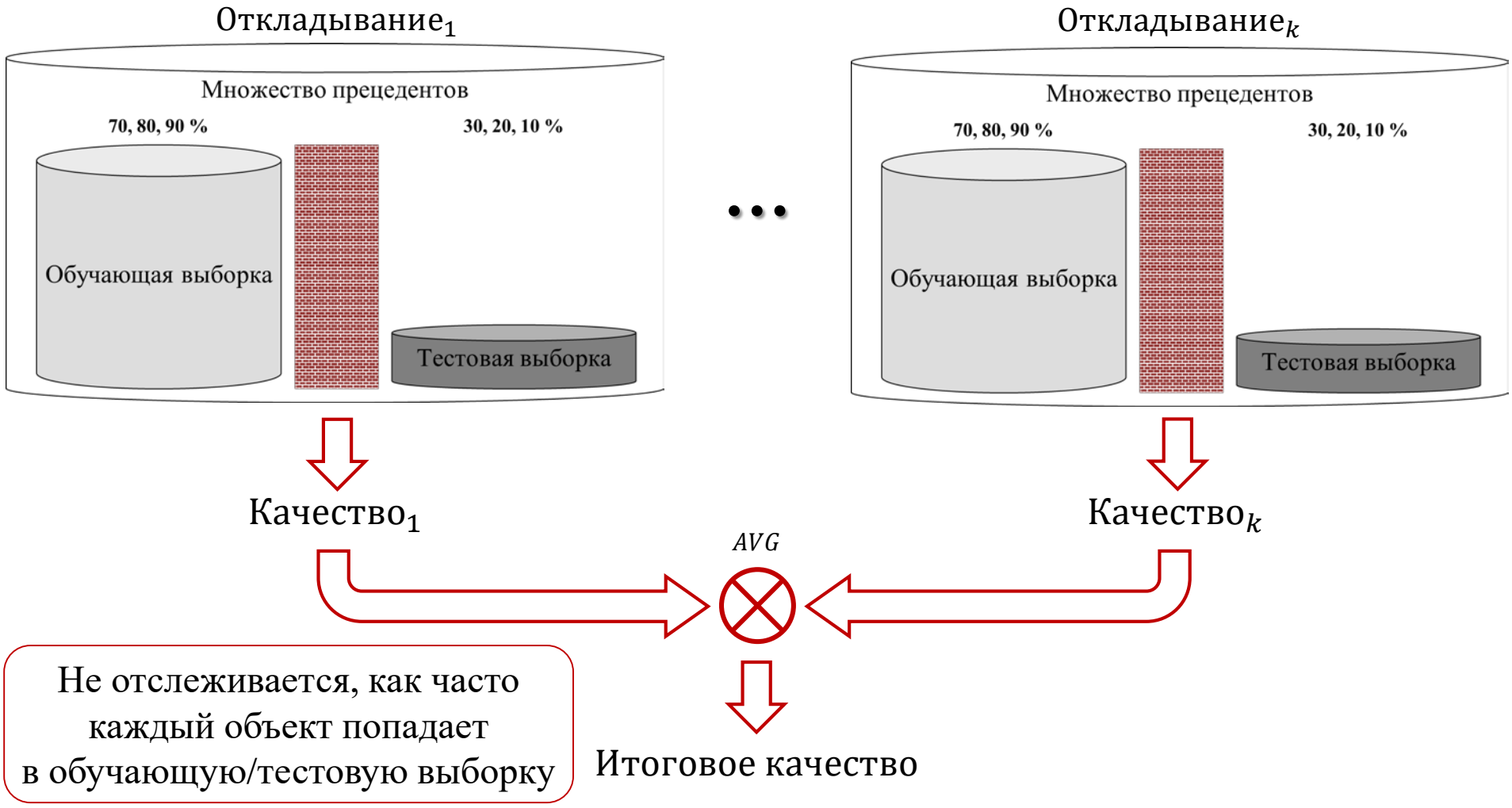
Откладывание (hold-out)



Ограничения откладывания

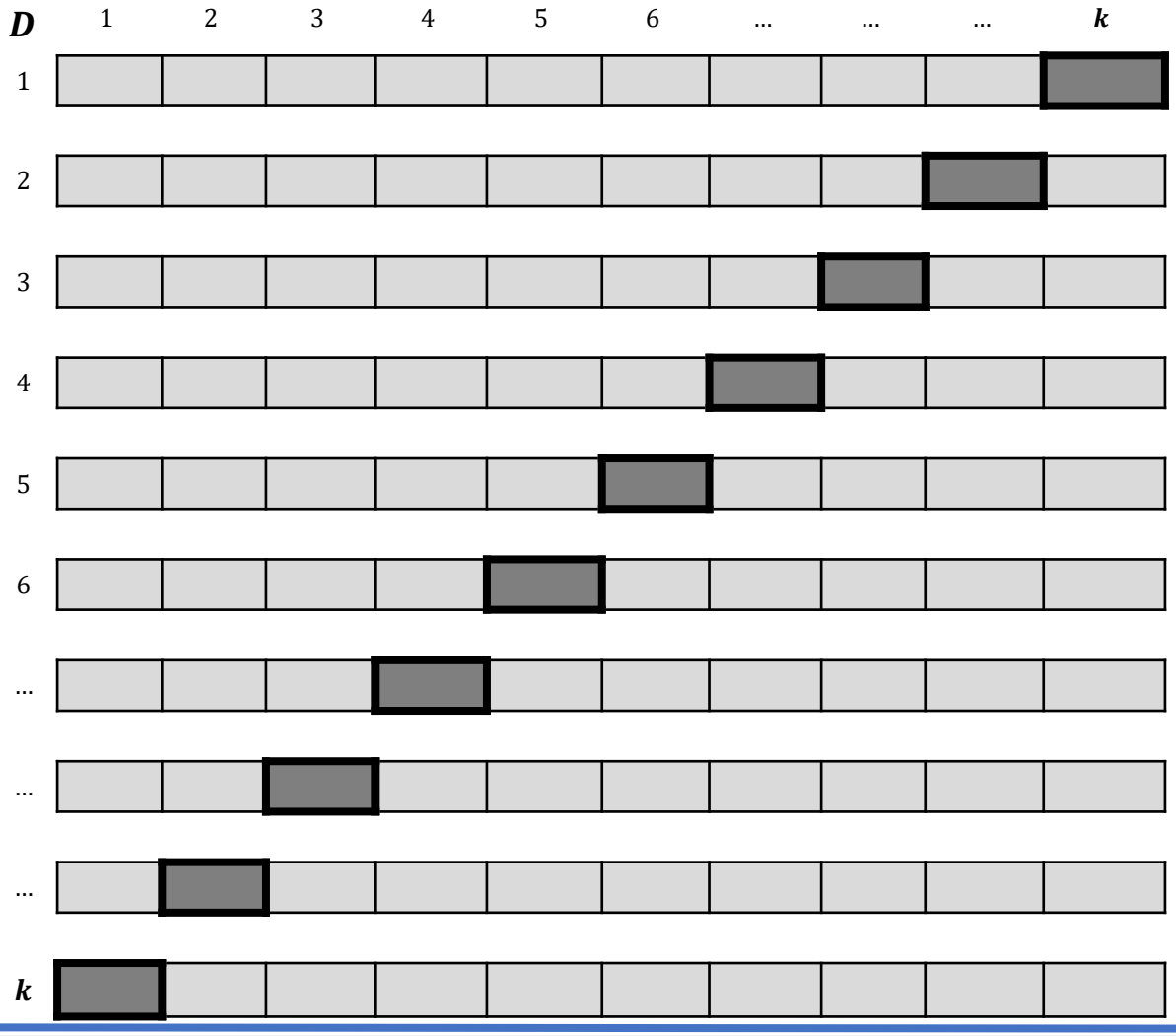
1. Обучающая выборка меньше, чем могла бы быть, из-за направления части данных в тестовую выборку. Модель, обученная на 100% имеющихся данных, могла бы быть лучше
2. Модель может сильно зависеть от состава обучающей и тестовой выборок. Чем меньше размер обучающей выборки, тем больше разброс в модели. Если обучающая выборка слишком велика, то точность, вычисленная на основе меньшей тестовой выборки, менее надежна
3. Обучающая и тестовая выборки не являются независимыми друг от друга, в них возможны перекосы баланса классов: класс, чрезмерно представленный в одной выборке, будет недостаточно представлен в другой выборке, и наоборот

Случайный отбор (random sampling)



Не отслеживается, как часто каждый объект попадает в обучающую/тестовую выборку

Перекрестная проверка (*k*-fold cross validation)



Обучающая выборка | Тестовая выборка

Типичное значение $k = 10$

AVG

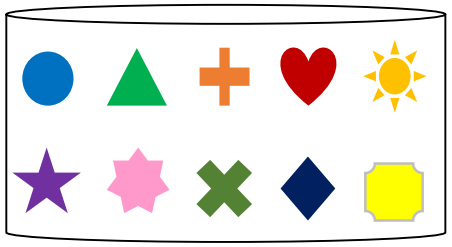
Итоговое качество

$k = |D|$
leave-one-out

- накладные расходы на вычисления
- большой разброс значения точности

Самонастройка (0.632 bootstrapping)

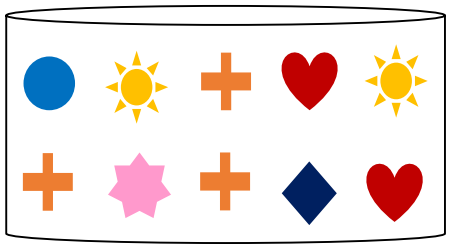
Исходная выборка

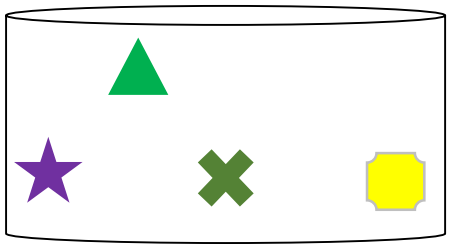


n объектов

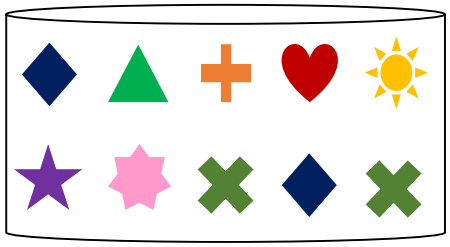
Обучающие выборки **Тестовые выборки**

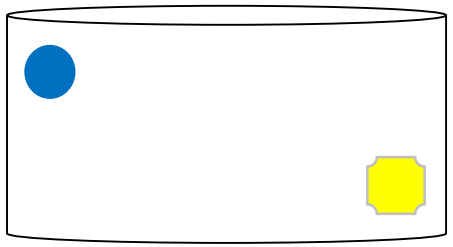
1.





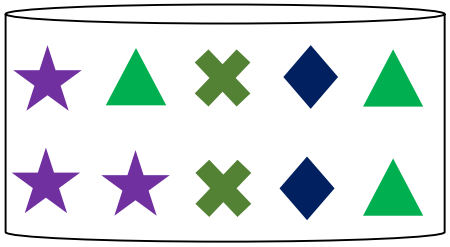
2.

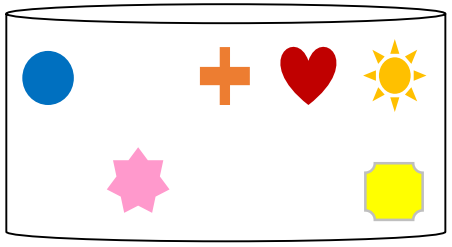




... ...

k.





$$P(o \notin train_{set}) = \left(1 - \frac{1}{n}\right)^n,$$

$$\lim_{n \to +\infty} \left(1 - \frac{1}{n}\right)^n = \frac{1}{e} = 0.368$$

$$Accuracy(M) = \frac{1}{k} \sum_{i=1}^{k} (0.632 \cdot Accuracy(M_i)_{test_{set}} + 0.368 \cdot Accuracy(M_i)_{train_{set}})$$

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN: 978-0123814791
  - 8.5 Model Evaluation and Selection, pp. 364-377
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN: 978-0-13-312890-1
  - 3.5 Model Selection, pp. 156-162, 3.6 Model Evaluation, pp. 164-165