

March 02, Kyiv

Data Science & Mathematical Modeling Bachelor Program

Course “Basics of Machine Learning”

Lecture 2-3: Classification Metrics



Oleg CHERTOV

Professor, Sc.D. (Doctor Habilitatus),
Head of the Applied Mathematics Department



Applied Mathematics Department
Igor Sikorsky Kyiv Polytechnic Institute
Ukraine



Lecture 2. Classification Metrics (класифікаційні метрики)

1

Задача
медичної
діагностики як
задача
класифікації

2

1. Confusion
matrix,
2. Accuracy,
3. Precision
and Recall

3

4. F-Scores
5. MCC
6. Cohen's
Kappa
7. Balanced
Accuracy

4

Comparison of
metrics #3-7

5

8. AUC: Precision-
Recall curve
9. AUC: ROC curve
10. Youden's J
statistic
11. Gini coefficient

Задача медичної діагностики (пошуку хворих) як задача класифікації

- **Дано: задача класифікації**

$X^\ell = \{x_1, \dots, x_\ell\}$ — вибірка

$y_i = y(x_i) \in \{0, 1\}$, $i = 1, \dots, \ell$ — відомі бінарні відповіді

- $a : X \rightarrow Y$ — алгоритм (розв'язувальна функція, стратегія), що наближує y на всій множині об'єктів X

- **Питання: як виміряти якість $a(x)$ на вибірці X^ℓ ?**

Задача медичної діагностики (пошуку хворих) як задача класифікації

- **Дано: задача класифікації**

$X^\ell = \{x_1, \dots, x_\ell\}$ — вибірка

$y_i = y(x_i) \in \{0, 1\}$, $i = 1, \dots, \ell$ — відомі бінарні відповіді

- $a: X \rightarrow Y$ — алгоритм (розв'язувальна функція, стратегія), що наближує y на всій множині об'єктів X

- **Питання: як виміряти якість $a(x)$ на вибірці X^ℓ ?**

- **Інтуїтивна відповідь:**

доля правильних
відповідей алгоритму
(accuracy, акуратність)

$$\text{Accuracy} = \frac{\text{\#correctly classified items}}{\text{\#all classified items}}$$

Задача медичної діагностики (пошуку хворих) як задача класифікації



- **Дано: задача класифікації**

$X^\ell = \{x_1, \dots, x_\ell\}$ — вибірка

$y_i = y(x_i) \in \{0, 1\}$, $i = 1, \dots, \ell$ — відомі бінарні відповіді

- $a: X \rightarrow Y$ — алгоритм (розв'язувальна функція, стратегія), що наближує y на всій множині об'єктів X

- **Питання: як виміряти якість $a(x)$ на вибірці X^ℓ ?**

- **Інтуїтивна відповідь:**

доля правильних
відповідей алгоритму
(accuracy, акуратність)

$$\frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

Кеннет Айверсон
(Kenneth Iverson;
17.12.1920 — 19.10.2004)

дельта-функція
Кронекера
(Kronecker delta):

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

дужки Айверсона
(Iverson bracket):
 $\delta_{ij} = [i=j]$.

Accuracy is a start, but not an end

- Accuracy is usually the end result, and what the model end user will see and care about
- However, it's not very useful for diagnosing a classifier because it doesn't tell us **where** the model is making errors
- Answering this "where" question is an essential part of model-building. Hence accuracy is a start, but not an end

Classification Metrics

1. Confusion matrix
2. Accuracy
3. Precision and Recall
4. F-Scores
5. Matthews Correlation Coefficient (MCC)
6. Cohen's Kappa
7. Balanced Accuracy
8. Area under curve (AUC): Precision-Recall curve
9. AUC: Receiver operating characteristic (ROC) curve
10. Youden's J statistic
11. Gini coefficient (Gini index)

Two kinds:
binary & multi-class
classification

1. Матриця помилок (confusion matrix)

		True label	
		Хвора людина	Здорова людина
Pre-dicted label	Хвора	True Positive (TP)	False Positive (FP)
	Здорова	False Negative (FN)	True Negative (TN)

- TP = True Positive, i.e., when the actual value was 'yes' the model predicted 'yes' (i.e., correct prediction = **коректно (правильно) спрацював**)
- FP = False Positive, i.e., when the actual value was 'no' the model predicted 'yes' (i.e., wrong prediction = **хибно спрацював**)
- TN = True Negative, i.e. when the actual value was 'no' the model predicted 'no' (i.e., correct prediction = **коректно (правильно) пропустив**)
- FN = False Negative, i.e., when the actual value was 'yes' the model predicted 'no' (i.e., wrong prediction = **хибно пропустив**)

1. Матриця помилок (правильна термінологія українською)

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	правильне спрацьовування	хибне спрацьовування
	Здорова	хибний пропуск	правильний пропуск

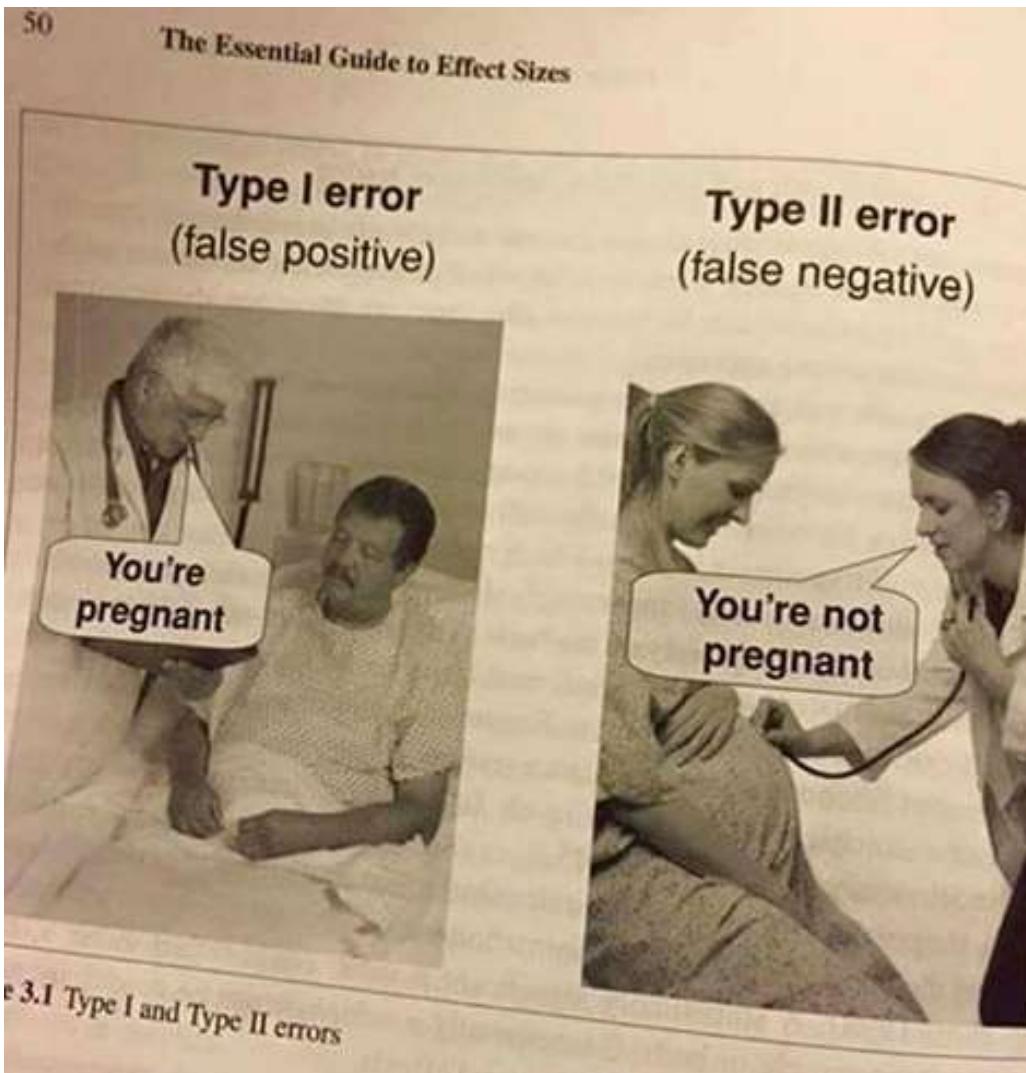
- False Positive (FP) – «хибно-позитивний»??
- False Positive (FP) – хибно спрацював,
хибне спрацьовування!

1. Матриця помилок (confusion matrix)

		True label	
		Хвора людина	Здорова людина
Pre-dicted label	Хвора	True Positive (TP)	False Positive (FP)
	Здорова	False Negative (FN)	True Negative (TN)

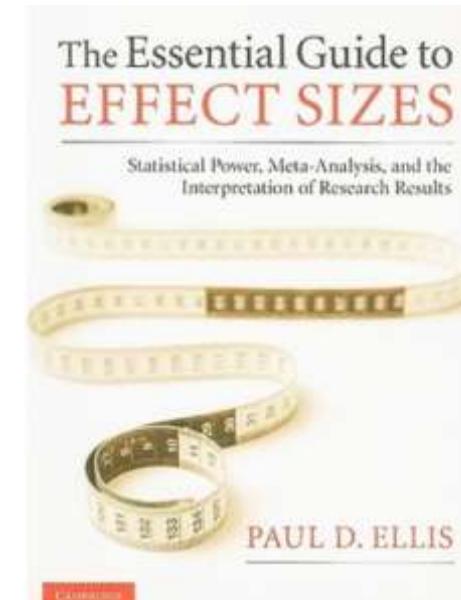
Нульова гіпотеза: всі здорові		Реальна ситуація	
		$y=1$	$y=0$
Відповідь алгоритму	$y=1$	True Positive (TP)	False Positive (FP) = помилка I роду = «хибна тривога»
	$y=0$	False Negative (FN) = помилка II роду = «пропуск цілі»	True Negative (TN)

Type I & II error (how to remember?)



null hypothesis: everyone is not pregnant

Source: Paul D. Ellis.
The Essential Guide to Effect Sizes:
Statistical Power, Meta-Analysis, and the
Interpretation of Research Results.
Cambridge University Press (2010)



Multi-class confusion matrix

Example with the MNIST dataset of handwritten digits



Sample images from MNIST dataset.

Source: <https://goo.gl/RVg4Vi>

Multi-class confusion matrix

Predictions

Reals

	0	1	2	3	4	5	6	7	8	9
0	902	0	10	5	1	12	3	3	7	3
1	0	1057	8	4	2	6	4	6	14	7
2	14	11	826	25	23	5	17	17	25	4
3	7	6	15	900	2	39	2	16	33	11
4	1	3	13	1	893	3	5	7	3	52
5	14	7	7	25	13	814	29	3	26	15
6	8	5	25	1	21	18	875	3	7	4
7	6	10	10	9	9	1	0	893	0	44
8	9	21	8	24	13	42	10	5	822	31
9	7	6	4	6	40	5	0	39	11	885

Multi-class confusion matrix

		Estimate		
		$c_0 \dots c_{k-1}$	c_k	$c_{k+1} \dots c_n$
annotated ground truth	$c_{k+1} \dots c_n$	TN	FP	TN
	c_k	FN	TP	FN
	$c_0 \dots c_{k-1}$	TN	FP	TN

Legend:

- TN: true negative (orange)
- TP: true positive (green)
- FN: false negative (red)
- FP: false positive (brown)

$$tp_i = c_{ii}$$

$$fp_i = \sum_{l=1}^n c_{li} - tp_i$$

$$fn_i = \sum_{l=1}^n c_{il} - tp_i$$

$$tn_i = \sum_{l=1}^n \sum_{k=1}^n c_{lk} - tp_i - fp_i - fn_i$$

- The confusion matrix of a classification with n classes
- When considering the class k ($0 \leq k \leq n$), the four different classification results can be obtained:
 - true positive (green)
 - true negative (orange)
 - false positive (brown)
 - false negative (red)

2. Accuracy paradox

- The accuracy paradox is the paradoxical finding that accuracy is not a good metric for predictive models when classifying in predictive analytics
- This is because a simple model may have a high level of accuracy but be too crude to be useful

2. Accuracy — доля правильних відповідей алгоритму

- Найпростіша метрика: $accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

- Проблеми на незбалансованих вибірках:** є 1075

пацієнтів, 1020 із яких наш класифікатор визначив правильно (True Positive = 20, True Negative = 1000), і 55 неправильно (False Negative = 5, Positive = 50)

$$accuracy = \frac{20 + 1000}{20 + 1000 + 50 + 5} = 94,88\%$$

	$y = 1$	$y = 0$
$a(x) = 1$	20	50
$a(x) = 0$	5	1000

- Але примітивний класифікатор, який вважатиме всіх користувачів здоровими, дасть краще значення цієї метрики!
- Причому він нічого не може передбачити!

	$y = 1$	$y = 0$
$a(x) = 1$	0	0
$a(x) = 0$	25	1050

$$accuracy = \frac{0 + 1050}{0 + 1050 + 0 + 25} = 97,67\%$$

3. Метрики precision (точність) і recall (повнота)

- Precision and recall originate from the field of information retrieval. In information retrieval, we are tasked with providing users with records which are relevant to their search query, and not records which are irrelevant
- Suppose, for example, that we are running a search engine. Our search engine returns 30 pages, of which 20 are relevant and 10 are irrelevant
- Our search engine also fails to return 40 other relevant pages

3. Метрики precision (точність) і recall (повнота)

- Suppose, for example, that we are running a search engine. Our search engine returns 30 pages, of which 20 are relevant and 10 are irrelevant. Our search engine also fails to return 40 other relevant pages
- In this toy example, **precision** is the percentage of results we've returned which are relevant: $20/30$, or $2/3$
- **Recall**, meanwhile, the "completeness" of the information we've returned with respect to all possible relevant results: $20/60$, or $1/3$

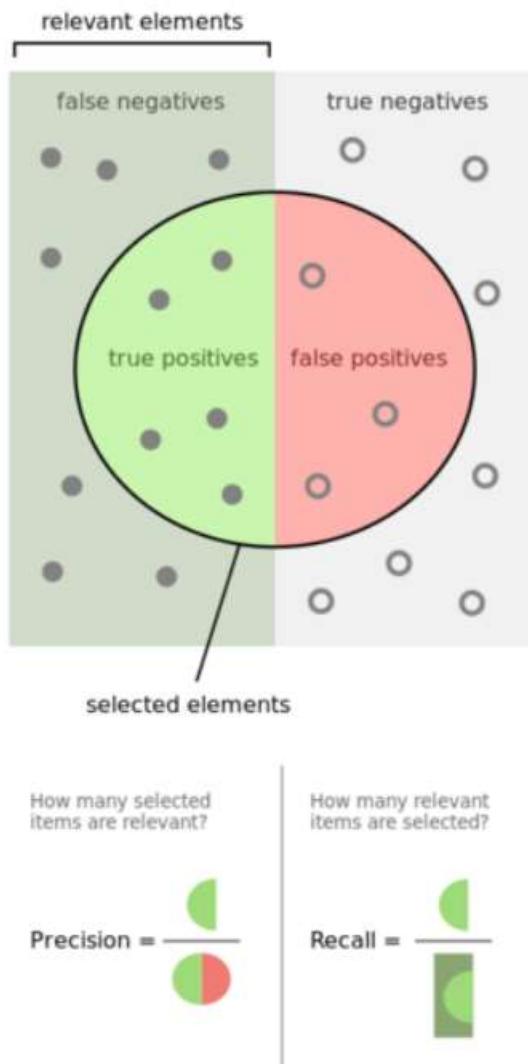
3. Метрики precision (точність) i recall (повнота)

- If you read any literature on search, you will learn that finding the right balance between precision and recall is an important part of search
- **Lightweight** users care less about recall and more about precision, because they are mainly interested in finding a few hits that satisfy their lightweight needs
- **Power** users meanwhile prioritize recall; they can tolerate a few false positives

3. Метрики precision (точність) i recall (повнота)

- At the same time, precision and recall are fundamentally a trade-off
- Returning **all** of the relevant results naturally requires returning at least some edge cases that turn out to be useless
- And returning **only** relevant results naturally requires excluding some edge cases that turn out to be useful

За. Метрика precision (точність)



$$precision = \frac{TP}{TP + FP} = \frac{\text{TP}}{\text{predicted Yes}}$$

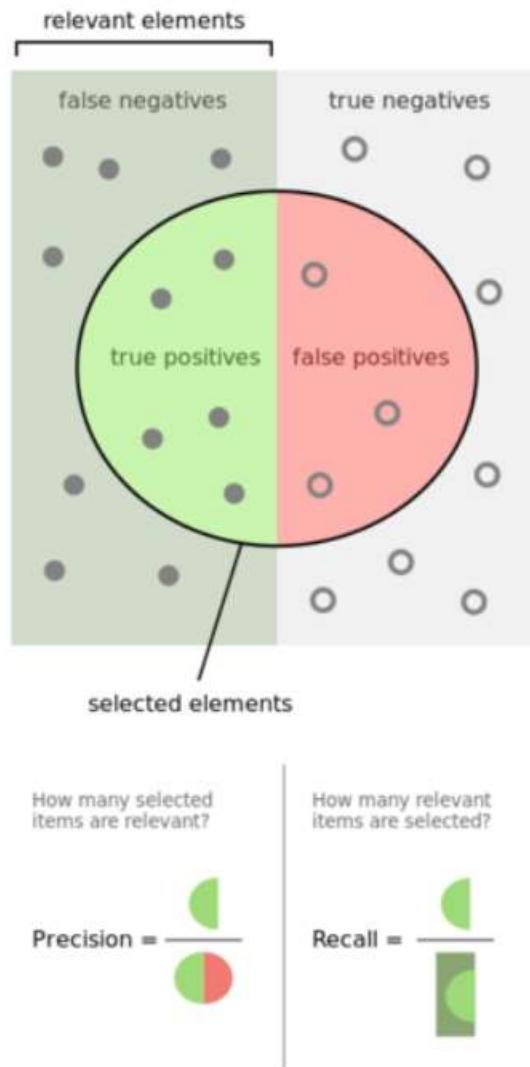
precision (точність) характеризує здатність відрізняти потрібний клас від інших; наскільки можна довіряти класифікатору $precision \in [0; 1]$
(=«доля дійсно хворих серед усіх тих, кого порахували хворими»)

	$y = 1$	$y = 0$
$a(x) = 1$	20	50
$a(x) = 0$	5	1000

measure of exactness or quality

Точність класифікатора: 28,57 %
Точність сталого класифікатора («усі здорові»): 0 %

За. Метрика precision (точність)



$$precision = \frac{TP}{TP + FP} = \frac{\text{TP}}{\text{predicted Yes}}$$

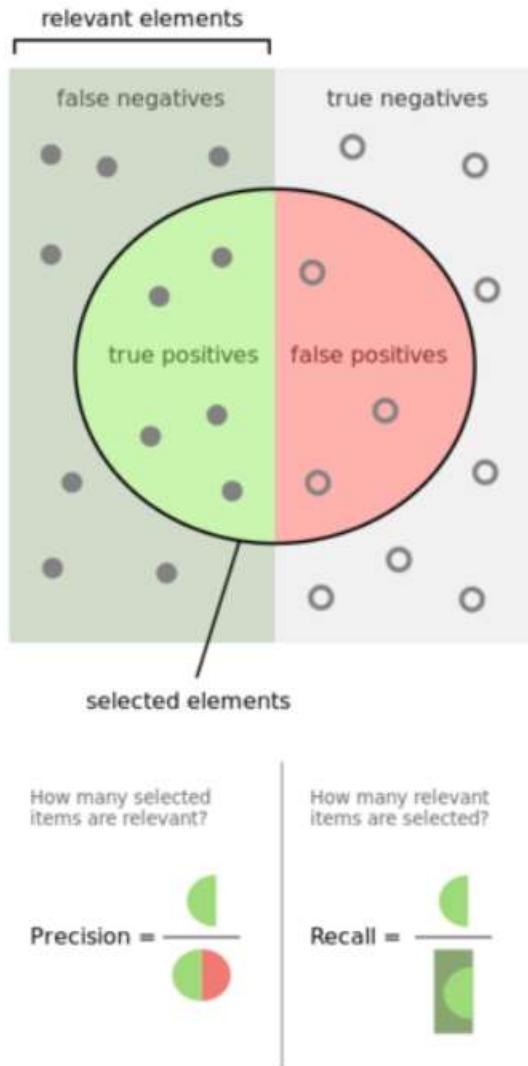
precision (точність) характеризує здатність відрізняти потрібний клас від інших; наскільки можна довіряти класифікатору (=«доля дійсно хворих серед усіх тих, кого порахували хворими»)

А якщо $TP+FP=0$, тобто не має людей, які класифікатор вважає хворими?

Тоді метрика «точність» не визначена!

scikit-learn: when $TP+FP=0$, `fbeta_score` returns 0 and raises `UndefinedMetricWarning`

За. Метрика precision (точність)



$$precision = \frac{TP}{TP + FP} = \frac{\text{TP}}{\text{predicted Yes}}$$

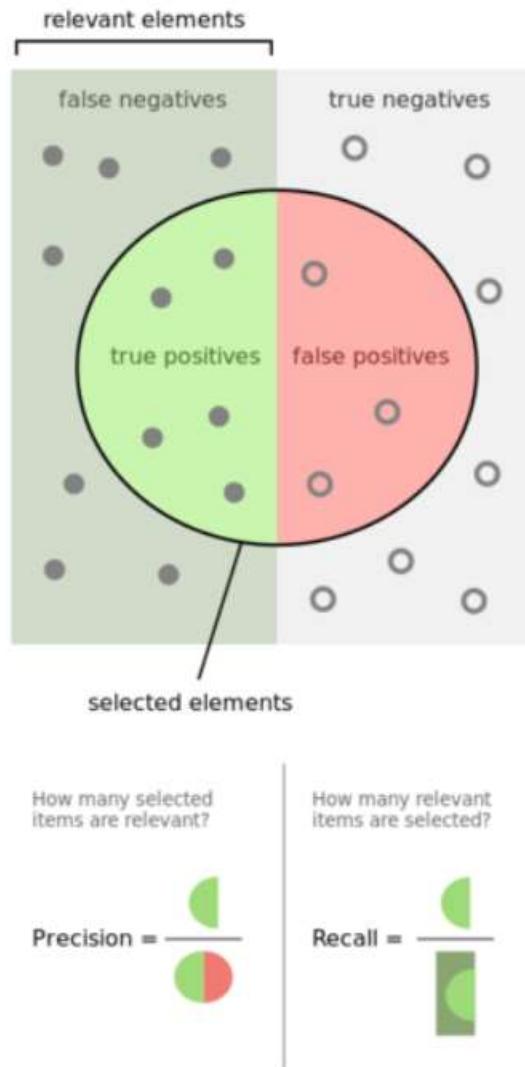
precision (точність) характеризує здатність відрізняти потрібний клас від інших; наскільки можна довіряти класифікатору (=«доля дійсно хворих серед усіх тих, кого порахували хворими»)

In Data Mining, precision is called **confidence (довіра, вірогідність)**

In biomedicine, precision is called **positive predictive value (PPV)**

Precision = Confidence = Positive predictive value

3b. Метрики recall (повнота)



$$recall = \frac{TP}{TP + FN} = \frac{\text{TP}}{\text{actual Yes}}$$

recall (повнота) демонструє здатність алгоритму виявляти потрібний клас взагалі
 $recall \in [0; 1]$
(=«доля правильно виявлених хворих серед усіх хворих»)

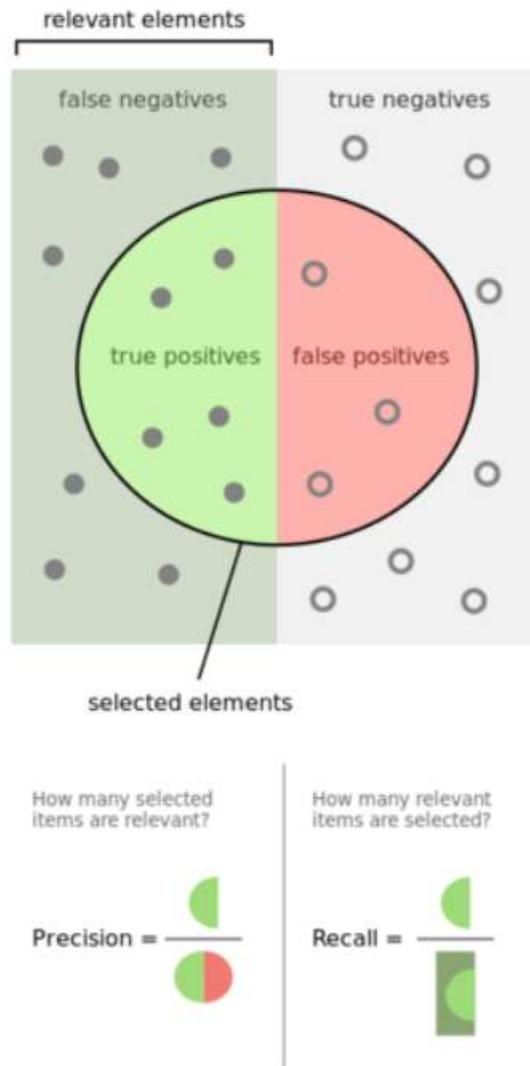
	$y = 1$	$y = 0$
$a(x) = 1$	20	50
$a(x) = 0$	5	1000

measure of completeness or quantity

Повнота класифікатора: 80 %

Повнота сталого класифікатора («усі здорові»): 0 %

3b. Метрики recall (повнота)



$$recall = \frac{TP}{TP + FN} = \frac{\text{TP}}{\text{actual Yes}}$$

recall (повнота) демонструє здатність алгоритму виявляти потрібний клас взагалі (=«доля правильно виявлених хворих серед усіх хворих»)

А якщо $\text{TP} + \text{FN} = 0$, тобто хворих людей не має?

Тоді метрика «повнота» не визначена!

In Psychology, recall is called **sensitivity (чутливість)**

Recall = Sensitivity = True Positive Rate = Hit Rate

Якому класу надати категорію 1, а якому – 0?

- **Кого шукати: хворих чи здорових?**

Якому класу надати категорію 1, а якому – 0?

- **Кого шукати: хворих чи здорових?**
- З точки зору алгоритмів машинного навчання – немає ніякої різниці, але з точки зору більшості метрик (точність, повнота тощо) різниця є
- Тому основна рекомендація – щоб була вірна інтерпретація. Зазвичай, це збігається з тим, що «найменший клас» позначають за 1
- Наприклад, рідкісну хворобу. Тоді повнота – який відсоток хворих ми знайшли, точність – який відсоток із знайдених є дійсно хворими

3. Яка характеристика матриці помилок відсутня в precision i recall?

$$\square \quad \text{precision} = \frac{TP}{TP + FP} \quad \text{recall} = \frac{TP}{TP + FN}$$

- А чому відсутня TN?
- In information retrieval only the positive class is of relevance, while number of negatives, in general, is large and unknown

Van Rijsbergen, C. J. (1979) Information Retrieval (2nd ed.)

3. Multi-class precision-recall

- Precision and recall scores can be defined in the multi-class setting
- The metrics can be "averaged" across all the classes in many possible ways (as in scikit-learn):
 - ❖ **None**: return the precision and recall scores for each class
 - ❖ **Another options??**

3. Multi-class precision-recall

- Precision and recall scores can be defined in the multi-class setting
- The metrics can be "averaged" across all the classes in many possible ways (as in scikit-learn):
 - ❖ **None**: return the precision and recall scores for each class
 - ❖ Calculate metrics for each class independently, and find their unweighted mean (**'macro'**) or their average weighted (**'weighted'**) by support (the number of true instances for each class)
 - ❖ **'micro'**: calculate metrics globally by counting the total TPs, FPs and FNs

3. Multi-class precision-recall

Recall (macro average)

predictions →

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

TP: 100, FN: 100 R(A) = 100 / 200
TP: 9, FN: 1 R(B) = 9/10
TP: 8, FN: 2 R(C) = 8/10
TP: 9, FN: 1 R(D) = 9/10

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

$$\text{macro average recall} = \frac{R(A) + R(B) + R(C) + R(D)}{4} = 0.775$$

the number of classes

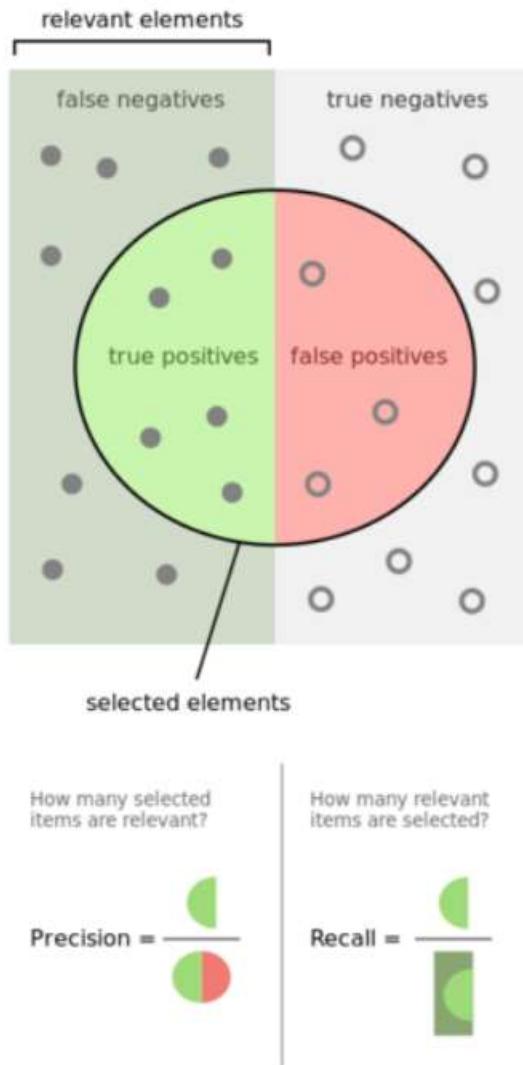
$$\text{weighted average recall} = \frac{100R(A)}{126} + \frac{9R(B)}{126} + \frac{8R(C)}{126} + \frac{9R(D)}{126} = 0.576$$

$$\text{micro average recall} = \frac{126}{126 + 104} = 0.548$$

3. Multi-class precision-recall

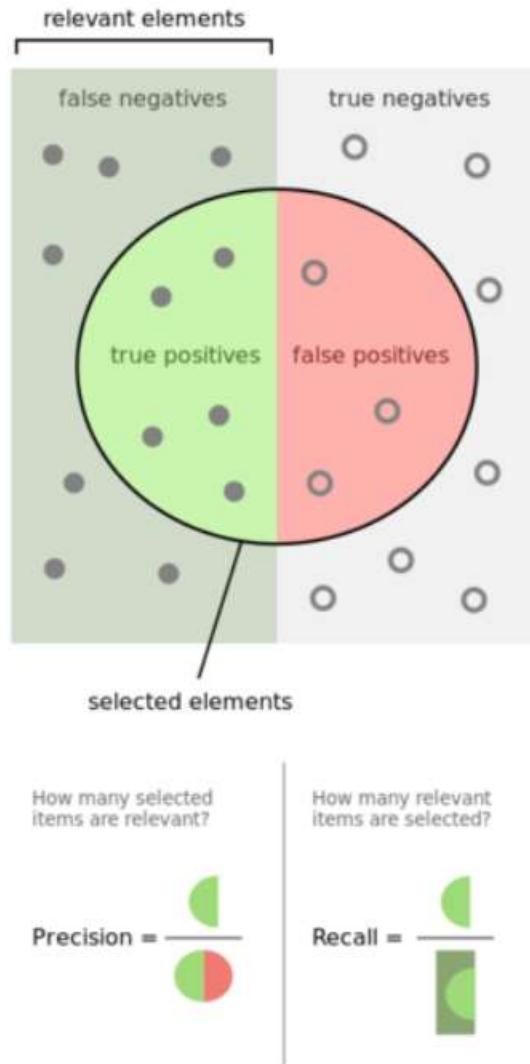
- ❖ Calculate metrics for each class independently, and find their unweighted mean (**'macro'**) by support (the number of true instances for each class)
- ❖ **'micro'**: calculate metrics globally by counting the total TPs, FPs and FNs
- ❖ Якщо класи відрізняються по потужності, то при мікро-усередненні малі за кількістю класи практично ніяк не впливатимуть на результат, оскільки їх внесок в середні TP, FP, FN і TN буде незначний
- ❖ У випадку ж із макро-усередненням кожен клас внесе рівний внесок до підсумкової метрики

3. Метрики precision (точність) і recall (повнота)



- Точність та повнота характеризують різні сторони якості класифікатора
- Чим вищі точність, тим менше **хібних спрацьовувань** $\frac{TP}{TP + FP}$
- Чим вищі повнота, тим менше **хібних пропусків** $\frac{TP}{TP + FN}$
- Пріоритет у бік точності чи повноти вибирається в залежності від задачі

А чи можна одночасно застосовувати і точність, і повноту?



□ Тобто як усереднити ці дві якісні метрики?

Comparing Systems

System 1

- Precision: 70%
- Recall: 60%



System 2

- Precision: 80%
- Recall: 50%

Усереднення точності та повноти. Спроба №1

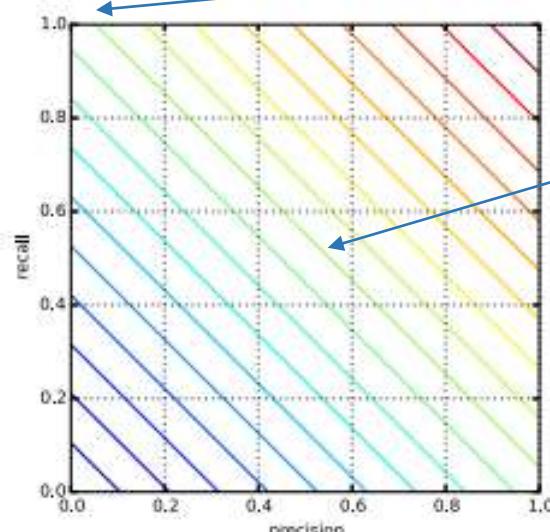
- Як **арифметичне середнє**: $A = \frac{1}{2} (\text{precision} + \text{recall})$ або як запропонував Harold Borko: $\text{precision} + \text{recall} - 1$

- (абстрактний) приклад №1:

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

сталий класифікатор (завжди клас 1): якщо $\text{precision}=0,05$, $\text{recall}=1$, то $A=0,525$

непоганий класифікатор: якщо $\text{precision}=0,525$, $\text{recall}=0,525$, то $A=0,525$



- **Лінії рівня:** $\text{Const} = \frac{1}{2} (\text{precision} + \text{recall})$,
- звідки $\text{recall} = 2 * \text{Const} - \text{precision}$
- Чим червоніша лінія рівня, тим більша Const

Усереднення точності та повноти. Спроба №2

□ Як **мінімум**: $M = \min(\text{precision}, \text{recall})$

□ (абстрактний) приклад №1:

сталий класифікатор: якщо $\text{precision}=0,05$, $\text{recall}=1$, то $M=0,05$

непоганий класифікатор: якщо $\text{precision}=0,525$, $\text{recall}=0,525$, то $M=0,525$

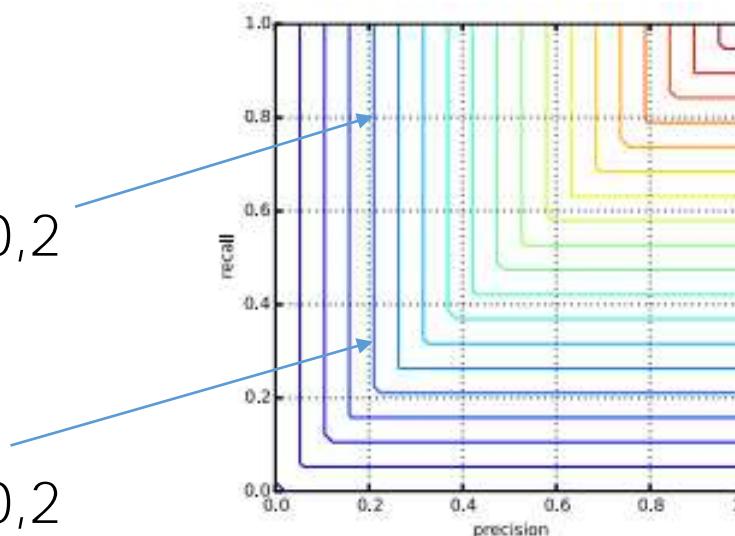
□ (абстрактний) приклад №2:

перший класифікатор:

якщо $\text{precision}=0,2$, $\text{recall}=0,8$, то $M=0,2$

другий класифікатор:

якщо $\text{precision}=0,2$, $\text{recall}=0,3$, то $M=0,2$



□ **Лінії рівня**

Усереднення точності та повноти.

Спроба №3. Дві задачки із фізики

- If the vehicle travels for a certain amount of time t at a speed x (e.g. 60 km/h) and then the same amount of time t at a speed y (e.g. 20 km/h), then its average speed is ...

$$v_{avr} = \frac{\text{total distance traveled}}{\text{total time taken}} = \frac{xt+yt}{t+t} = \frac{x+y}{2}$$

- ... the **arithmetic mean** of x and y , which is 40 km/h

- If a vehicle travels a certain distance d outbound at a speed x (e.g. 60 km/h) and returns the same distance at a speed y (e.g. 20 km/h), then its average speed is ...

$$v_{avr} = \frac{\text{total distance traveled}}{\text{total time taken}} = \frac{2d}{\frac{d}{x} + \frac{d}{y}} = \frac{2}{\frac{1}{x} + \frac{1}{y}}$$

- ... the **harmonic mean** of x and y , which is 30 km/h

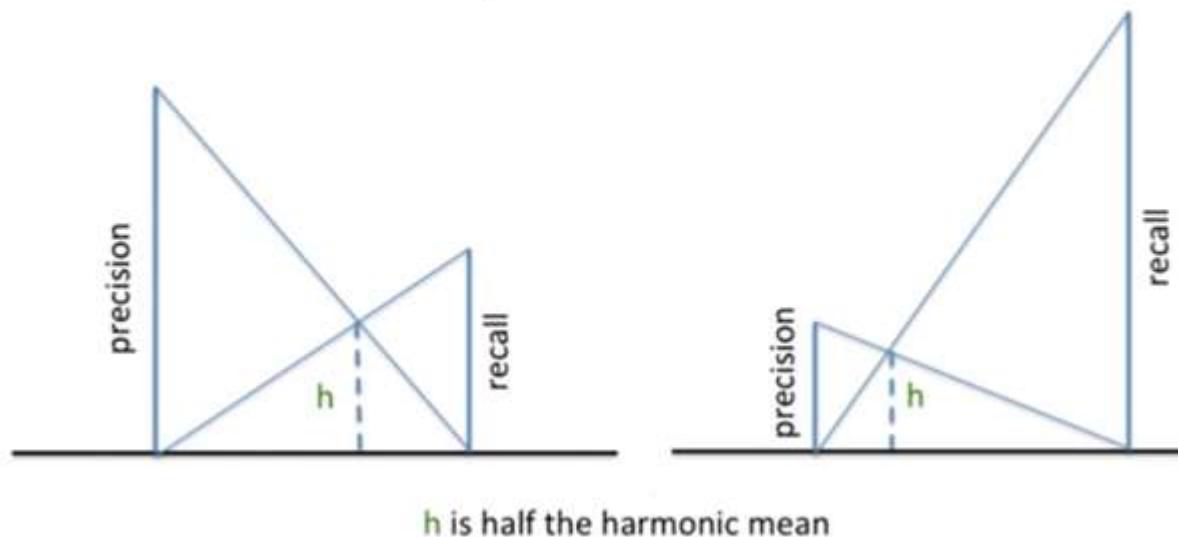
Усереднення точності та повноти. Спроба №3 (4. F-scores)

□ Як гармонічне середнє: $F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \\ = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^n x_i^{-1}}{n} \right)^{-1}$$

тобто гармонічне середнє є оберненою величиною середнього арифметичного обернених величин

Harmonic Mean punishes extreme value more



The harmonic mean H is a Schur-concave function (вогнута функція Шура), and dominated by the minimum of its arguments, in the sense that for any positive set of arguments,

$\min(x_1, \dots, x_n) \leq H(x_1, \dots, x_n) \leq n * \min(x_1, \dots, x_n)$. Thus, the harmonic mean cannot be made arbitrarily large by changing some values to bigger ones (if at least one unchanged)

Усереднення точності та повноти. Спроба №3 (4. F-scores)

□ Як гармонічне середнє: $F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$

□ (абстрактний) приклад №1:

сталий класифікатор: якщо $precision=0,05$, $recall=1$, то $F_1 \approx 0,095$

непоганий класифікатор: якщо $precision=0,525$, $recall=0,525$, то $F_1=0,525$

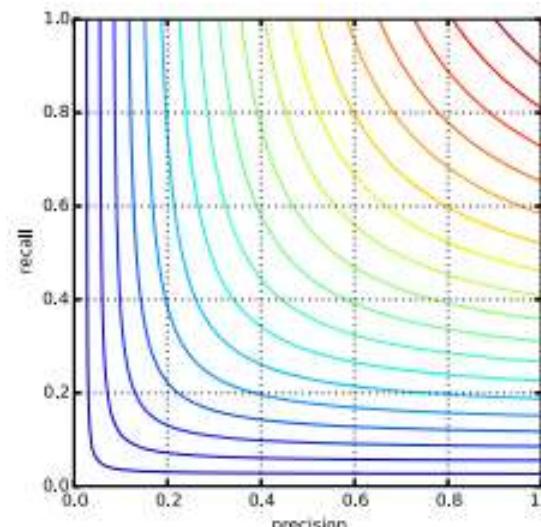
□ (абстрактний) приклад №2:

перший класифікатор:

якщо $precision=0,2$, $recall=0,8$, то $F_1=0,32$

другий класифікатор:

якщо $precision=0,2$, $recall=0,3$, то $F_1=0,24$

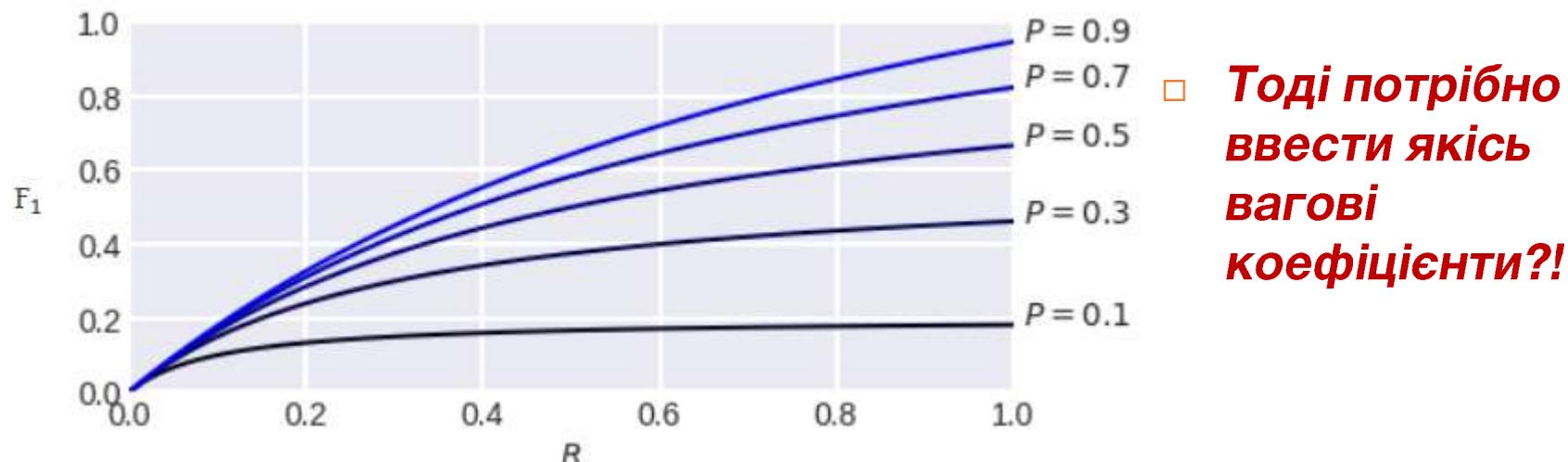


□ **Лінії рівня**
□ $recall = \frac{Const * precision}{2 * precision - Const}$

Усереднення точності та повноти.

Спроба №3 (4. F-scores)

- Як гармонічне середнє: $F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$
- Якщо, наприклад, точність дуже мала, то збільшення повноти, нехай і в два рази, не сильно змінює значення метрики F_1
- Наприклад, при точності 10% F_1 -міра не може бути більше 20%:



- Графік залежності F_1 -міри від повноти при фіксованій точності

□ **Тоді потрібно ввести якісь вагові коефіцієнти?!**

4. F-scores

□ F-міра: $F_{\beta} = (1 + \beta^2) \frac{precision * recall}{\beta^2 * precision + recall}$

Nancy Chinchor, MUC-4 Evaluation Metrics, in Proc.of the Fourth Message Understanding Conference, pp. 22-29, 1992

<http://www.aclweb.org/anthology-new/M/M92/M92-1002.pdf>

- β визначає вагу точності в усередненій (агрегованій) метриці
- На практиці, зазвичай, беруть $\beta=0,5$ або $\beta=2$, коли хочуть надати перевагу одній із складових

4. F-scores

- F-міра: $F_{\beta} = (1 + \beta^2) \frac{precision*recall}{\beta^2*precision+recall}$
- β визначає вагу точності в усередненій (агрегованій) метриці
- **Коли F-міра (F-score) досягає свого максимуму і мінімуму? Як це знайти?**

4. F-scores

- F-міра: $F_\beta = (1 + \beta^2) \frac{precision*recall}{\beta^2*precision+recall}$
- **Коли F-міра (F-score) досягає свого максимуму і мінімуму? Як це знайти?**
 - $\frac{\partial F_\beta}{\partial p} = (1 + \beta^2) \frac{r*(\beta^2*p+r)-\beta^2*p*r}{(\beta^2*p+r)^2} = (1 + \beta^2) \frac{r^2}{(\beta^2*p+r)^2}$
 - $\frac{\partial F_\beta}{\partial r} = (1 + \beta^2) \frac{p*(\beta^2*p+r)-p*r}{(\beta^2*p+r)^2} = (1 + \beta^2) \frac{\beta^2*p^2}{(\beta^2*p+r)^2}$
 - $\frac{\partial F_\beta}{\partial p} = 0 \rightarrow recall=0 \quad \text{и} \quad \frac{\partial F_\beta}{\partial r} = 0 \rightarrow precision=0$

4. F-scores

- Стационарна точка ($precision=0$, $recall=0$) є точкою, підозрілою на екстремум
- $$F_{\beta} = \frac{1 + 1/\beta^2}{\frac{1}{\beta^2 * precision} + \frac{1}{recall}}$$
- F_{β} ($precision=0$, $recall=0$) = 0
- Оскільки F-міра, очевидно, приймає тільки невід'ємні значення, то точка ($precision=0$, $recall=0$) є точкою мінімуму
- **А чи є максимум?**

4. F-scores

- **А чи є максимум функції** $F_\beta = \frac{\frac{1+1/\beta^2}{1}}{\beta^2 * precision + \frac{1}{recall}}$?
- Обидві змінні лежать в діапазоні [0;1], тому максимум повинен бути
- І якщо екстремуму не має всередині області, то він лежить на її межі
- Коли precision і recall будуть збільшуватися від 0 до 1, то знаменник F_β зменшується, а саме F_β збільшується, причому максимум буде досягатися, коли змінні дорівнююватимуть 1
- Тоді максимум F_β дорівнює 1

4. F-scores

- F-міра: $F_{\beta} = (1 + \beta^2) \frac{precision*recall}{\beta^2*precision+recall}$
- F-міра (F-score) досягає свого максимуму, коли точність і повнота дорівнюють одиниці,
- і мінімальна (=близька до нуля), якщо один із аргументів теж близький до нуля
- β визначає вагу точності в усередненій (агрегованій) метриці

4. F-scores

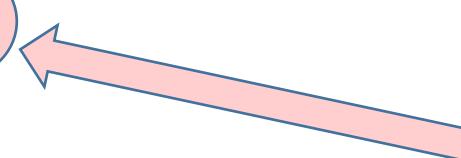
- Як F-міра: $F_{\beta} = (1 + \beta^2) \frac{precision*recall}{\beta^2*precision+recall}$
- β визначає вагу точності в усередненій (агрегованій) метриці
- **Найбільш поширені значення коеф-ту β :**
 $\frac{1}{2}$; 1; 2
- **Коли який брати?**

4. F-scores

- Як F-міра: $F_{\beta} = (1 + \beta^2) \frac{precision*recall}{\beta^2*precision+recall}$
- β визначає вагу точності в усередненій метриці
- **Найбільш поширені значення коеф-ту β :** $\frac{1}{2}$; 1; 2
- **Коли який брати?**
- **Завдання:** хочемо **зменшити** складські витрати, прогнозуючи, коли товар, що швидко псується, закінчиться у магазині
- **Які помилки потрібно зменшити більше: хибні спрацьовування чи хибні пропуски?**

4. F-scores

- Завдання: хочемо зменшити складські витрати, прогнозуючи, коли товар, що швидко псується, закінчиться у магазині
- Які помилки потрібно зменшити більше: **хибні спрацьовування** чи хибні пропуски?

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$


- Тобто повинна збільшитися точність, а для цього потрібно взяти $\beta=1/2$ чи $\beta=2$? Як це визначити?

4. F-scores

- Пояснення залежності F score від значення коефіцієнту β :

$$\begin{aligned} \square F_{\beta} &= (1 + \beta^2) \frac{precision * recall}{\beta^2 * precision + recall} = \\ &= (1 + \beta^2) precision * recall * \frac{1}{\beta^2 * precision + recall} \end{aligned}$$

- Таким чином, щоб збільшити F_{β} , потрібно зменшити $\beta^2 * precision + recall$
- Отже, якщо $\beta^2 < 1$, то вплив *precision* буде більшим, інакше – більше впливатиме *recall*

4. F-scores

- Завдання: хочемо мінімізувати складські витрати, прогнозуючи, коли товар, що швидко псується, закінчиться у магазині
- Які помилки потрібно зменшити більше: **хибні спрацьовування** чи хибні пропуски?

$$precision = \frac{TP}{TP + \cancel{FP}}$$
$$recall = \frac{TP}{TP + FN}$$

- Тобто повинна збільшитися точність, а для цього потрібно взяти $\beta=1/2$ чи $\beta=2$?

4. F-scores

- Найбільш поширені значення коеф-ту β : $\frac{1}{2}$; 1; 2
- Завдання: служба безпеки аеропорту хоче мінімізувати пропуск терористів на борт літака
- **Які помилки потрібно зменшити більше: хибні спрацьовування чи хибні пропуски?**

4. F-scores

- Найбільш поширені значення коеф-ту β : $\frac{1}{2}$; 1; 2
- Завдання: служба безпеки аеропорту хоче мінімізувати пропуск терористів на борт літака
- Які помилки потрібно зменшити більше: хибні спрацьовування чи **хибні пропуски**?

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

$$recall = \frac{TP}{TP + \textcolor{red}{FN}}$$

- Тобто повинна збільшитися повнота
- Як це вплине на точність? Чому будемо брати $\beta=2$?

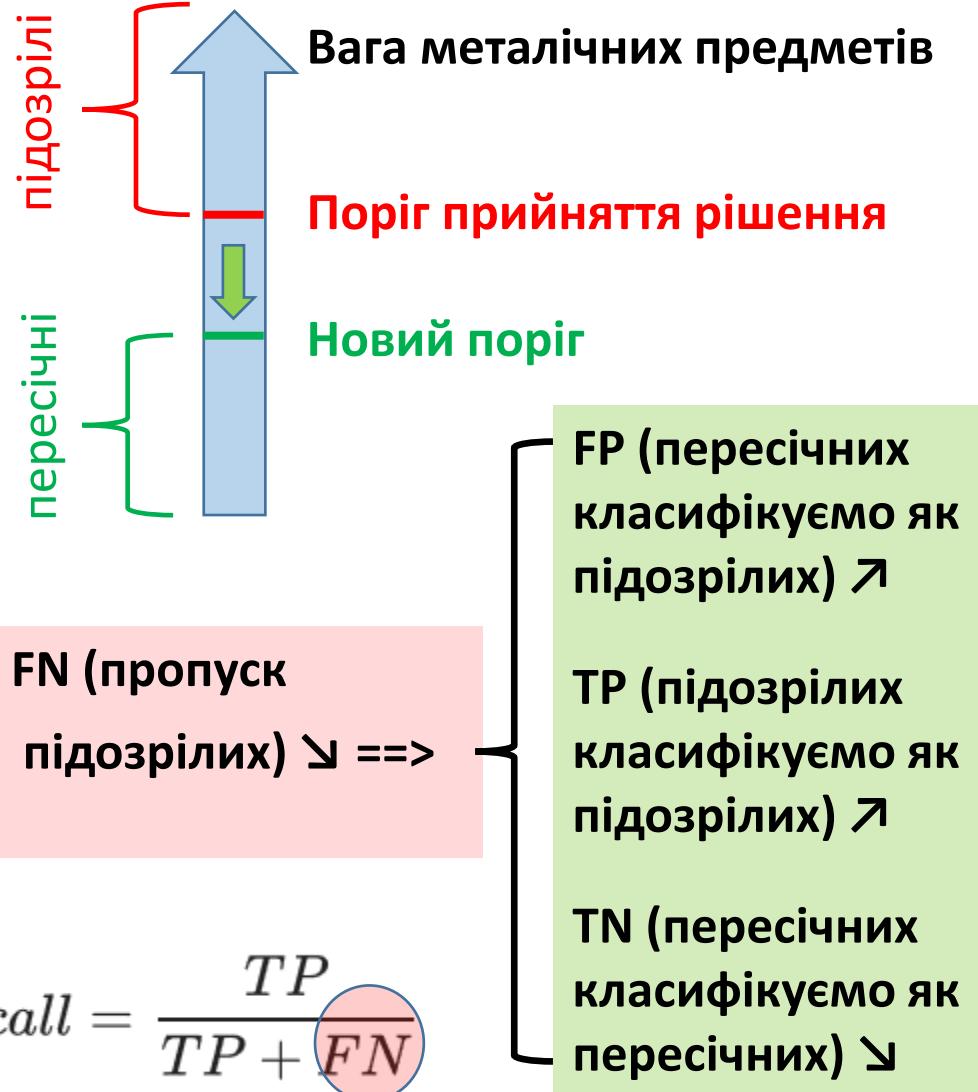
4. F-scores



- Завдання: служба безпеки аеропорту хоче мінімізувати пропуск терористів на борт літака, тобто потрібно зменшити **хибні пропуски**

$$precision = \frac{TP}{TP + \text{FP}}$$

$$recall = \frac{TP}{TP + \text{FN}}$$



4. F-scores

1. FN (пропуск підозрілих) ↘



2. FP (пересічних класифікуємо як підозрілих) ↗

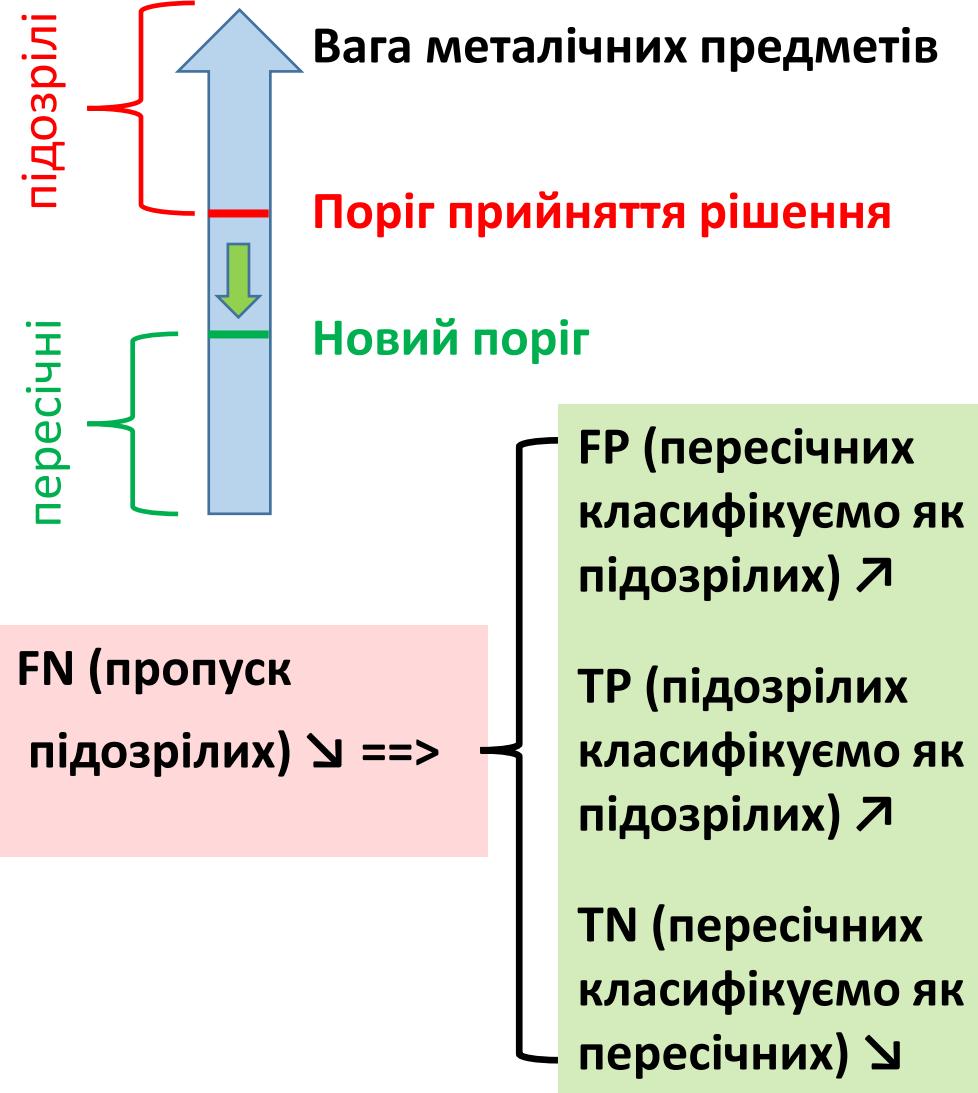


3. Precision ↘

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



4. Для зменшення ваги precision потрібно збільшити коефіцієнт β , наприклад, $\beta=2$



4. F-scores

□ **F-міра:** $F_{\beta} = (1 + \beta^2) \frac{precision*recall}{\beta^2*precision+recall}$

Nancy Chinchor, MUC-4 Evaluation Metrics, in Proc.of the Fourth Message Understanding Conference, pp. 22-29,1992

<http://www.aclweb.org/anthology-new/M/M92/M92-1002.pdf>

- β визначає вагу точності в усередненій (агрегованій) метриці
- На практиці, зазвичай, беруть $\beta=0,5$ або $\beta=2$, коли хочуть надати перевагу одній із складових

4. F-scores. Трішки історії

В основі F-measure лежить функція *E* (effectiveness), запропонована (1975) в книзі "Пошук інформації" (Information Retrieval) ван Рійсбергена (Cornelis Joost van Rijsbergen), професора інформатики та керівника групи пошуку інформації в Університеті Глазго



Nancy Chinchor першою (в 1992) запропонувала застосовувати F-measure як метрику для оцінки технологій видобутку інформації (розуміння повідомлень)

4. F-scores. E (effectiveness) function

ван Рійсберген (van Rijsbergen)

$$E = 1 - \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}, \text{ where } \alpha = \frac{1}{\beta^2 + 1}.$$

Let's remove α using β .

$$\begin{aligned} E &= 1 - \frac{1}{\frac{1}{\beta^2 + 1} \frac{1}{P} + \left(1 - \frac{1}{\beta^2 + 1}\right) \frac{1}{R}} = \\ &= 1 - \frac{(\beta^2 + 1)PR}{R + \beta^2 P}. \end{aligned}$$

Now you see that $E = 1 - F_\beta$.

- Note that F rises if R or P gets better whereas E becomes small if R or P improves
- This seems the reason why F is more commonly used than E

4. F-score. Why β^2

Still, some of you are not sure why β^2 is used instead of β in $\alpha = \frac{1}{\beta^2+1}$.

β is the parameter that controls the weighting between P and R. Formally, β is defined as follows:

$$\beta = R/P, \quad \text{where} \quad \frac{\partial E}{\partial P} = \frac{\partial E}{\partial R}.$$

The motivation behind this condition is that at the point where the gradients of E w.r.t. P and R are equal, the ratio of R against P should be a desired ratio β .

$$\begin{aligned} \frac{\partial E}{\partial P} &= -\frac{R(\alpha R + (1 - \alpha)P) - PR(1 - \alpha)}{g^2}. & \text{Then, } \frac{\partial E}{\partial P} = \frac{\partial E}{\partial R} \text{ is equivalent to:} \\ \frac{\partial E}{\partial R} &= -\frac{P(\alpha R + (1 - \alpha)P) - PR\alpha}{g^2}. & R(\alpha R + (1 - \alpha)P) - PR(1 - \alpha) = \\ & & = P(\alpha R + (1 - \alpha)P) - PR\alpha. \end{aligned}$$

which can be simplified to: $\alpha R^2 = (1 - \alpha)P^2$.

As $\beta = R/P$, we can replace R with βP .²

$$\alpha \beta^2 P^2 = (1 - \alpha)P^2 \Rightarrow \alpha = \frac{1}{\beta^2 + 1}$$

Yutaka Sasaki. The truth of the F-measure (2007)

4. F-score. It's name

- There is one thing that remains unsolved, which is why the F-measure is called F. A personal communication with David D. Lewis several years ago revealed that when the F-measure was introduced to MUC-4, the name was accidentally selected by the consequence of regarding a different F function in van Rijsbergen's book as the definition of the “F-measure”

Yutaka Sasaki. The truth of the F-measure (2007)

4. Multi-class F-scores

- F-scores can be defined in the multi-class setting
- The metrics can be "averaged" across all the classes in many possible ways (as in scikit-learn):

None: return the F-score for each class

Calculate metrics for each class independently, and find their unweighted mean ('macro') or their average weighted ('weighted') by support (the number of true instances for each class)

'micro': calculate metrics globally by counting the total TPs, FPs and FNs

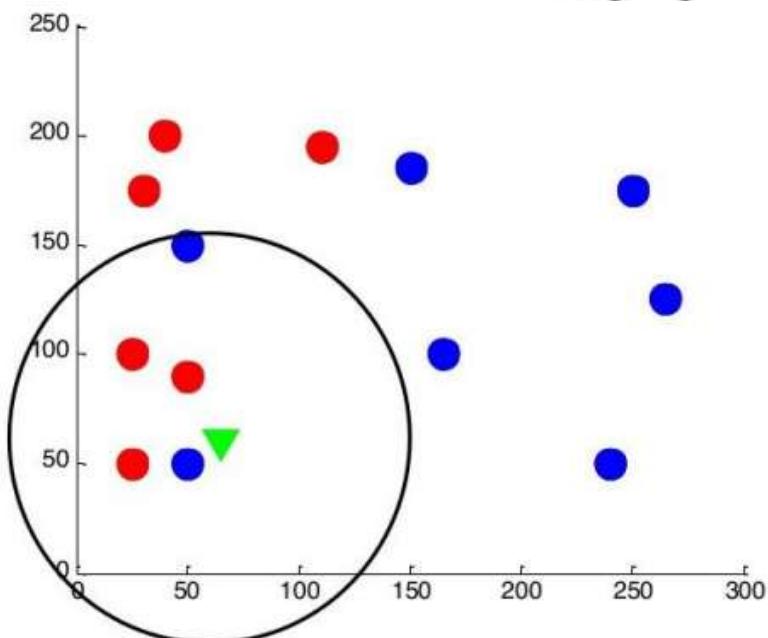
Two types of classification algorithms

- In classification problems, we use two types of algorithms (dependent on the kind of output it creates):
- **Class output:** Algorithms like SVM and KNN create a class output. For instance, in a binary classification problem, the outputs will be either 0 or 1
- **Probability output:** Algorithms like Logistic Regression, Random Forest, Gradient Boosting, Adaboost etc. give probability outputs. Converting probability outputs to class output is just a matter of creating a threshold probability
- In regression problems, we do not have such inconsistencies in output. The output is always continuous in nature and requires no further treatment

Оцінка належності та класифікатор у методі k найближчих сусідів

- Класифікатор для методу k найближчих сусідів:

$$a(x) = \left[\sum_{i=1}^k [y^{(i)} = 1] > k/2 \right]$$



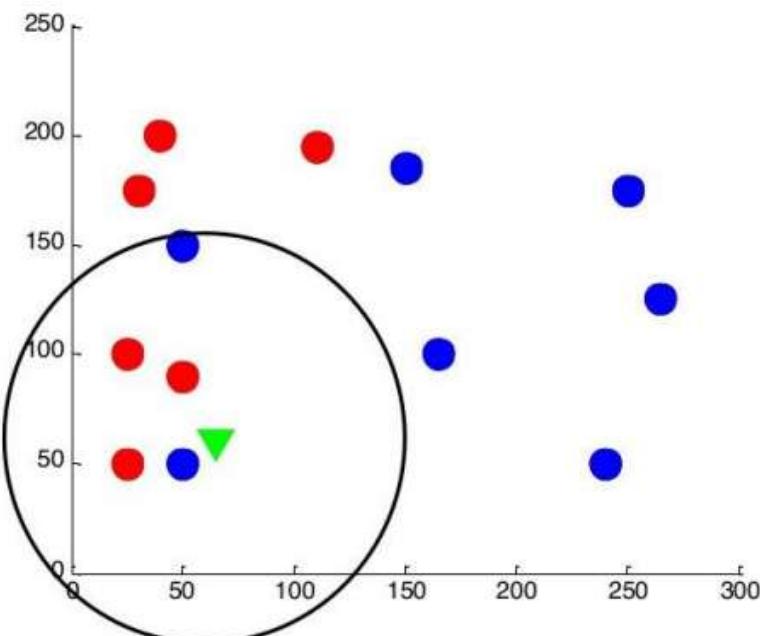
□ Давайте зрозуміємо – звідки може братися на виході ймовірність приналежності об'єкту до певного класу?

<https://habrahabr.ru/company/yandex/blog/206058/>

Оцінка належності та класифікатор у методі k найближчих сусідів

- Класифікатор для методу k найближчих сусідів:

$$a(x) = \left[\sum_{i=1}^k [y^{(i)} = 1] > k/2 \right]$$



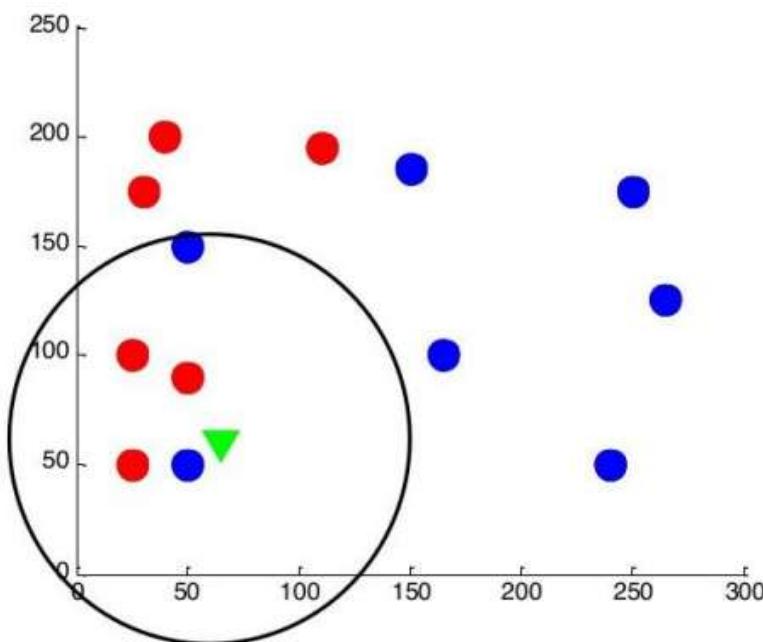
- Якщо серед сусідів об'єкту всі відносяться до одного класу, то він упевнений у класифікації та дає високу оцінку,
- Інакше – оцінка знижується

<https://habrahabr.ru/company/yandex/blog/206058/>

Оцінка належності та класифікатор у методі k найближчих сусідів

- Класифікатор для методу k найближчих сусідів:

$$a(x) = \left[\sum_{i=1}^k [y^{(i)} = 1] > k/2 \right]$$

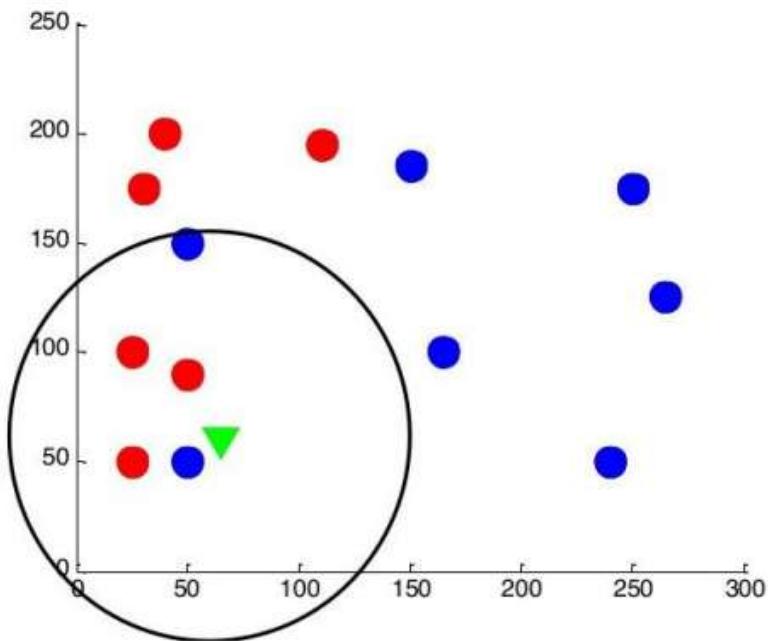


□ А бувають задачі, де поріг потрібно взяти менше половини?

Оцінка належності та класифікатор у методі k найближчих сусідів

- Класифікатор для методу k найближчих сусідів:

$$a(x) = \left[\sum_{i=1}^k [y^{(i)} = 1] > k/2 \right]$$



- А якщо ці класи нерівноцінні?
- Наприклад, щось із цих кружечків – золоті піщинки, а щось – звичайний пісок?
- Тоді поріг спрацьовування, мабуть, потрібно зменшувати!

<https://habrahabr.ru/company/yandex/blog/206058/>

Two types of classification algorithms

- It can be more **flexible** to predict probabilities of an observation belonging to each class in a classification problem rather than predicting classes directly
- This **flexibility** comes from the way that probabilities may be interpreted using different **thresholds** that allow the operator of the model to trade-off concerns in the errors made by the model, such as the number of false positives compared to the number of false negatives
- This is required when using models where the cost of one error outweighs the cost of other types of errors

Постановка задачі

- **Дано: задача класифікації**

$X^\ell = \{x_1, \dots, x_\ell\}$ — вибірка

$y_i = y(x_i) \in \{0, 1\}$, $i = 1, \dots, \ell$ — відомі бінарні відповіді

Попередня задача	Нова задача
$a: X \rightarrow Y$ — алгоритм, розв'язувальна функція, що наближує y на всій множині об'єктів X	$b: X \rightarrow \mathbb{R}$ — алгоритм, розв'язувальна функція, що оцінює належність x до класу 1
Питання: як виміряти якість класифікатора $a(x)$ на вибірці X^ℓ?	Питання: як виміряти якість оцінки $b(x)$ на вибірці X^ℓ?

Співпраця оцінки належності та класифікатора через порогову ймовірність

- Зазвичай, класифікатор має вигляд:

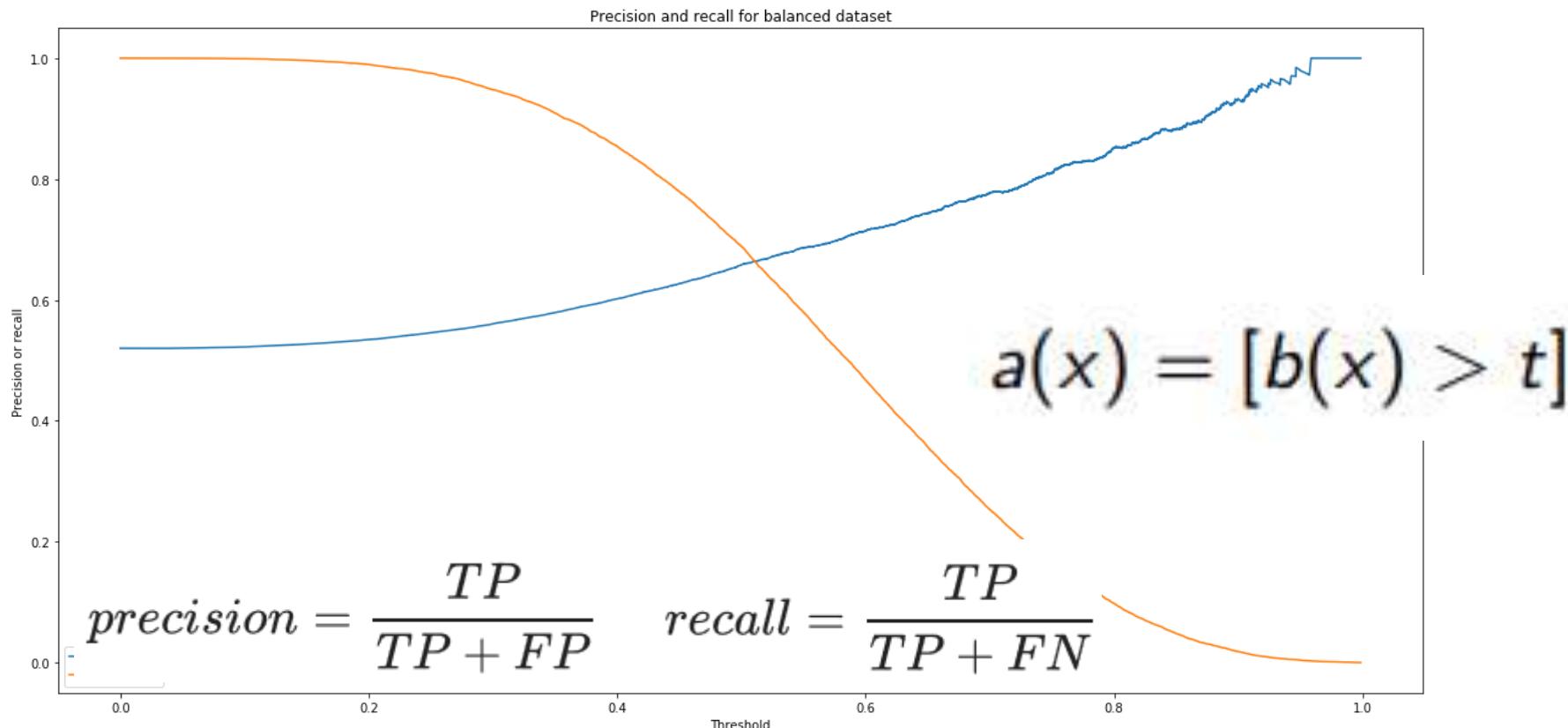
$$a(x) = [b(x) > t]$$

де

- $b(x)$ — оцінка належності x до класу 1
- t — поріг (threshold) класифікації (=порогова ймовірність = точка відсічення, cut-off value)

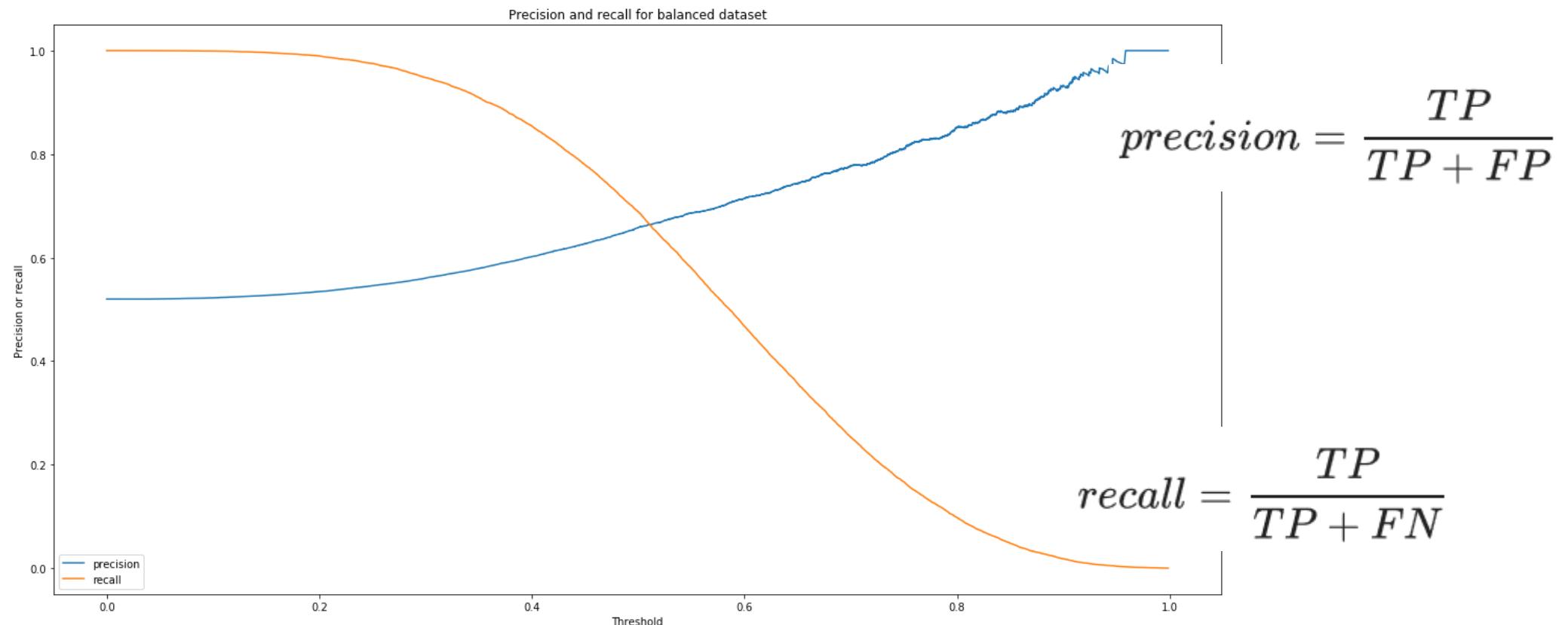
Співпраця оцінки належності та класифікатора через порогову ймовірність

- Зазвичай, зі збільшенням порогу t ??????? буде зростати, а ??????? – падати



Співпраця оцінки належності та класифікатора через порогову ймовірність

- Зазвичай, зі збільшенням порогу т **точність** буде зростати, а **повнота** – падати



4. F-scores

1. FN (пропуск підозрілих) ↗

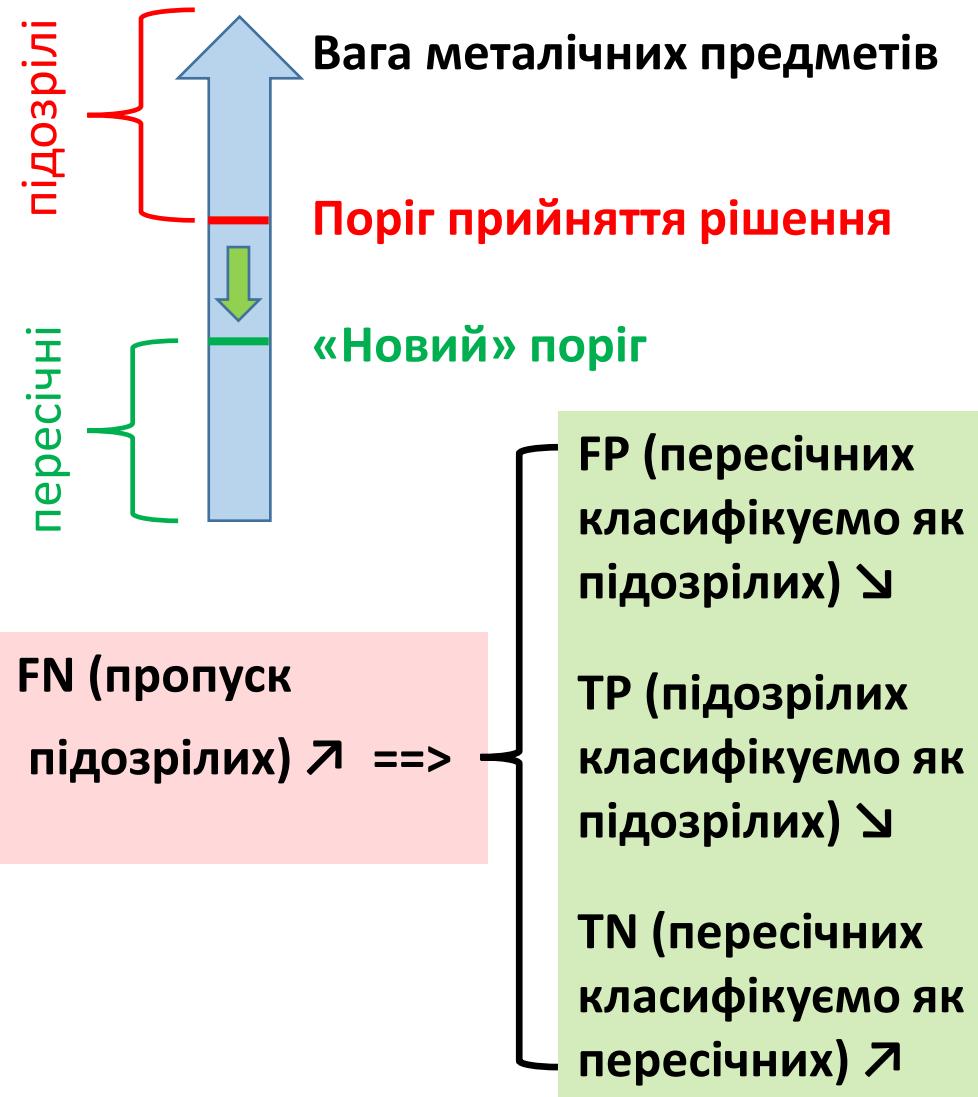


2. FP (пересічних класифікуємо як підозрілих) ↘



3. Precision ↗

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



Яка негативна риса об'єднує метрики precision, recall і F-scores?

- $precision = \frac{TP}{TP + FP}$
- $recall = \frac{TP}{TP + FN}$
- $F_\beta = (1 + \beta^2) \frac{precision*recall}{(\beta^2*precision)+recall}$
- Всі ці три метрики не враховують TN, тому можуть давати зміщену (biased) оцінку

5. Matthews Correlation Coefficient (MCC)

Коефіцієнт кореляції Метьюса

- Unlike the other metrics discussed above, MCC takes all the cells of the confusion matrix into consideration in its formula

$$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

- Similar to Correlation Coefficient, the range of values of MCC lie **between -1 to +1**. A model with a score of +1 is a perfect model and -1 is a poor model. This property is one of the key usefulness of MCC as it leads to easy interpretability
- The Matthews correlation coefficient has been introduced by Brian Matthews to evaluate the predicted structure of an enzyme, in a biochemical study in 1975:

B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta - Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975

5. Matthews Correlation Coefficient (MCC)

Коефіцієнт кореляції Метьюса

- Unlike the other metrics discussed above, MCC takes all the cells of the confusion matrix into consideration in its formula

$$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

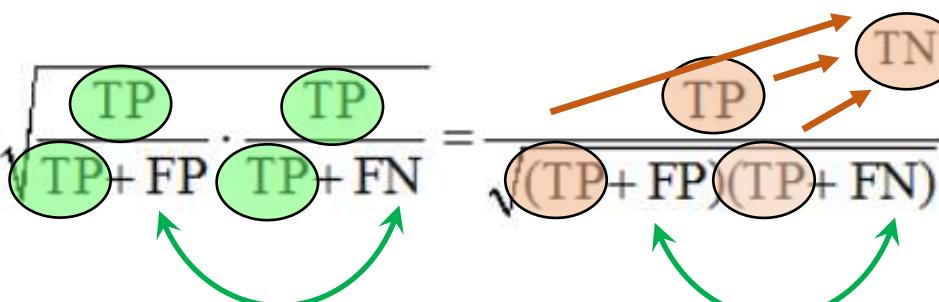
- The MCC can be undefined when a pair of confusion matrix values are **both 0**, but these cases can be handled with some mathematical steps:

D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics, vol. 21, no. 1, p. 6, Dec. 2020

5. Matthews Correlation Coefficient (MCC). Пояснення складових

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \in [-1, +1]$$

$$\sqrt{P \cdot R} = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}} = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}$$

$$\sqrt{P \cdot R} = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}} = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}$$


$$\sqrt{P_1 R_1 P_0 R_0} = \frac{TP \cdot TN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \in [0, 1]$$

- Якщо поміняти класи місцями

□ Якщо хочемо максимум для P і R

Приклад № 1 оцінки належності

Приклад: кредитний скоринг

- Потрібно передбачити, чи поверне клієнт кредит
- Сортуємо клієнтів по оцінці $b(x)$ ймовірності повернення затребуваного кредиту:

$$b(x_1) \leq \dots \leq b(x_\ell)$$

- Банк отримує ранжований список клієнтів
- Поріг вибирається залежно від стратегії банку
- Поріг може багаторазово переглядатися

Приклад № 2 оцінки належності

Приклад: спам

- Потрібно передбачити, чи є даний електронний лист спамом
- Сортуємо листи по оцінці $b(x)$ ймовірності того, що лист є спамом:

$$b(x_1) \leq \dots \leq b(x_\ell)$$

- Отримуємо ранжований список листів
- Поріг вибирається залежно від нашої стратегії
- Поріг може багаторазово переглядатися

5. Matthews Correlation Coefficient (MCC)

- **Нехай лише 100 листів із 1 млн є спамом**
- **Приклад (поріг=0,8):** TP=90, FP=10, FN=10, TN=999890
- Recall (=Sensitivity) = $\frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0,9$
- Specificity = $\frac{TN}{TN + FP} = \frac{999890}{999890 + 10} = 1$
- Precision = $\frac{TP}{TP + FP} = \frac{90}{90 + 10} = 0,9$
- F1 score = $2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = 2 * \frac{0,9 * 0,9}{0,9 + 0,9} = 0,9$
- $$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} =$$
- $$= \frac{90 * 999890 - 10 * 10}{\sqrt{(90 + 10) * (90 + 10) * (999890 + 10) * (999890 + 10)}} = 0,9$$

5. Matthews Correlation Coefficient (MCC)

- Нехай матриця помилок при зміні порогу класифікації змінюється як в таблиці:

Що прогнозуємо?	Поріг	TP	FP	FN	TN
спам	0,7	90	1910	10	997990
спам	0,8	90	10	10	999890
спам	0,9	90	10	1910	997990
не спам	0,8	999890	10	10	90

5. Matthews Correlation Coefficient (MCC)

Що прогнозуємо?	Поріг	Recall	Specificity	Precision	F1 score	MCC
спам	0,7	0,90	1	0,05 (FP ↑)	0,09	0,20
спам	0,8	0,90	1	0,90	0,90	0,90
спам	0,9	0,05 (FN ↑)	1	0,90	0,09	0,20
не спам	0,8	1	0,90	1	1	0,90

- Except for MCC all the other testing metrics have changed
- MCC is not only easily interpretable but also robust to changes in the prediction goal: MCC is class symmetric: switching positives and negatives would lead to the same result

5. Matthews Correlation Coefficient (MCC). Multiclass case

- The Matthews correlation coefficient has been generalized to the multiclass case. This generalization was called the R_K statistic (for K different classes), and defined in terms of a K confusion matrix C :

$$\text{MCC} = \frac{\sum_k \sum_l \sum_m C_{kk}C_{lm} - C_{kl}C_{mk}}{\sqrt{\sum_k (\sum_l C_{kl})(\sum_{k' \neq k} \sum_{l'} C_{k'l'})} \sqrt{\sum_k (\sum_l C_{lk})(\sum_{k' \neq k} \sum_{l'} C_{l'k'})}}$$

- When there are more than two labels the MCC will no longer range between -1 and +1. Instead, the minimum value will be between -1 and 0 depending on the true distribution. The maximum value is always +1

Gorodkin, Jan (2004). "Comparing two K-category assignments by a K-category correlation coefficient". Computational Biology and Chemistry. **28** (5): 367–374

6. Каппа Коена (Cohen's Kappa)

- Коефіцієнт каппа Коена — це статистика, яка вимірює узгодження рішень двох експертів про якісні (категоріальні) об'єкти
- Скажімо, два експерти класифікують електронні листи — це спам чи не спам
- Зазвичай вважається, що це більш надійна міра, ніж простий підрахунок відсотка співпадань думок (=рішень) експертів, оскільки каппа Коена враховує можливість випадкового співпадання їх рішень
- У класичному варіанті (який реалізований у бібліотеці scikit-learn) каппа Коена підраховує узгодженість рішень двох експертів (=оцінювачів, raters, observers, annotators), кожен з яких класифікує N предметів на M взаємовиключних категорій

6. Каппа Коена (Cohen's Kappa)

- Cohen's kappa: a statistic that measures inter-annotator agreement, i.e. a score that expresses the level of agreement between two annotators on a classification problem
- It's not a classifier versus a ground truth
- It is defined as $\kappa = \frac{p_0 - p_e}{1 - p_e}$,

where p_0 is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio, **коєфіцієнт очікуваного співпадання**),

p_e is the expected agreement when both annotators assign labels randomly (the chance agreement ratio, **коєфіцієнт випадкового співпадання**). p_e is estimated using a per-annotator empirical prior over the class labels

6. Каппа Коена (Cohen's Kappa)

- Cohen's kappa: $\kappa = \frac{p_0 - p_e}{1 - p_e}$,
- The Cohen's kappa is a number between -1 and 1
- If the raters are in complete agreement (i.e. $p_0 = 1$) then $\kappa = 1$
- If there is no agreement among the raters other than what would be expected by chance (i.e. $p_0 = p_e$, i.e. the prediction made was similar to random guessing), $\kappa = 0$
- It is possible for the statistic to be negative (if $p_0 < p_e$), which implies that there is no effective agreement between the two raters or the agreement is worse than random
- scikit-learn documentation (https://scikit-learn.org/stable/modules/model_evaluation.html#cohen-kappa):

“The kappa score ... is a number between -1 and 1. Scores above .8 are generally considered good agreement; zero or lower means no agreement (practically random labels).”

6. Каппа Коена (Cohen's Kappa). Приклад розрахунку

- Два члени кредитного комітету оцінювали 50 заявок на отримання кредиту. Результати їх рішень зведемо до таблиці

		Експерт 2	
		дати кредит	не давати кредит
Експерт 1	дати кредит	20	5
	не давати кредит	10	15

- observed agreement ratio $p_0 = \frac{20+15}{50} = 0,7$
- Експерт 1 сказав «Так» 25 заявникам і «Ні» також 25 заявникам, тобто рекомендував дати кредит в 50% випадків
- Експерт 2 сказав «Так» 30 заявникам і «Ні» 20 заявникам, тобто рекомендував дати кредит в 60% випадках

6. Каппа Коена (Cohen's Kappa). Приклад розрахунку

- observed agreement ratio $p_0 = \frac{20+15}{50} = 0,7$
- Рахуючи ці події незалежними, ймовірність того, що експерти одночасно скажуть «Так» дорівнює $0,5 * 0,6 = 0,3$
- Аналогічно ймовірність того, що вони одночасно скажуть «Ні» дорівнює $0,5 * 0,4 = 0,2$
- Тоді спільна ймовірність випадкової згоди рішень (chance agreement ratio): $p_e = 0,3 + 0,2 = 0,5$
- $\kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0,7 - 0,5}{1 - 0,5} = 0,4$

6. Каппа Коена (Cohen's Kappa) як метрика класифікації

- Ідея: оскільки використання точності (Accuracy) викликає сумнів у задачах з сильному дисбалансом класів, треба її значення дещо перенормувати
- Робиться це за допомогою статистики chance adjusted index (<https://github.com/jmgirard/mReliability/wiki/Chance-adjusted-index>): ми точність нашого рішення (Accuracy) пронормуємо за допомогою точності, яку можна було б отримати випадково ($Accuracy_{chance}$)
- Під випадковою тут розуміємо точність рішення, яке отримано з нашого випадковою перестановкою відповідей (=рішень експертів)

6. Каппа Коена (Cohen's Kappa) як метрика класифікації

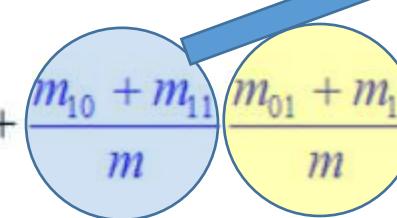
$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

$$\kappa = \frac{\text{Accuracy} - \text{Accuracy}_{\text{chance}}}{1 - \text{Accuracy}_{\text{chance}}}$$

$$\text{Accuracy} = \frac{m_{00} + m_{11}}{m}$$

$$\text{Accuracy}_{\text{chance}} = \frac{m_{00} + m_{01}}{m} \frac{m_{00} + m_{10}}{m} +$$

	$a = 0$	$a = 1$
$y = 0$	m_{00}	m_{01}
$y = 1$	m_{10}	m_{11}



- тут червоним виділена ймовірність вгадати **клас 0**, а синім - **клас 1**
- Дійсно, клас k вгадується, якщо алгоритм видає мітку k і об'єкт дійсно належить цьому класу
- Припускаємо, що це незалежні події. Ймовірність приналежності до класу k можна оцінити по матриці помилок як частку об'єктів класу k .
Аналогічно, ймовірність видати мітку оцінюємо як частку таких міток у відповідях побудованого алгоритму

- Імовірність приналежності до класу 1 = частка об'єктів класу 1
- Імовірність видати мітку класу 1 = частка таких міток у відповідях побудованого алгоритму

5. Matthews Correlation Coefficient & Cohen's Kappa

- MCC & Cohen's kappa take all the cells of the confusion matrix into consideration in its formula

$$\bullet \text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TN + FP) * (TP + FN) * (TN + FN)}}$$

- Chicco D., Warrens M.J., Jurman G. (June 2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access*. 9: 78368 – 78381:

$$\bullet \kappa = \frac{2 * (TP * TN - FP * FN)}{(TP + FP) * (TN + FP) + (TP + FN) * (TN + FN)}$$

- Obviously, κ can be undefined in some cases, but these cases can be handled with mathematical operations similar to the ones needed when MCC is undefined:

D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020

5. Matthews Correlation Coefficient & Cohen's Kappa

- $MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TN + FP) * (TP + FN) * (TN + FN)}}$
- $\kappa = \frac{2 * (TP * TN - FP * FN)}{(TP + FP) * (TN + FP) + (TP + FN) * (TN + FN)}$
- We have $MCC = \kappa$ if and only if $FP = FN$, that is, the metrics coincide when the 2×2 confusion matrix is symmetric
- Furthermore, MCC and Kappa are, respectively, the geometric mean and harmonic mean of the following quantities:

$$\frac{TP * TN - FP * FN}{(TP + FP) * (TN + FP)} \text{ and } \frac{TP * TN - FP * FN}{(TP + FN) * (TN + FN)}$$

- From the geometric-harmonic-means inequality we obtain the inequality
- $$\|MCC\| \geq \|\kappa\| \quad (\text{does not hold for the case of multi-class classification})$$

Chicco D., Warrens M.J., Jurman G. (June 2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. IEEE Access. 9: 78368 – 78381

7. Збалансована акуратність (Balanced Accuracy)

- У разі дисбалансу класів є спеціальний аналог акуратності — збалансована акуратність
- Для простоти запам'ятовування — це середнє повноти всіх класів, або в інших термінах: середнє чутливості (Sensitivity) і специфічності (Specificity)
- Відзначимо, що чутливість і специфічність теж, неформально кажучи, «ортогональні критерії». Легко зробити специфічність 100% -ю, віднісши всі об'єкти до класу 0, при цьому буде 0% -а чутливість, і навпаки, якщо віднести всі об'єкти до класу 1, то буде 0% -а специфічність і 100% -а чутливість
- Якщо в бінарній задачі класифікації представників двох класів приблизно порівну, то $TP + FN \approx TN + FP \approx 1/2$ і збалансована акуратність приблизно дорівнює акуратності звичайній (Accuracy)

7. Збалансована акуратність (Balanced Accuracy). Multiclass case

1. From scikit-learn [I. Guyon, K. Bennett, G. Cawley, H.J. Escalante, S. Escalera, T.K. Ho, N. Macià, B. Ray, M. Saeed, A.R. Statnikov, E. Viegas, Design of the 2015 ChaLearn AutoML Challenge, IJCNN 2015]: random predictions have a score of 0 and perfect predictions have a score of 1

If y_i is the true value of the i -th sample, and w_i is the corresponding sample weight, then we adjust the sample weight to:

$$\hat{w}_i = \frac{w_i}{\sum_j 1(y_j = y_i)w_j}$$

where $1(x)$ is the [indicator function](#). Given predicted \hat{y}_i for sample i , balanced accuracy is defined as:

$$\text{balanced-accuracy}(y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \sum_i 1(\hat{y}_i = y_i) \hat{w}_i$$

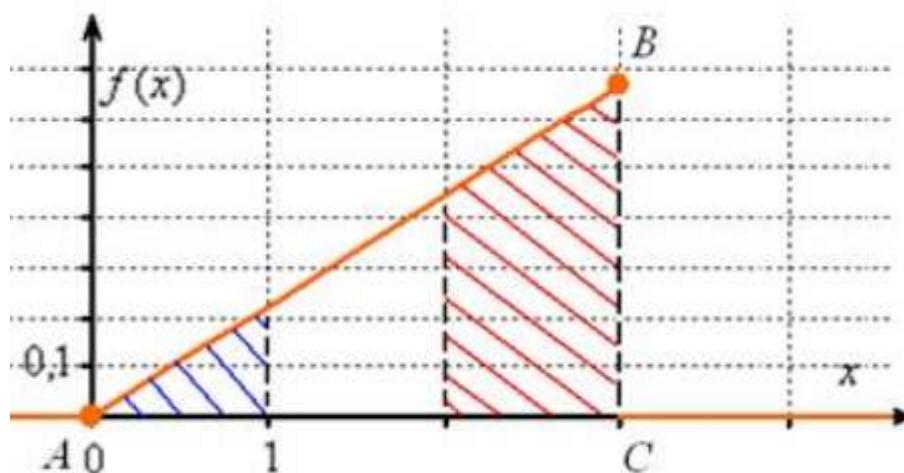
2. [L. Mosley, A balanced approach to the multi-class imbalance problem, IJCV 2010]: Class balanced accuracy is the minimum between the precision and the recall for each class is computed. Those values are then averaged over the total number of classes to get the balanced accuracy.
3. [Urbanowicz R.J., Moore, J.H. ExSTraCS 2.0: description and evaluation of a scalable learning classifier system, Evol. Intel. (2015) 8: 89]: the average of sensitivity and specificity is computed for each class and then averaged over total number of classes

Порівняння метрик №3-7. Функція розподілу випадкової величини

- Порівняння розглянутих метрик будемо проводити на модельному прикладі. Тому згадаємо деякі поняття з курсу теорії ймовірностей
- На відміну від дискретної випадкової величини, неперервна випадкова величина може приймати будь-які дійсні значення на певному проміжку ненульової довжини. Тому її не можна визначити через таблицю
- **Функція розподілу** і для дискретних, і для неперервних випадкових величин визначається однаково — як ймовірність того, що випадкова величина ξ приймає значення, менше, ніж змінна x :
$$F(x) = P(\xi < x)$$
- Функція розподілу, очевидно, є неспадною функцією, що приймає значення на відрізку $[0; 1]$

Порівняння метрик №3-7. Щільність розподілу ймовірності

- Функція **щільності розподілу ймовірності** – це похідна функції розподілу:
 $f(x)=F'(x)$

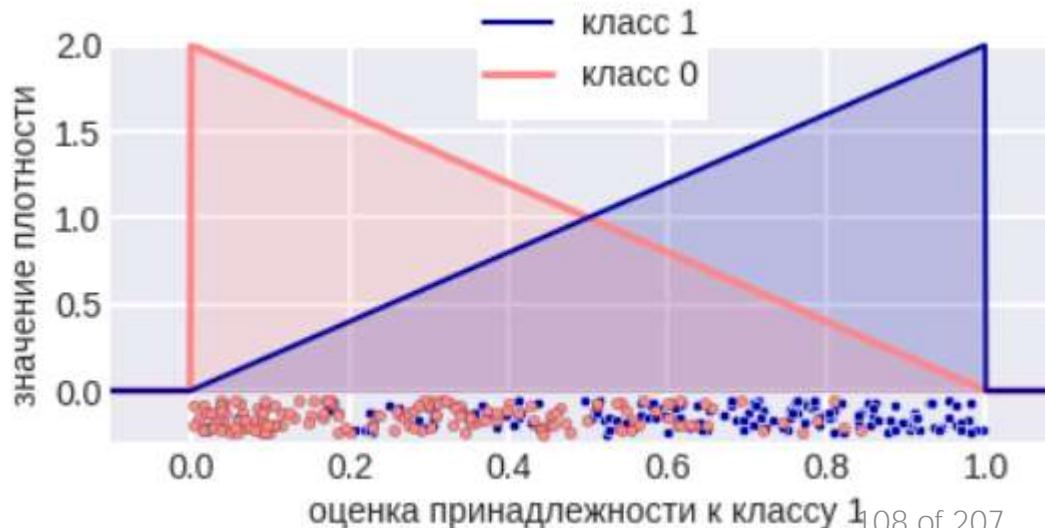


- Властивості щільності розподілу ймовірності:
 - $f(x) \geq 0$
 - $\int_{-\infty}^{+\infty} f(x)dx = 1$

- Чим більша площа під графіком функції щільності, тим більш ймовірнішими є значення із відповідного відрізку
- Оскільки дійсних чисел нескінченно багато, то ймовірність того, що випадкова величина ξ приймає якесь конкретне значення, прямує до нуля. Тому ймовірність обчислюється тільки для числових проміжків

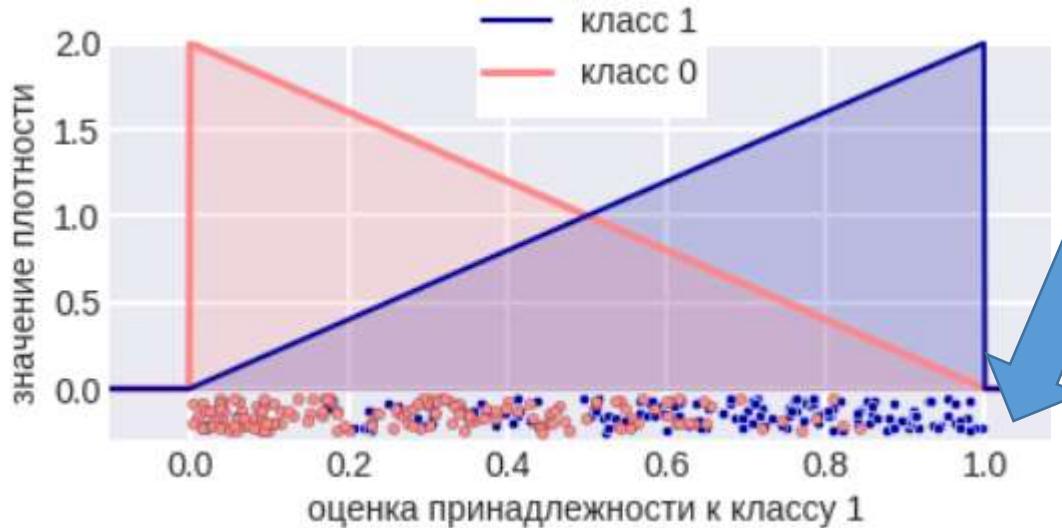
Порівняння метрик №3-7. Модельна задача (1)

- Розв'язуємо задачу бінарної класифікації.
- Нехай наш алгоритм видає оцінки $b(x)$ приналежності об'єкту x до класу 1 на відрізку $[0; 1]$
- І нехай функції щільності розподілу класів на оцінках, породжених цим алгоритмом, є лінійними:
 - за відповідями алгоритму $b(x)$ об'єкти x **класу 0** розподілені зі щільністю $f(x)=2-2x$, а об'єкти **класу 1** – зі щільністю $f(x)=2x$

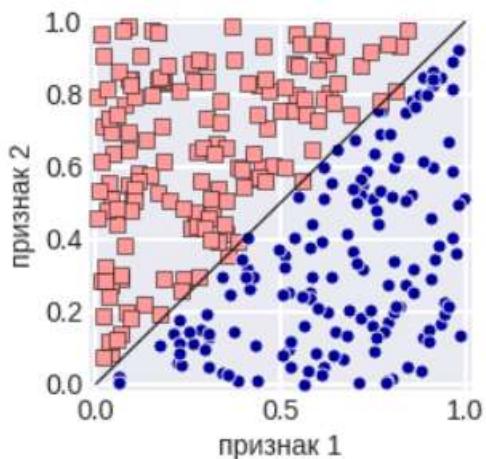


- Інтуїтивно зрозуміло, що алгоритм має певну роздільну здатність: більшість об'єктів класу 0 мають оцінку менше 0.5, а більшість об'єктів класу 1 – більше 0.5

Порівняння метрик №3-7. Модельна задача (2)

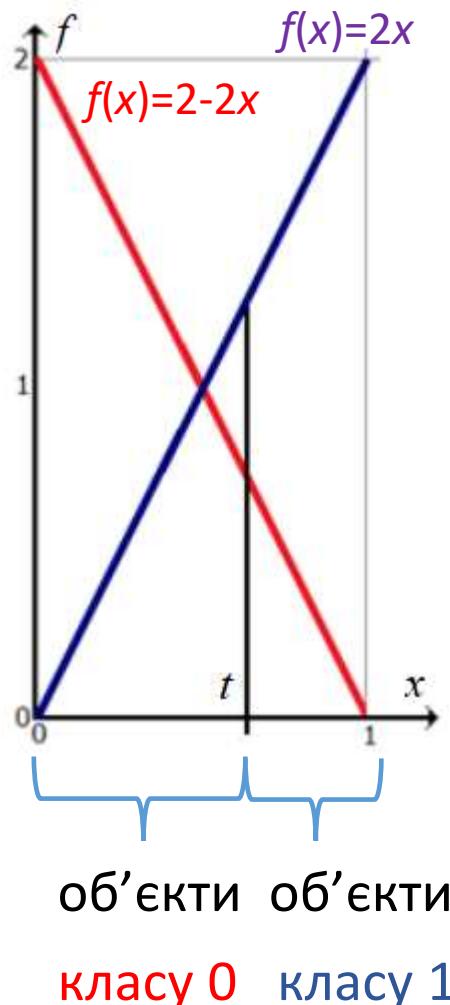


- На рис. показана скінчена (маленька) вибірка, яка відповідає зображенім щільностям
- Хоча ми будемо вважати, що **вибірка нескінчена**, тобто що ми знаємо розподіл об'єктів усіх класів



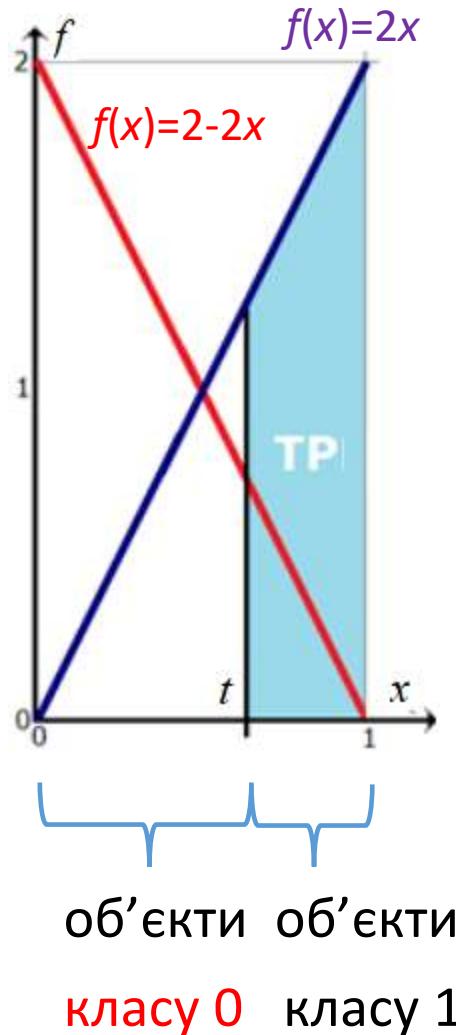
- Така ситуація (модельна задача) **можлива на практиці**, коли об'єкти лежать в квадраті $[0;1] \times [0;1]$, об'єкти вище однієї з діагоналей належать класу 0, нижче – класу 1, для класифікації використовується логістична регресія
- Рішення залежить тільки від першої ознаки (при другій коефіцієнт дорівнює нулю)

Порівняння метрик №3-7. Розрахунок матриці помилок (1)



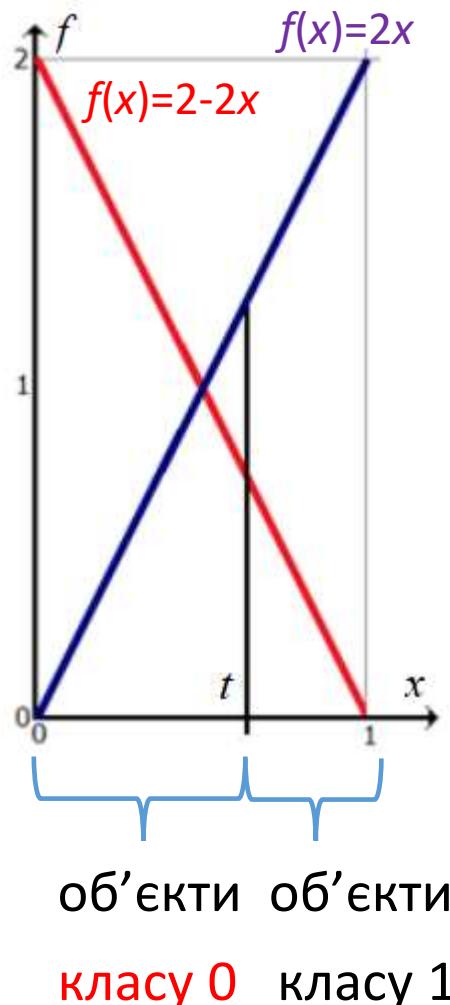
- Класифікатор має вигляд: $a(x) = [b(x) > t]$
де
 - $b(x)$ — оцінка належності x до класу 1
 - t — поріг класифікації
- Потрібно розрахувати (показати на графіку) елементи матриці помилок:
 - TP - ?
 - FP
 - FN
 - TN

Порівняння метрик №3-7. Розрахунок матриці помилок (2)



- Класифікатор має вигляд: $a(x) = [b(x) > t]$
де
 - $b(x)$ — оцінка належності x до класу 1
 - t — поріг класифікації
- Потрібно розрахувати (показати на графіку) елементи матриці помилок:
 - TP — площа прямокутної трапеції під графіком щільності розподілу об'єктів класу 1 від порогу t до 1, тобто $TP = (1-t) \frac{2t+2}{2} = 1-t^2$
 - FP
 - FN
 - TN

Порівняння метрик №3-7. Розрахунок матриці помилок (3)



- Класифікатор має вигляд: $a(x) = [b(x) > t]$

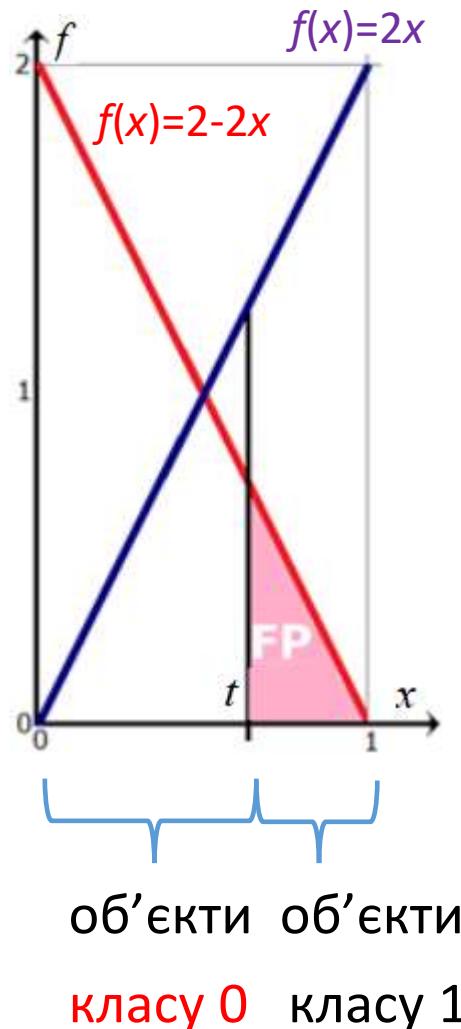
де

- $b(x)$ — оцінка належності x до класу 1
- t — поріг класифікації

- Потрібно розрахувати (показати на графіку) елементи матриці помилок:

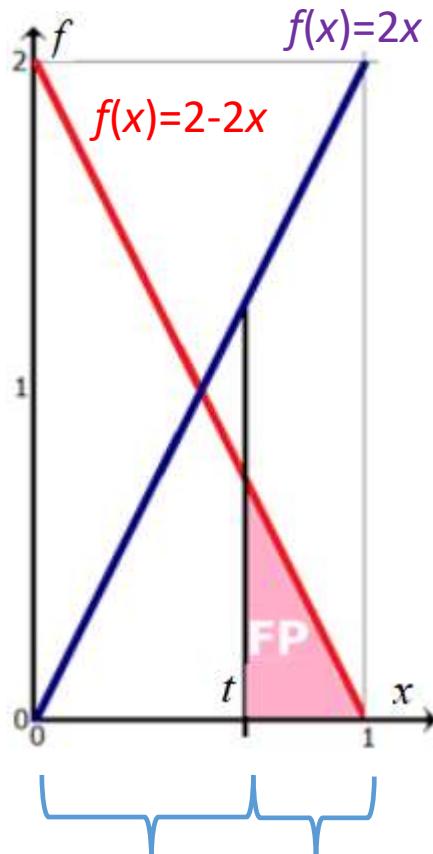
- $TP = 1-t^2$
- FP - ?
- FN
- TN

Порівняння метрик №3-7. Розрахунок матриці помилок (4)



- Класифікатор має вигляд: $a(x) = [b(x) > t]$
де
 - $b(x)$ — оцінка належності x до класу 1
 - t — поріг класифікації
- Потрібно розрахувати (показати на графіку) елементи матриці помилок:
 - $TP = 1-t^2$
 - FP — площа трикутника під графіком щільності розподілу об'єктів класу 0 від порогу t до 1, тобто $FP = \frac{1}{2} (1-t) (2-2t) = (1-t)^2$
 - FN
 - TN

Порівняння метрик №3-7. Розрахунок матриці помилок (5)



об'єкти об'єкти
класу 0 класу 1

- Класифікатор має вигляд: $a(x) = [b(x) > t]$
де
 - $b(x)$ — оцінка належності x до класу 1
 - t — поріг класифікації
- Потрібно розрахувати (показати на графіку) елементи матриці помилок:
 - $TP = 1-t^2$
 - $FP = (1-t)^2$
 - $FN = t^2 = 1-TP$
 - $TN = 1-(1-t)^2 = 1-FP$

Порівняння метрик №3-7. Розрахунок метрик (1)

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

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

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

$$BA = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad A = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\kappa = \frac{\text{Accuracy} - \text{Accuracy}_{\text{chance}}}{1 - \text{Accuracy}_{\text{chance}}}$$

$$\begin{array}{ccc} a = 0 & a = 1 \\ \hline y = 0 & m_{00} & m_{01} \\ & & \\ y = 1 & m_{10} & m_{11} \end{array}$$

$$\text{Accuracy} = \frac{m_{00} + m_{11}}{m}$$

$$\text{Accuracy}_{\text{chance}} = \frac{m_{00} + m_{01}}{m} \frac{m_{00} + m_{10}}{m} + \frac{m_{10} + m_{11}}{m} \frac{m_{01} + m_{11}}{m}$$

□ Розрахуємо значення метрик, якщо:

- $TP = 1-t^2$ $FP = (1-t)^2$
- $FN = t^2$ $TN = 1-(1-t)^2$

Порівняння метрик №3-7. Розрахунок метрик (2)

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

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

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

$$BA = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad A = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\kappa = \frac{Accuracy - Accuracy_{chance}}{1 - Accuracy_{chance}}$$

$$Accuracy = \frac{m_{00} + m_{11}}{m}$$

$$Accuracy_{chance} = \frac{m_{00} + m_{01}}{m} \frac{m_{00} + m_{10}}{m} + \frac{m_{10} + m_{11}}{m} \frac{m_{01} + m_{11}}{m}$$

□ Розрахуємо значення метрик, якщо:

- $TP = 1-t^2 \quad FP = (1-t)^2$
- $FN = t^2 \quad TN = 1-(1-t)^2$

$$P = (1+t)/2 \quad R = 1-t^2$$

$$F_1 = \frac{1-t^2}{3/2-t}$$

$$MCC = \sqrt{t(1-t)}$$

$$BA = Accuracy = \frac{1+2t-2t^2}{2}$$

$$\kappa = \frac{\frac{1+2t-2t^2}{2} - \frac{1}{2}}{1 - \frac{1}{2}} = 2t(1-t)$$

Порівняння метрик №3-7. Пошук максимуму метрик (1)

Метрика	Максимум метрики	За якого порога t досягається максимум метрики?
P	?	?
R	?	?
F_1		
MCC		
BA=Accuracy		
κ		

$P=(1+t)/2$ $R=1-t^2$ $F_1=\frac{1-t^2}{3/2-t}$ $MCC=\sqrt{t(1-t)}$

$BA=Accuracy=\frac{1+2t-2t^2}{2}$ $\kappa=\frac{\frac{1+2t-2t^2}{2}-\frac{1}{2}}{1-\frac{1}{2}}=2t(1-t)$

Порівняння метрик №3-7. Пошук максимуму метрик (2)

Метрика	Максимум метрики	За якого порога t досягається максимум метрики?
P	1	1
R	1	0
F_1		
MCC		
BA=Accuracy		
K		

□ $P=(1+t)/2$ $R=1-t^2$ $F_1=\frac{1-t^2}{3/2-t}$ $MCC=\sqrt{t(1-t)}$

□ $BA=Accuracy=\frac{1+2t-2t^2}{2}$ $K=\frac{\frac{1+2t-2t^2}{2}-\frac{1}{2}}{1-\frac{1}{2}}=2t(1-t)$

Порівняння метрик №3-7. Пошук максимуму метрик (3)

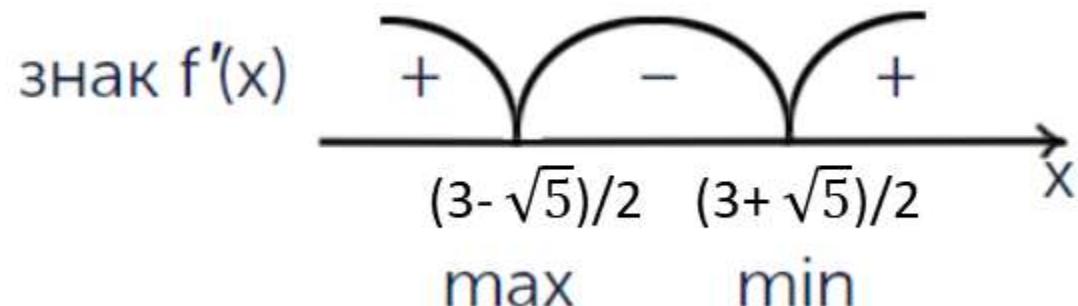
Метрика	Максимум метрики	За якого порога t досягається максимум метрики?
F_1	$3-\sqrt{5}$	$(3-\sqrt{5})/2 \approx 0.38$

- $F_1(t) = \frac{1-t^2}{3/2-t}$
- $F'_1(t) = \frac{-2t(3/2-t)-(-1)(1-t^2)}{(3/2-t)^2} = \frac{t^2-3t+1}{(3/2-t)^2} \implies t^2-3t+1=0 \implies t=(3\pm\sqrt{5})/2$

Порівняння метрик №3-7. Пошук максимуму метрик (3)

Метрика	Максимум метрики	За якого порога t досягається максимум метрики?
F_1	$3-\sqrt{5}$	$(3-\sqrt{5})/2 \approx 0.38$

- $\square F_1(t) = \frac{1-t^2}{3/2-t}$
- $\square F'_1(t) = \frac{-2t(3/2-t)-(-1)(1-t^2)}{(3/2-t)^2} = \frac{t^2-3t+1}{(3/2-t)^2} \implies t^2-3t+1=0 \implies t=(3\pm\sqrt{5})/2$



Порівняння метрик №3-7. Пошук максимуму метрик (2)

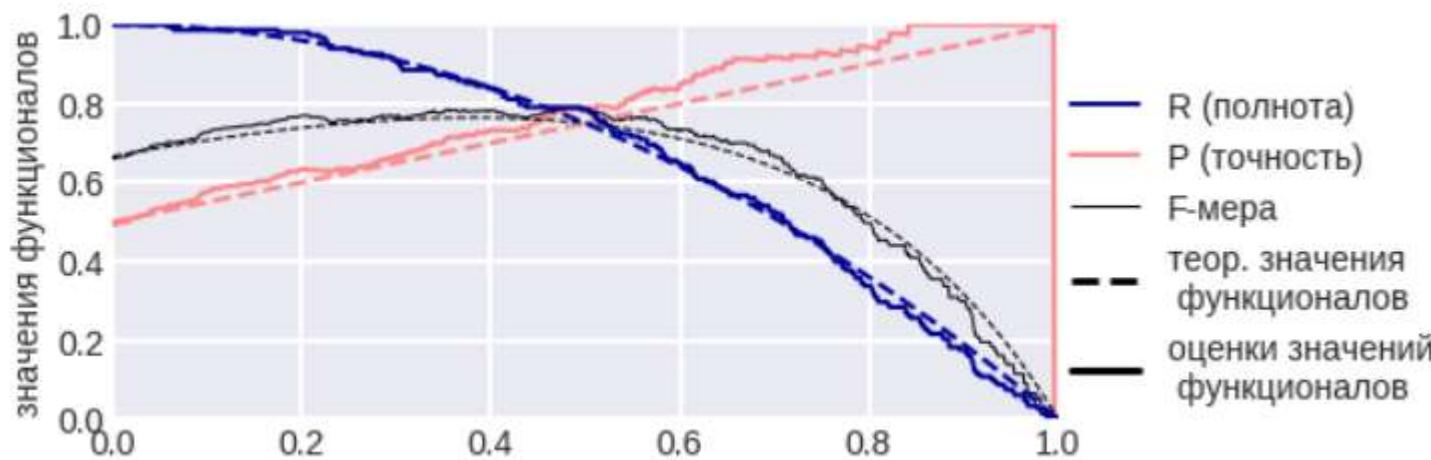
Метрика	Максимум метрики	За якого порога t досягається максимум метрики?
P	1	1
R	1	0
F_1	$3-\sqrt{5}$	$(3-\sqrt{5})/2$
MCC	$\frac{1}{2}$	$\frac{1}{2}$
BA=Accuracy	$\frac{3}{4}$	$\frac{1}{2}$
κ	$\frac{1}{2}$	$\frac{1}{2}$

□ $P=(1+t)/2$ $R=1-t^2$ $F_1=\frac{1-t^2}{3/2-t}$ $MCC=\sqrt{t(1-t)}$

□ $BA=Accuracy=\frac{1+2t-2t^2}{2}$ $\kappa=\frac{\frac{1+2t-2t^2}{2}-\frac{1}{2}}{1-\frac{1}{2}}=2t(1-t)$

Порівняння метрик №3-7. Нескінчених вибірок не має

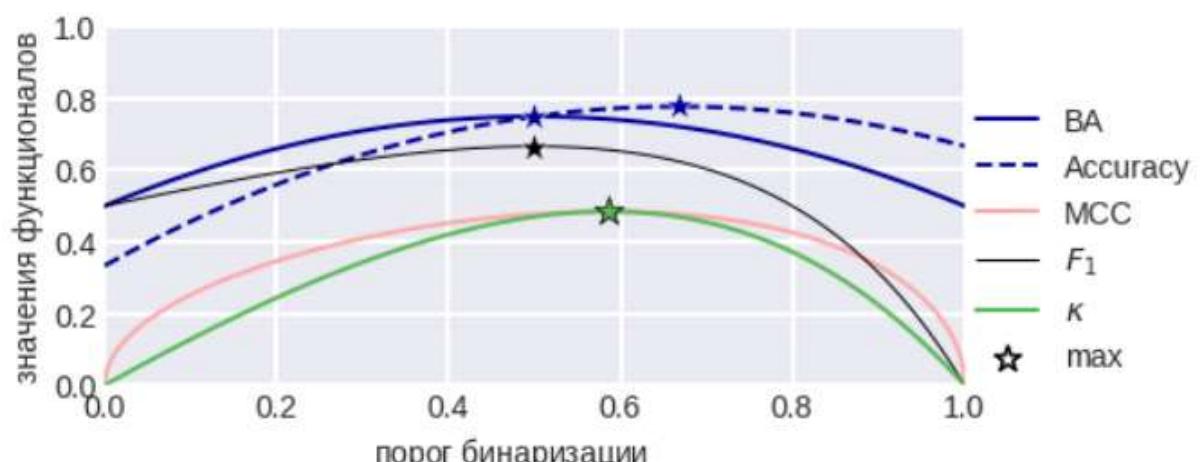
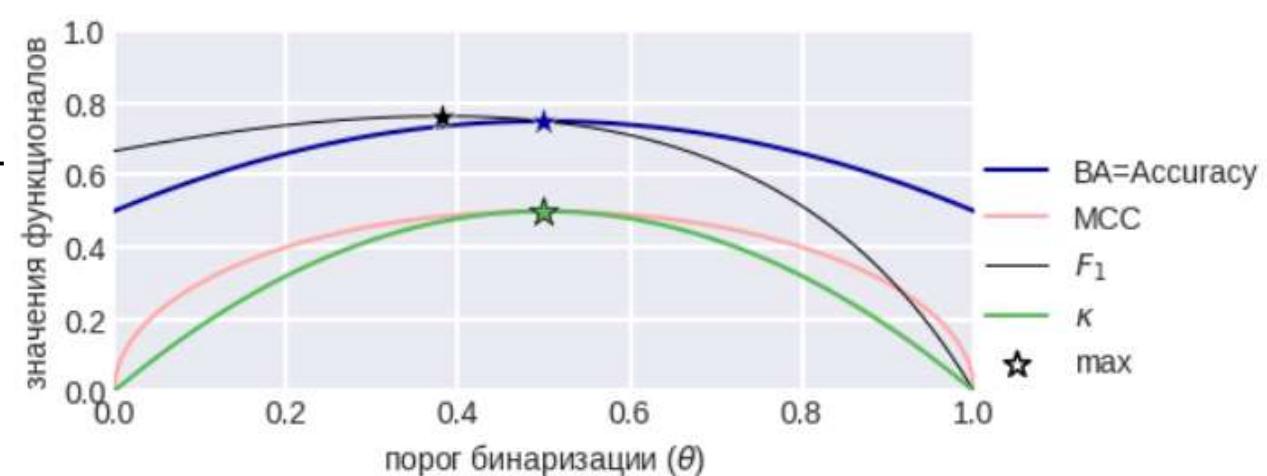
- $P = (1+t)/2$
- $R = 1-t^2$
- $F_1 = \frac{1-t^2}{3/2-t}$



- На практиці у нас немає нескінчених вибірок
- Що зміниться, якщо ми обчислимо значення метрик на скінченній вибірці (наприклад, у 300 об'єктів), об'єкти якої згенеровані відповідно до зазначених розподілів?
- Як видно, криві досить близькі до теоретичних при $t = 300$, при збільшенні вибірки ще у 10 разів практично збігаються

Порівняння метрик №3-7. А якщо вибірка незбалансована?

- $BA = \frac{1+2t-2t^2}{2}$
- $MCC = \sqrt{t(1-t)}$
- $F_1 = \frac{1-t^2}{3/2-t}$
- $\kappa = 2t(1-t)$

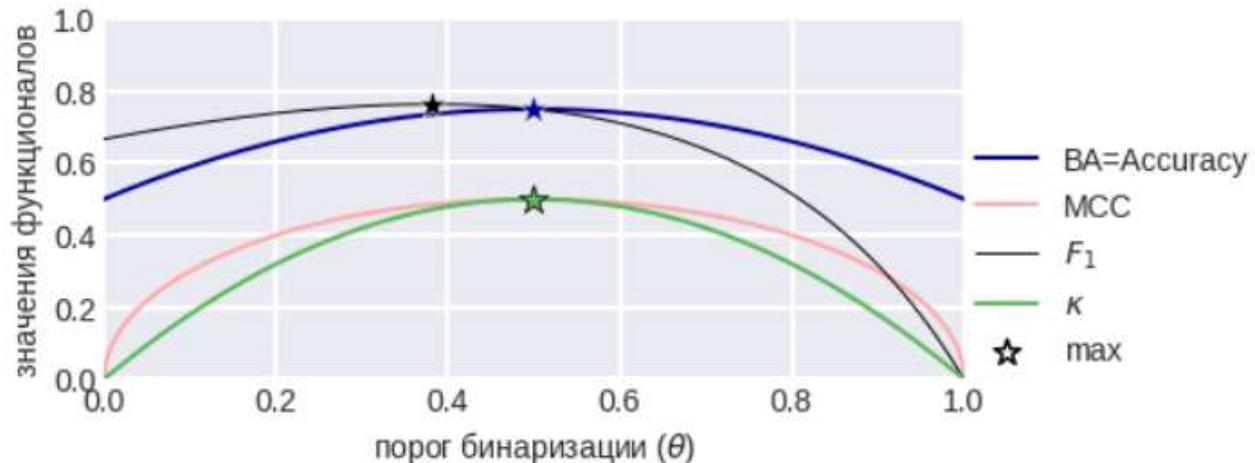


□ Якщо класи у нашій виборці рівномірні, тобто вона збалансована

□ Якщо у виборці об'єктів класу 1 удвічі більше
□ Тільки графік ВА залишився симетричним!

Порівняння метрик №3-7. А якщо вибірка незбалансована?

збалансована

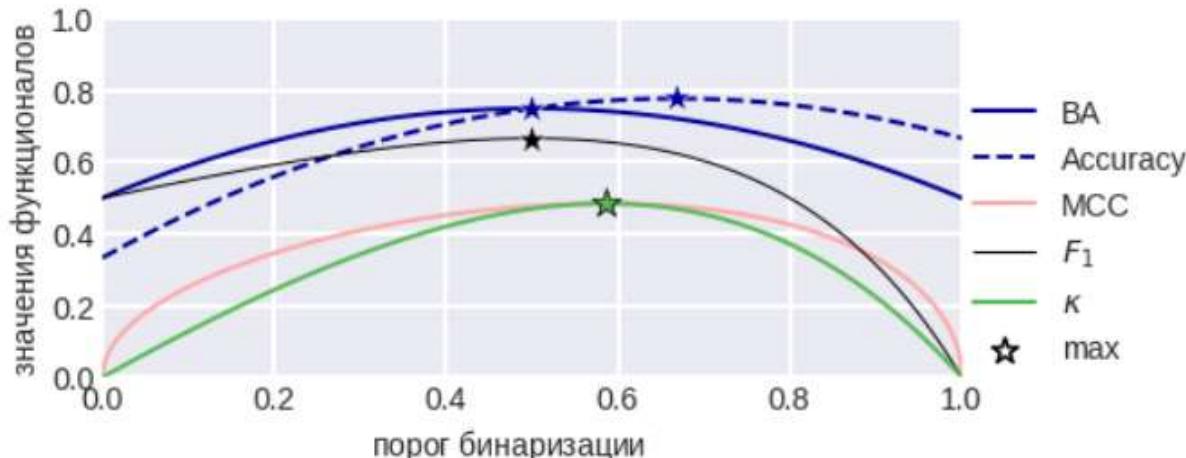


- Якщо вибірка незбалансована, то яку метрику має сенс застосовувати?

- Стандартні поради: ВА, МСС, κ , F_1 мають різні властивості:

- модельна задача показує, що каппа Коена та МСС ведуть себе схоже, а ми раніше з'ясували (коли вводили МСС), що він гарно працює для незбалансованих задач, коли ми хочемо з однаковою успішністю як знаходити об'єкти класу 0, так і класу 1
- ВА дозволяє зберегти незмінним оптимальний поріг класифікації та має симетричний графік

незбалансована



Порівняння метрик №3-7. А якщо вибірка незбалансована?

- Якщо вибірка незбалансована, то яку метрику має сенс застосовувати?
- Наприклад, якщо порівнювати МСС і κ
- $\text{MCC} = \sqrt{t(1 - t)}$ $\kappa = 2t(1 - t)$
- Показано, що на симетричних матрицях помилок навіть для множинної класифікації метрики МСС і κ співпадають
- Але у цій же роботі показано, що для незбалансованих вибірок каппа Коена гірше класифікує, ніж МСС:

Delgado R, Tibau X-A (2019) Why Cohen's Kappa should be avoided as performance measure in classification. PLoS ONE 14(9): e0222916

- In some particular cases, especially when MCC and Cohen's Kappa generate negative discordant scores or when $TP = TN = 0$, the value produced by MCC is more reliable and informative:

Chicco D., Warrens M.J., Jurman G. (June 2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. IEEE Access. 9: 78368 – 78381

Приклад № 3 оцінки належності

Приклад: задача виявлення хворих людей

- $b(x)$ оцінює ймовірність того, що людина є хворою
- Нехай класифікатор має такий вигляд:

$$a(x) = [b(x) > 0.9]$$

- Якщо точність (precision) = 0.2, а повнота (recall) = 0.7, то як зрозуміти, в чому причина низької точності класифікації – неправильний вибір порогу чи погана функція оцінки $b(x)$?

Приклад № 3 оцінки належності

Приклад: задача виявлення хворих людей

- $b(x)$ оцінює ймовірність того, що людина є хворою
- Нехай класифікатор має такий вигляд:

$$a(x) = [b(x) > 0.9]$$

- Якщо точність (precision) = 0.2, а повнота (recall) = 0.7, то як зрозуміти, в чому причина низької точності класифікації – неправильний вибір порогу чи погана функція оцінки $b(x)$?
- **Потрібно перепробувати різні значення порогу!!**

8. Precision-Recall-крива

1. Відсортуємо об'єкти по зростанню оцінки $b(x)$:

$$b(x_1) \leq \dots \leq b(x_\ell)$$

2. Переберемо всі пороги класифікації, почавши з максимального:

$$t_\ell = b(x_\ell) \geq \dots \geq t_1 = b(x_1) \geq t_0 = b(x_1) - \varepsilon$$

віднімаємо ε , щоб розглянути і випадок порогів, менших за наявні

3. Для кожного порога порахуємо точність і повноту

4. Нанесемо відповідну точку в осіх «повнота-точність»

5. З'єднаємо точки, отримавши Precision-Recall-криву

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.90$

$a(x) = [b(x) > 0.9]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=? (правильне спрацьовування)	FP=? (хибне спрацьовування)
	Здорова	FN=? (хибний пропуск)	TN=? (правильний пропуск)

8. Precision-Recall-криива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	0

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.90$

$a(x) = [b(x) > 0.9]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=0 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=3 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	0

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.90$

$a(x) = [b(x) > 0.9]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=0 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=3 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-криива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	0

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.90$

$a(x) = [b(x) > 0.9]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=0 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=3 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	0

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.90$

$a(x) = [b(x) > 0.9]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=1 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=2 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

2. Переберемо всі пороги класифікації, почавши з максимального

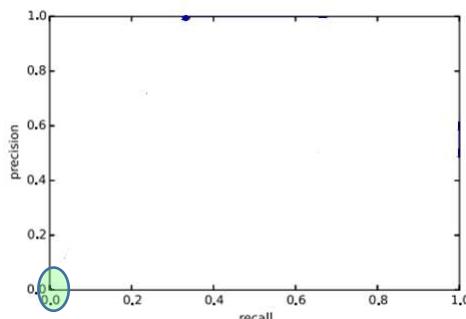
2.1. Беремо поріг $t = 0.90$

TP=0	FP=0
FN=3	TN=3

2.2. Рахуємо точність і повноту:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{0}{0 + 0} = (?) = 0 \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{0}{0 + 3} = 0$$

3. Нанесемо відповідну точку в осіх «повнота-точність»



8. Precision-Recall-криива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	1

2. Переберемо всі пороги класифікації, почавши з максимального

2.2. Беремо поріг $t = 0.73$

$a(x) = [b(x) > 0.73]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=1 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=2 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	1

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.73$

$a(x) = [b(x) > 0.73]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=1 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=2 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	1

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.73$

$a(x) = [b(x) > 0.73]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=1 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=2 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-криива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1
$a(x)$	0	0	0	0	0	1

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = 0.73$

$a(x) = [b(x) > 0.73]$

		Реальна ситуація	
		Хвора людина	Здорова людина
Наш прогноз	Хвора	TP=1 (правильне спрацьовування)	FP=0 (хибне спрацьовування)
	Здорова	FN=2 (хибний пропуск)	TN=3 (правильний пропуск)

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

2. Переберемо всі пороги класифікації, почавши з максимального

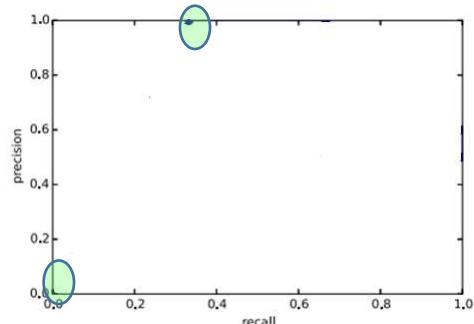
2.1. Беремо поріг $t = 0.73$

2.2. Рахуємо точність і повноту:

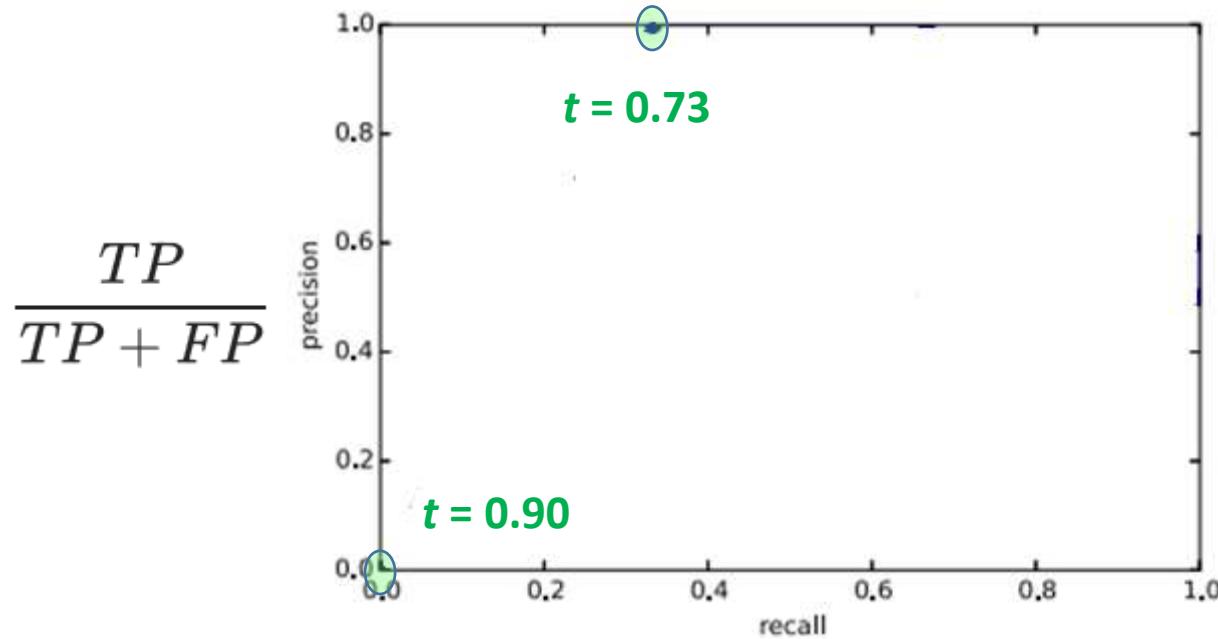
TP=1	FP=0
FN=2	TN=3

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{1}{1 + 0} = 1 \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{1}{1 + 2} = 1/3$$

3. Нанесемо відповідну точку в осіх «повнота-точність»



8. Precision-Recall-крива (приклад побудови)



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

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

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

Закінчіть
побудову
Precision-
Recall-
кривої

8. Precision-Recall-крива (приклад побудови)

1. Відсортували об'єкти по зростанню оцінки $b(x)$:

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

2. Переберемо всі пороги класифікації, почавши з максимального

2.1. Беремо поріг $t = ?$

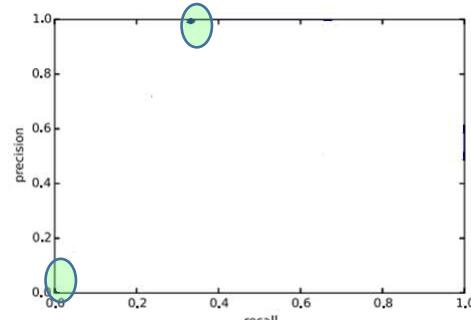
$$a(x) = [b(x) > t]$$

2.2. Рахуємо точність і повноту:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

TP=?	FP=?
FN=?	TN=?

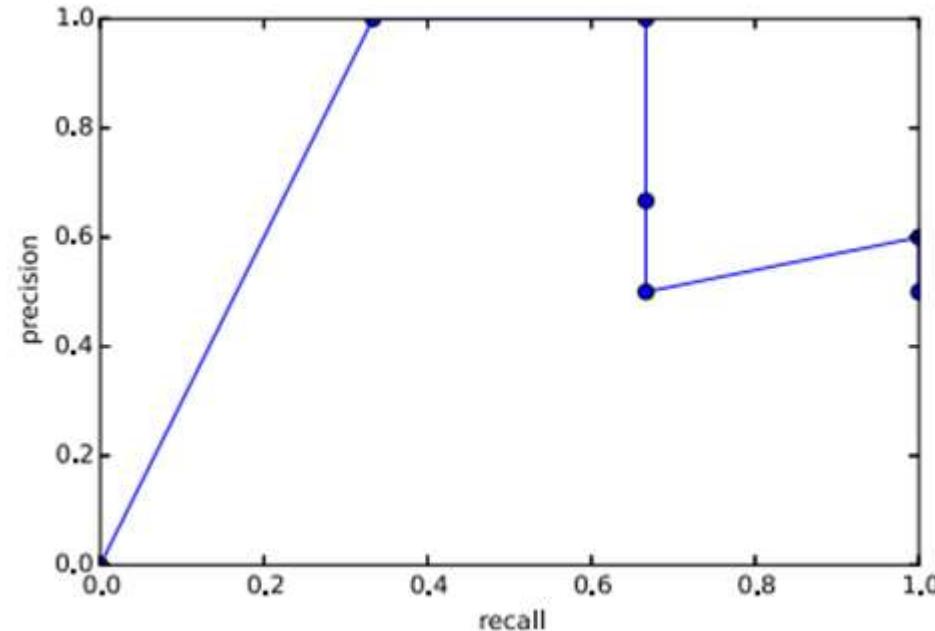
3. Нанесемо відповідну точку в осіх «повнота-точність»



Поріг	Виконавці
0.52	КМ-01
0.39	КМ-02
0.23	КМ-03
0.14	КМ-01, 02
0.10	КМ-03

8. Precision-Recall-криива

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

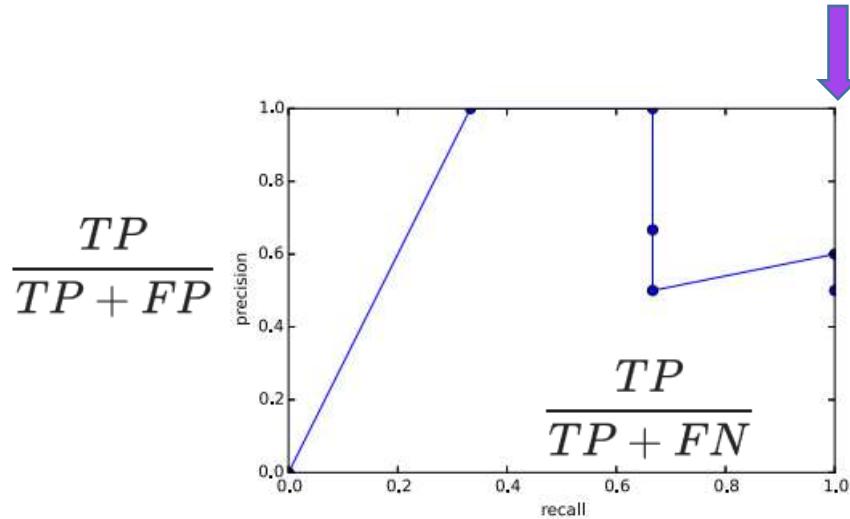


$$\frac{TP}{TP + FN}$$

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

8. Precision-Recall-крива. Властивості

Perfect classifier



$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

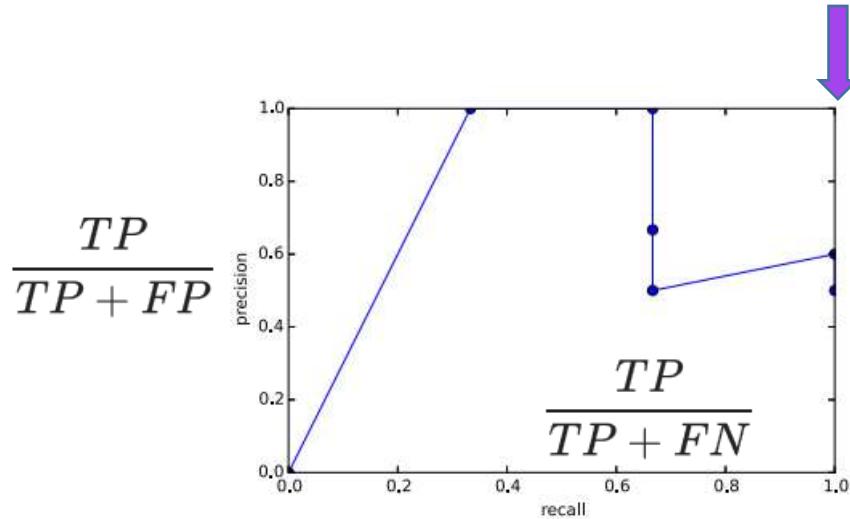
AUC-PRC (Area Under Precision-Recall curve) – міра якості для $b(x)$

Властивості:

- Якщо поріг $t=1$, то класифікатор всіх вважає здоровими – точка $(0, 0)$
- Ліва точка кривої: завжди $(0, 0)$ (не вгадуємо жодного хворого, тобто $TP=0$)
- Якщо поріг $t=0$, то класифікатор всіх вважає хворими, тобто $FN=0$ – точка $(1, 0.5)$
- Права точка: $(1, \ell_+ / \ell)$, ℓ_+ – число об'єктів класу 1 (хворих) у вибірці
- Якщо вибірка ідеально роздільна, тобто $FP=FN=0$, то крива пройде через точку $(1, 1)$
- Чим більше площа під кривою (AUC), тим краще класифікатор. **А чому так?**

8. Precision-Recall-крива. Властивості

Perfect classifier



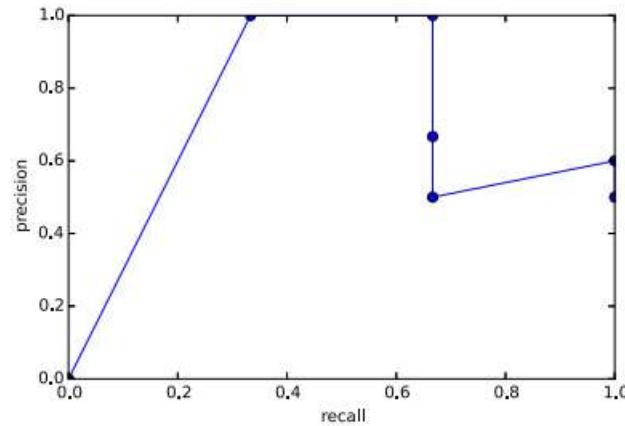
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

AUC-PRC (Area Under Precision-Recall curve) – міра якості для $b(x)$

Властивості:

- Якщо поріг $t=1$, то класифікатор всіх вважає здоровими – точка $(0, 0)$
- Ліва точка кривої: завжди $(0, 0)$ (не вгадуємо жодного хворого, тобто $TP=0$)
- Якщо поріг $t=0$, то класифікатор всіх вважає хворими, тобто $FN=0$ – точка $(1, 0.5)$
- Права точка: $(1, \ell_+ / \ell)$, ℓ_+ – число об'єктів класу 1 (хворих) у вибірці
- Якщо вибірка ідеально роздільна, тобто $FP=FN=0$, то крива пройде через точку $(1, 1)$
- Чим більше площа під кривою (AUC), тим краще класифікатор. **Бо при фіксованій повноті вища точність!**

8. Precision-Recall-крива. Як вибрати поріг класифікації?

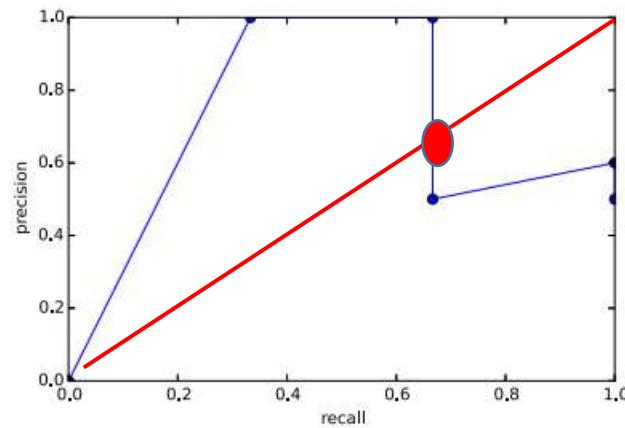


А як по PR-кривій
вибрати поріг
 класифікації?

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

AUC-PRC (Area Under
Precision-Recall curve) –
міра якості для $b(x)$

8. Precision-Recall-крива. Як вибрати поріг класифікації?



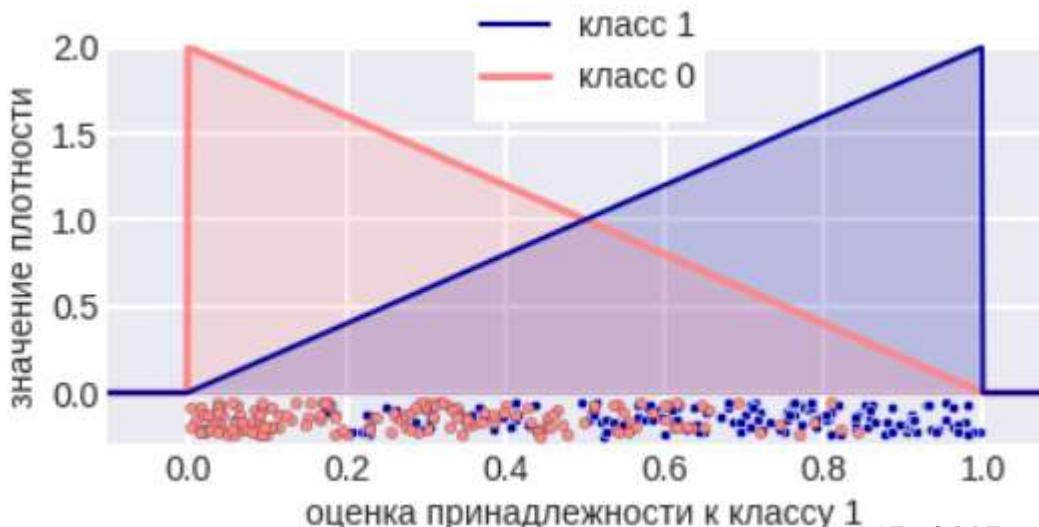
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

AUC-PRC (Area Under Precision-Recall curve) – міра якості для $b(x)$

- ❑ А як по PR-кривій вибрати поріг класифікації?
- ❑ Залежить від того, який результат Ви хочете досягти
- ❑ Якщо точність і повнота рівноправні, то на перетині PR-кривої і діагоналі між точками (0,0) і (1,1)

8. Precision-Recall-крива на нашій модельній задачі (1)

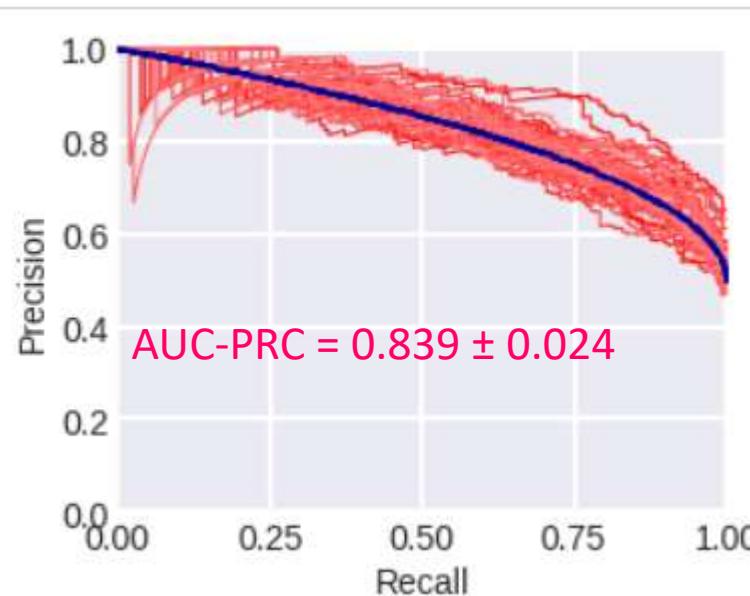
- Розв'язуємо задачу бінарної класифікації
- Нехай наш алгоритм видає оцінки $b(x)$ приналежності об'єкту x до класу 1 на відрізку $[0; 1]$
- І нехай функції щільності розподілу класів на оцінках, породжених цим алгоритмом, є лінійними:
 - за відповідями алгоритму $b(x)$ об'єкти x **класу 0** розподілені зі щільністю $f(x)=2-2x$, а об'єкти **класу 1** – зі щільністю $f(x)=2x$



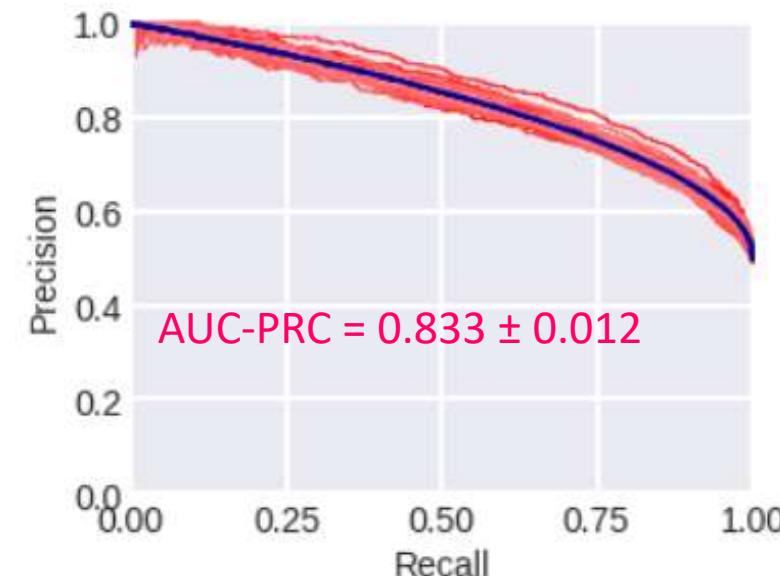
- Інтуїтивно зрозуміло, що алгоритм має певну роздільну здатність: більшість об'єктів класу 0 мають оцінку менше 0.5, а більшість об'єктів класу 1 – більше 0.5

8. Precision-Recall-крива на нашій модельній задачі (2)

- PR-крива в модельній задачі: теоретична (синя) і емпіричні (червоні) для вибірок із певної кількості об'єктів



- Для вибірки із 300 об'єктів



- Для вибірки із 3000 об'єктів

8. Precision-Recall-крива на нашій модельній задачі (3)

- Програмний код для обчислення площі під PR-кривою:

```
from sklearn.metrics import precision_recall_curve
precision, recall, thresholds = precision_recall_curve(y_test, a)
plt.plot(recall, precision)
# вычисление площади методом трапеций
from sklearn.metrics import auc
auc(recall, precision)
# или готовую функцию использовать
from sklearn.metrics import average_precision_score
```

- А давайте підрахуємо площину під теоретичною PR-кривою для модельного прикладу («синій графік»)?!

8. Precision-Recall-крива на нашій модельній задачі (4)

- $P=(1+t)/2$ $R= 1 - t^2$
- Потрібно знайти залежність P від R

$$\left. \begin{array}{l} P = (1+t)/2 \implies t = 2P - 1 \\ R = 1 - t^2 \end{array} \right\} \implies R = 1 - (2P - 1)^2 \implies$$

$$\implies 4P^2 - 4P + R = 0 \implies P = \frac{1}{2}(1 \pm \sqrt{1 - R})$$

- Знаходимо площину під цією кривою на відрізку $[0; 1]$:
- Який знак взяти у формулі: «+» чи «-»?
- AUC-PRC = ??

8. Precision-Recall-крива на нашій модельній задачі (4)

- $P=(1+t)/2$ $R= 1 - t^2$
- Потрібно знайти залежність P від R

$$\begin{aligned} P = (1+t)/2 &\implies t = 2P - 1 \\ R = 1 - t^2 & \end{aligned} \quad \left. \begin{aligned} &\implies R = 1 - (2P - 1)^2 \\ &\implies 4P^2 - 4P + R = 0 \end{aligned} \right\} \implies P = \frac{1}{2}(1 \pm \sqrt{1 - R})$$

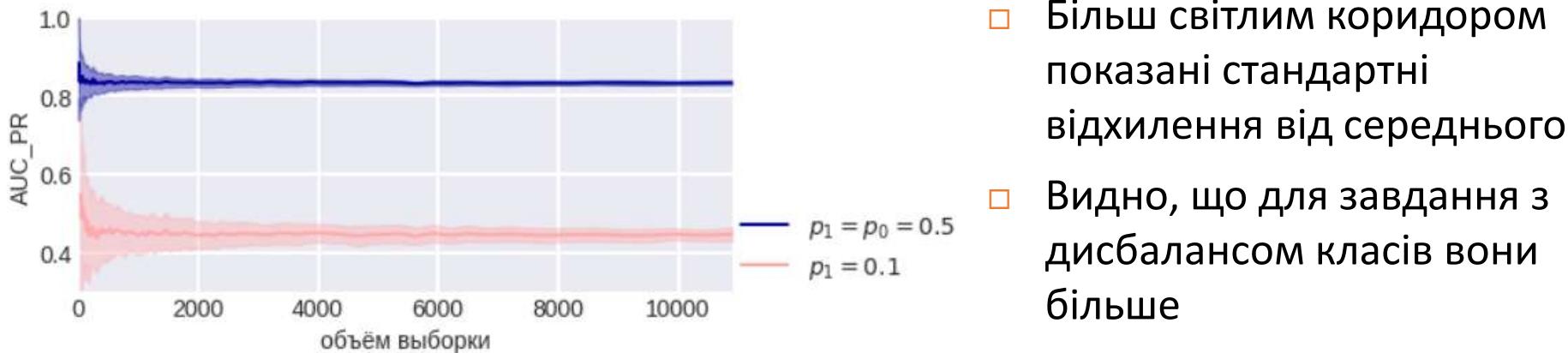
Беремо «+», а не «-», бо інакше при $R=0$ матимемо, що $P=0$.

Звідки $t=-1$, а в модельній задачі поріг змінюється від 0 до 1.

- Знаходимо площину під цією кривою на відрізку $[0; 1]$:
- $AUC(PRC) = \int_0^1 \left(\frac{1}{2} (1 + \sqrt{1 - R}) \right) dR = \frac{5}{6} = 0,8(3)$

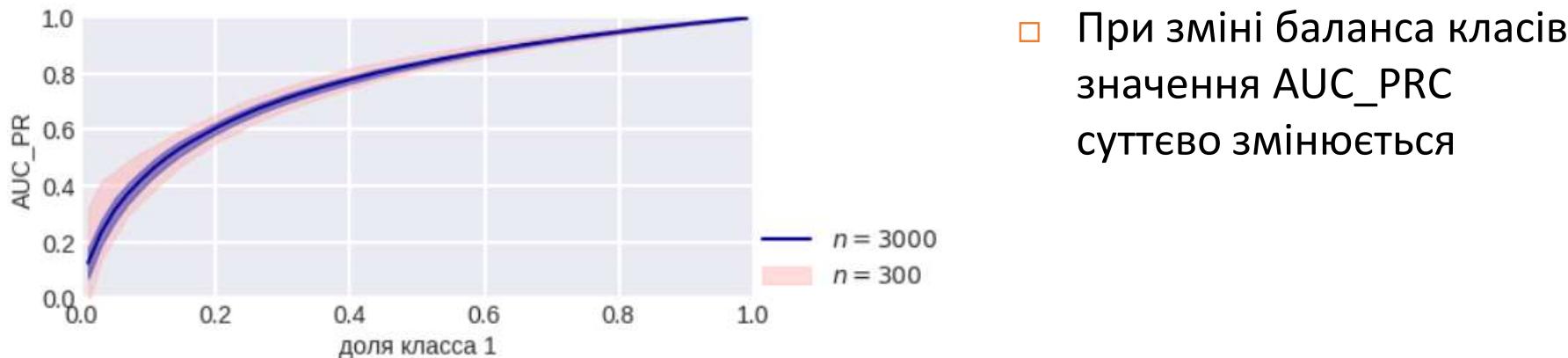
8. Precision-Recall-крива на нашій модельній задачі (5)

- Оцінка AUC_PRC для різного розміру вибірки і різних балансів класів



- Більш світлим коридором показані стандартні відхилення від середнього
- Видно, що для завдання з дисбалансом класів вони більше

- Оцінка AUC_PRC для різного балансу класів



- При зміні баланса класів значення AUC_PRC суттєво змінюється

9. ROC («reciever operating characteristic»)-крива, або крива помилок

- по вісі X : False Positive Rate, доля хибних позитивних спрацьовувань («хибна тривога»):

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

$FPR = 1 - TNR$, де TNR називається **специфічністю (specificity)** алгоритму

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

- по вісі Y : True Positive Rate, доля правильних позитивних спрацьовувань (класифікацій):

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

TPR називається **чутливістю (sensitivity) (=повнотою)** алгоритму

$TPR \equiv Recall$

9. ROC («reciever operating characteristic»)- крива

- ROC-криву будують не по абсолютним значенням (TP і FP), а по відносним — часткам (rates), вираженим у відсотках:

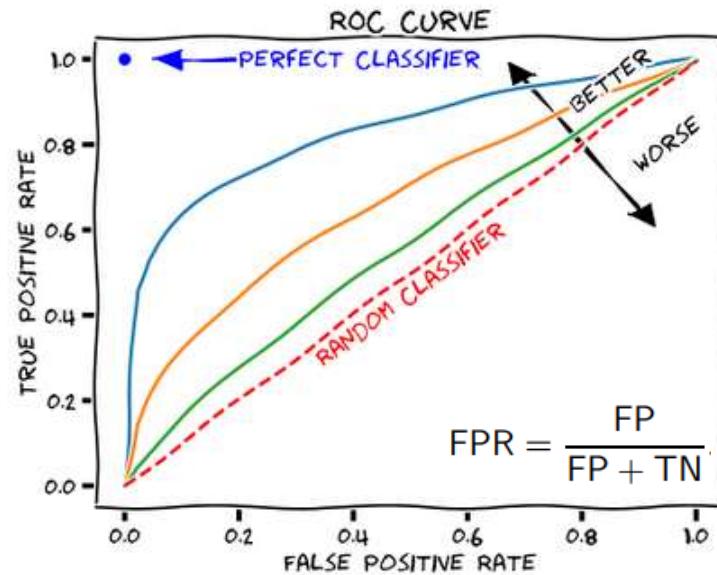
$$TPR = \frac{TP}{TP+FN} \cdot 100 \%$$

$$FPR = \frac{FP}{TN+FP} \cdot 100 \%$$

- Для кожного значення порогу класифікації, яке змінюється від 0 до 1 з певним кроком (скажімо, 0.01) розраховуємо TPR і FPR
- Альтернативно поріг можна рахувати для кожного наступного значення прикладу із вибірки

9. The ROC space for a "better" and "worse" classifier

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



- The best possible prediction method (**perfect classifier**) would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives)

- A **random guess** would give a point along a diagonal line (the so-called **line of no-discrimination**) from the left bottom to the top right corners
- An intuitive example of random guessing is a decision by flipping coins
- As the size of the sample increases, a random classifier's ROC point tends towards the diagonal line
- In the case of a balanced coin, it will tend to the point (0.5, 0.5)

From Wikipedia

9. ROC curve. Random classifier

- s – number of sick people
- h – number of healthy people
- Random classifier for balanced coin

		Fact	
		<i>sick</i>	<i>healthy</i>
Pre- dic- tion	<i>sick</i>	TP≈	FP≈
	<i>healthy</i>	FN≈	TN≈

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

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

9. ROC curve. Random classifier

- s – number of sick people
- h – number of healthy people
- Random classifier
for balanced coin

		Fact	
		sick	healthy
Pre-diction	sick	$TP \approx s/2$	$FP \approx h/2$
	healthy	$FN \approx s/2$	$TN \approx h/2$

$$TPR = \frac{TP}{TP+FN} \approx \frac{s/2}{s} = \frac{1}{2}$$

$$FPR = \frac{FP}{FP+TN} \approx \frac{h/2}{h} = \frac{1}{2}$$

9. ROC curve. Random classifier

- s – number of sick people
- h – number of healthy people



for balanced coin

Random classifier

when "tails" (sick) falls more than
"heads" (healthy) k times

		Fact	
		sick	healthy
Pre-dic-tion	sick	$TP \approx s/2$	$FP \approx h/2$
	healthy	$FN \approx s/2$	$TN \approx h/2$

$$TPR = \frac{TP}{TP+FN} \approx \frac{s/2}{s} = \frac{1}{2}$$

$$FPR = \frac{FP}{FP+TN} \approx \frac{h/2}{h} = \frac{1}{2}$$

		Fact	
		sick	healthy
Pre-dic-tion	sick	$TP \approx$	$FP \approx$
	healthy	$FN \approx$	$TN \approx$

$$TPR = ?$$

$$FPR = ?$$

9. ROC curve. Random classifier

- s – number of sick people
- h – number of healthy people
- Random classifier

for balanced coin

$$\begin{cases} \text{TP} + \text{FN} = s \\ \text{TP} / \text{FN} = k \end{cases} \rightarrow \begin{cases} \text{TP} = ks / (k+1) \\ \text{FN} = s / (k+1) \end{cases}$$

when "tails" (sick) falls more than
"heads" (healthy) k times

		Fact	
		sick	healthy
Pre-dic-tion	sick	$\text{TP} \approx s/2$	$\text{FP} \approx h/2$
	healthy	$\text{FN} \approx s/2$	$\text{TN} \approx h/2$

		Fact	
		sick	healthy
Pre-dic-tion	sick	$\text{TP} \approx$	$\text{FP} \approx$
	healthy	$\text{FN} \approx$	$\text{TN} \approx$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \approx \frac{s/2}{s} = \frac{1}{2} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \approx \frac{h/2}{h} = \frac{1}{2}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \approx ? \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \approx ?$$

9. ROC curve. Random classifier

- s – number of sick people
- h – number of healthy people
-

for balanced coin

Random classifier

$$\begin{cases} \text{TP} + \text{FN} = s \\ \text{TP} / \text{FN} = k \end{cases} \rightarrow \begin{cases} \text{TP} = ks / (k+1) \\ \text{FN} = s / (k+1) \end{cases}$$

when "tails" (sick) falls more than
"heads" (healthy) k times

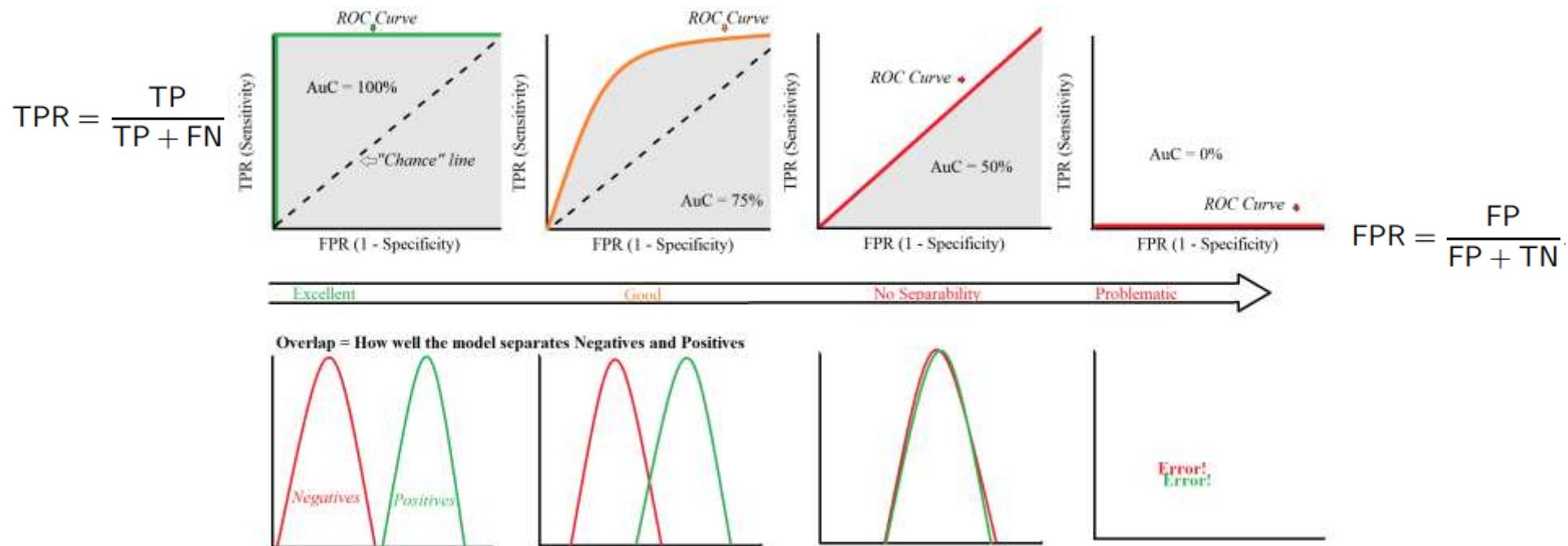
		Fact	
		sick	healthy
Pre-diction	sick	$\text{TP} \approx s/2$	$\text{FP} \approx h/2$
	healthy	$\text{FN} \approx s/2$	$\text{TN} \approx h/2$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \approx \frac{s/2}{s} = \frac{1}{2} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \approx \frac{h/2}{h} = \frac{1}{2}$$

		Fact	
		sick	healthy
Pre-diction	sick	$\text{TP} \approx ks/(k+1)$	$\text{FP} \approx kh/(k+1)$
	healthy	$\text{FN} \approx s/(k+1)$	$\text{TN} \approx h/(k+1)$

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \approx \frac{ks/(k+1)}{s} = \frac{k}{k+1} \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}} \approx \frac{kh/(k+1)}{h} = \frac{k}{k+1} \end{aligned}$$

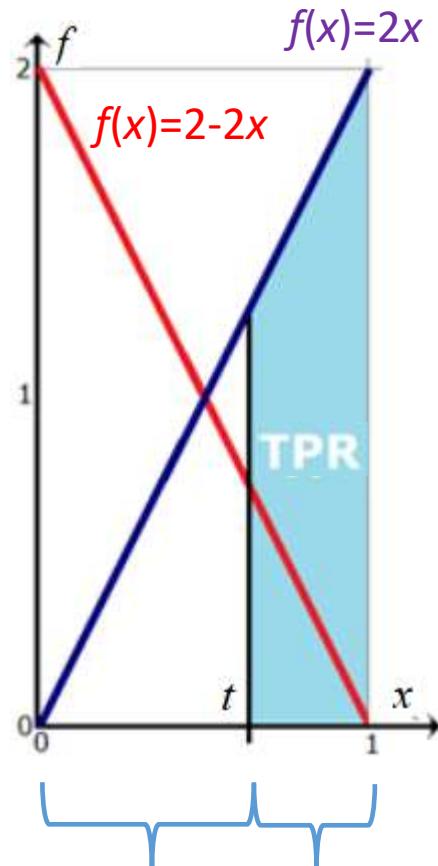
9. Границні випадки ROC-кривої



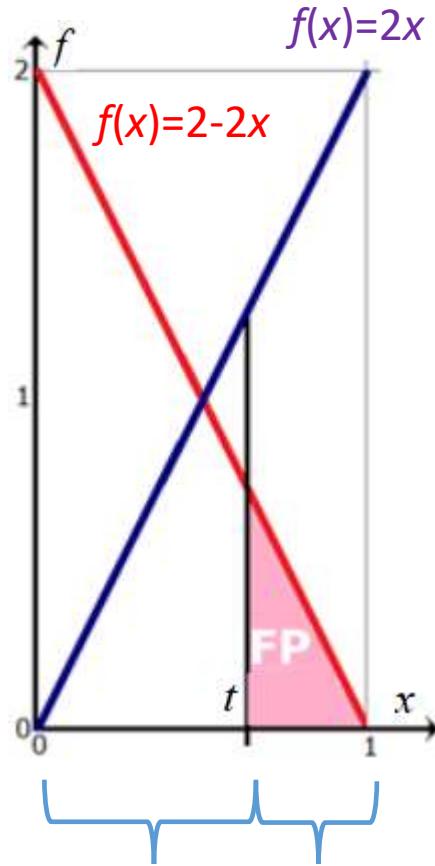
Площа під ROC-кривою для найкращого ($AUC=1$), непоганого ($AUC=0.75$), випадкового ($AUC=0.5$) та найгіршого ($AUC=0$) алгоритмів

- Площа під ROC-кривою AUC (Area Under Curve) є агрегованою характеристикою якості класифікатора
- Чим більше значення AUC, тим «краще» модель класифікації
- Даний показник часто використовується для порівняльного аналізу кількох моделей класифікації

9. Підрахуємо AUC (ROC) для модельної задачі (1)



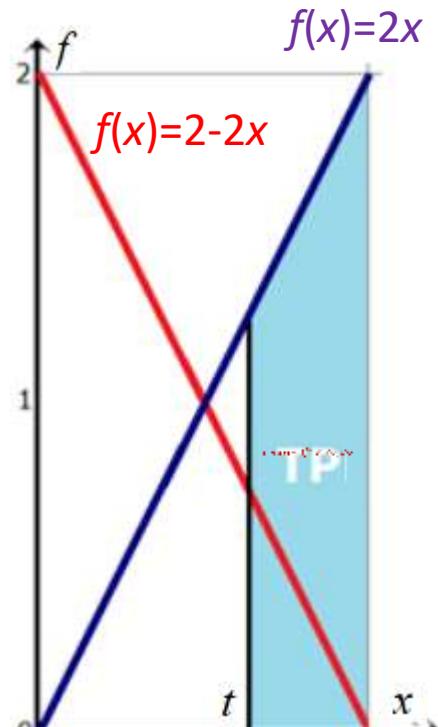
об'єкти
об'єкти
класу 0 класу 1



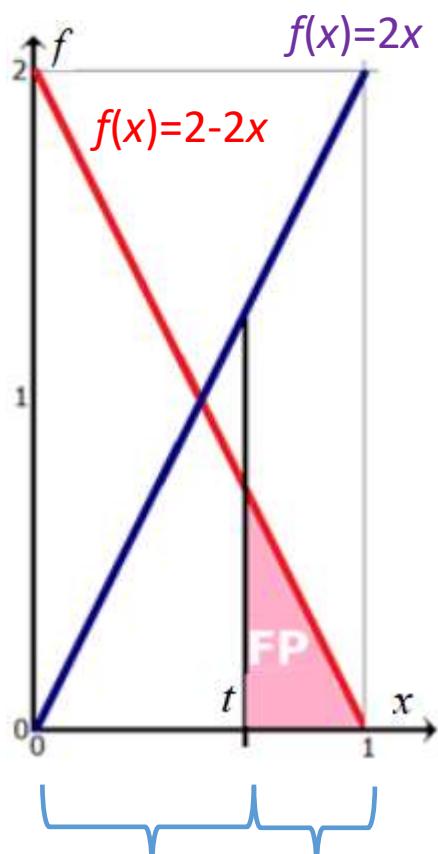
об'єкти
об'єкти
класу 0 класу 1

- Класифікатор має вигляд: $a(x) = [b(x) > t]$ де
 - $b(x)$ — оцінка належності x до класу 1
 - t — поріг класифікації
- Потрібно розрахувати (показати на графіку) елементи матриці помилок:
 - $TP = 1-t^2$ $TPR = 1-t^2$
 - $FP = (1-t)^2$ $FPR = (1-t)^2$
 - $FN = t^2$
 - $TN = 1-(1-t)^2$
- А якщо записати у загальному вигляді (через інтеграли)?

9. Підрахуємо AUC (ROC) для модельної задачі (2)



об'єкти об'єкти
об'єкти **об'єкти**
класу 0 **класу 1**

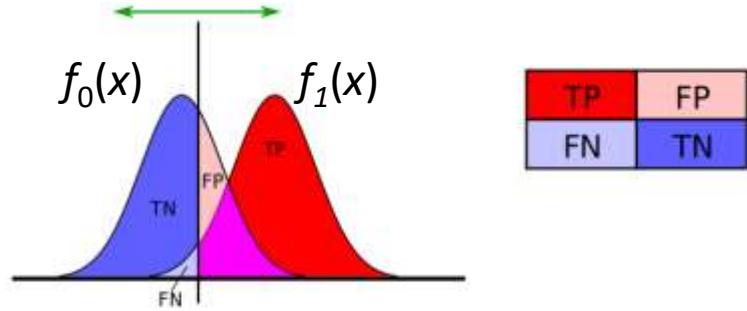


об'єкти об'єкти
об'єкти **об'єкти**
класу 0 **класу 1**

Classes 2-3. Classification Metrics

- Класифікатор має вигляд: $a(x) = [b(x) > t]$ де
 - $b(x)$ — оцінка належності x до класу 1
 - t — поріг класифікації
- Потрібно розрахувати (показати на графіку) елементи матриці помилок:
 - $TP = 1-t^2$ $TPR(t) = \int_t^{\infty} f_1(x)dx$
 - $FP = (1-t)^2$ $FPR(t) = \int_t^{\infty} f_0(x)dx$
- А якщо записати у загальному вигляді (через інтеграли)?

9. Class prediction for each instance based on a continuous random variable

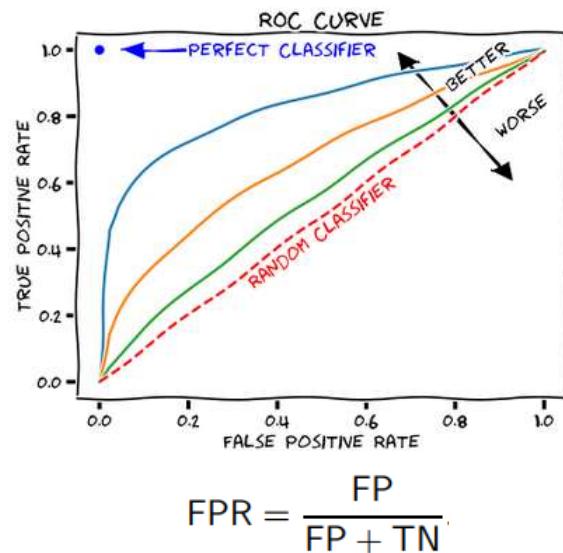


$$TPR(t) = \int_t^{\infty} f_1(x)dx$$

$$FPR(t) = \int_t^{\infty} f_0(x)dx$$

- In binary classification, the class prediction for each instance is often made based on a **continuous random variable** ξ , which is a "score" computed for the instance (e.g. the estimated probability in logistic regression)
- Given a threshold parameter t , the instance is classified as "positive" if $\xi > t$ and "negative" otherwise
- ξ follows a probability density $f_1(x)$ if the instance actually belongs to class "positive", and $f_0(x)$ if otherwise

9. Підрахуємо AUC (ROC) для модельної задачі (3)



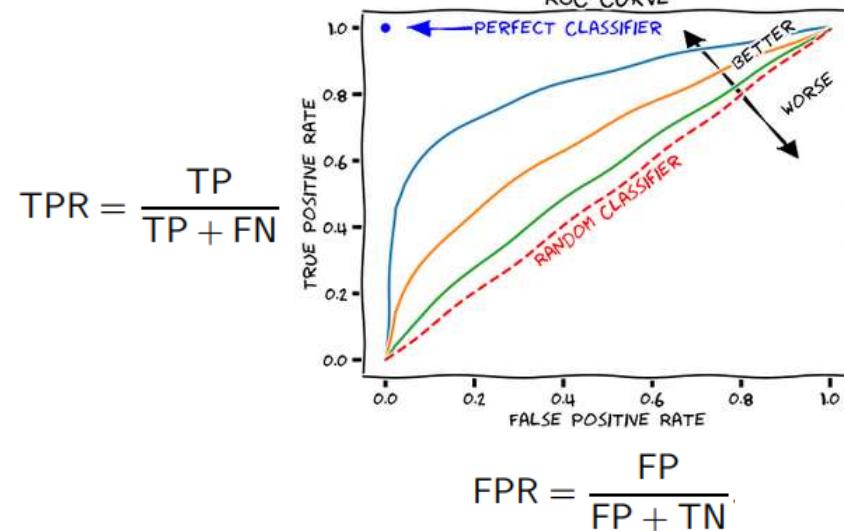
- Значення метрик, що визначають абсцису та ординату ROC-кривої:

- $TPR = 1-t^2$ $FPR = (1-t)^2$
- Абсциса $x = FPR(t)$
- Ордината $y = TPR(t)$

Чому межі інтегрування поміняні місцями?

$$AUC(ROC) = \int_0^1 y(x)dx = \int_{\infty}^{-\infty} (1-t^2)d(1-t)^2$$

9. Підрахуємо AUC (ROC) для модельної задачі (4)



- Значення метрик, що визначають абсцису та ординату ROC-кривої:

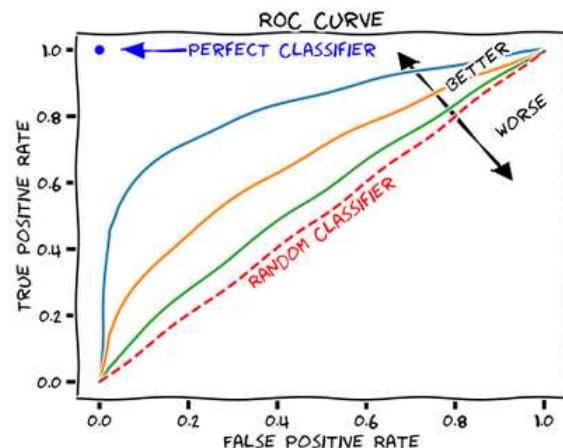
- $TPR = 1-t^2$ $FPR = (1-t)^2$
- Абсциса $x = FPR(t)$
- Ордината $y = TPR(t)$

Тому що абсциса $x = FPR(t)$
зростає, коли параметр t зменшується (від 1 до 0)

$$\begin{aligned} AUC(ROC) &= \int_0^1 y(x)dx = \int_{\infty}^{-\infty} (1-t^2)d(1-t)^2 = \\ &= -2 \int_{\infty}^{-\infty} (1-t^2)(1-t)dt = 2 \int_{-\infty}^{\infty} (1-t^2)(1-t)dt = \\ &= 2 \int_0^1 (1-t^2)(1-t)dt = 2 \int_0^1 (1-t-t^2+t^3)dt = \frac{5}{6} = 0,8(3) \end{aligned}$$

$AUC(ROC) = AUC(PPC)$ – Так, звичайно, не завжди

9. Підрахуємо AUC (ROC) у загальному випадку



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

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

$$TPR(t) = \int_t^{\infty} f_1(x)dx$$

$$FPR(t) = \int_t^{\infty} f_0(x)dx$$

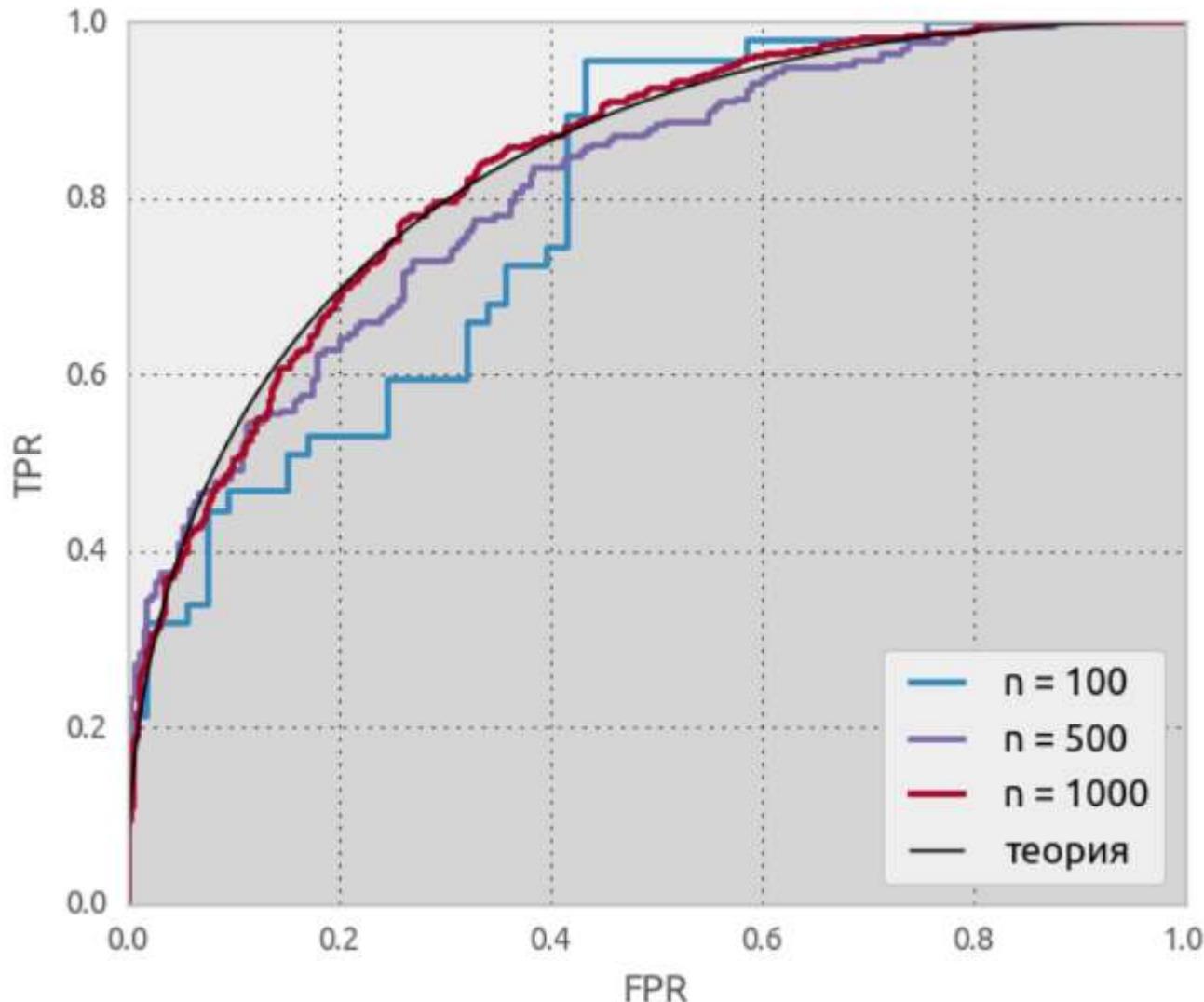
- Значення метрик, що визначають абсцису та ординату ROC-кривої:

- Абсциса $x = FPR(t)$
- Ордината $y = TPR(t)$

$$\begin{aligned} AUC(ROC) &= \int_{-\infty}^1 y(x)dx = \int_{-\infty}^{\infty} TPR(t)d(FPR(t)) = \\ &= \int_{-\infty}^{\infty} TPR(t)FPR'(t)dt = \int_{-\infty}^{\infty} \left(\int_t^{\infty} f_1(t^*)dt^* \right) (-f_0(t))dt = \\ &= \int_{-\infty}^{\infty} \left(\int_t^{\infty} f_1(t^*)dt^* \right) f_0(t)dt = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \lambda_{t^* > t} f_1(t^*) f_0(t)dt^*dt = \\ &= P(\xi_1 > \xi_0) \end{aligned}$$

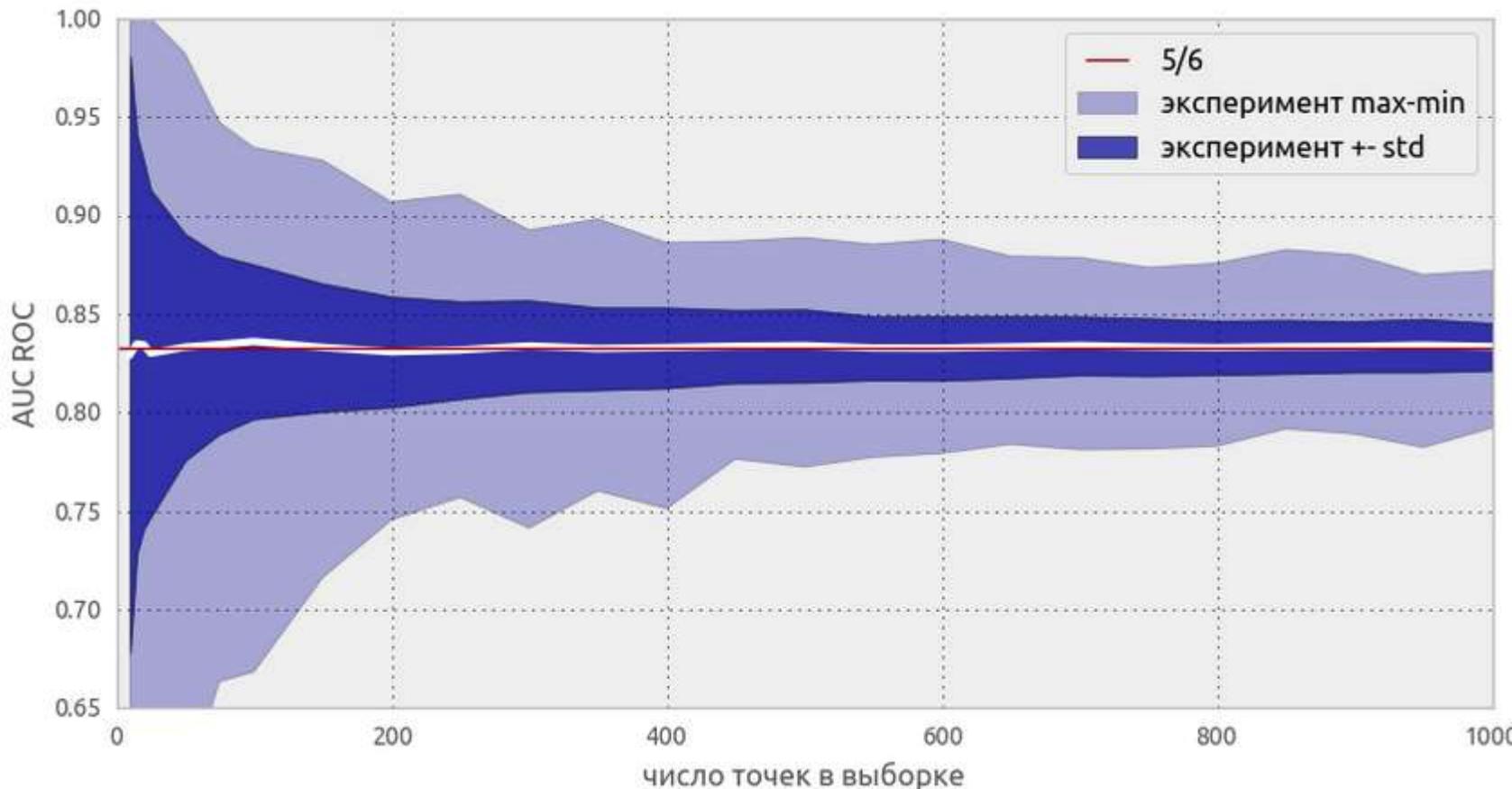
Якщо об'єкти двох класів описуються щільностями), то $AUC(ROC)$ можна трактувати як **ймовірність того, що випадково взятий об'єкт класу 1 має оцінку приналежності до класу 1 вище, ніж випадково взятий об'єкт класу 0**

9. ROC-крива для модельної задачі



- При збільшенні обсягу вибірки ROC-криві, побудовані по вибірці, будуть збігатися до теоретичної кривої, побудованої для розподілу

9. AUC (ROC) для модельної задачі

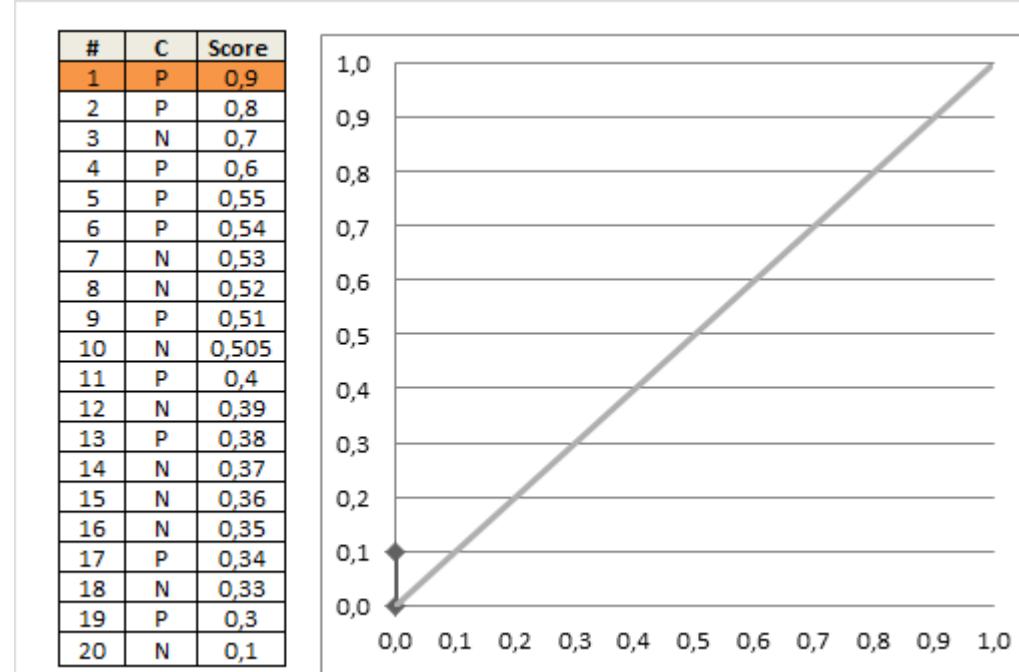


□ Для оцінки AUC (ROC) вибірка у кілька сотень об'єктів мала!

9. Optimizing area under the ROC curve

- ❑ Оптимізувати AUC ROC безпосередньо важко з кількох причин:
 - ця функція недиференційовна за параметрами алгоритму
 - вона в явному вигляді не розбивається на окремі складові, які залежать від відповіді тільки на одному об'єкті
- ❑ Є кілька підходів до оптимізації - див. розділ “5. Optimizing Area Under the Curve” в статті “Davis J., Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves. // Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA”
- ❑ C. Marrocco, R. P. W. Duin, F. Tortorella. Maximizing the area under the ROC curve by pairwise feature combination. Pattern Recognition, Volume 41, Issue 6, June, 2008, pp 1961–1974
<https://doi.org/10.1016/j.patcog.2007.11.017>

9. ROC-крива



http://mlwiki.org/index.php/ROC_Analysis

20 examples from: <https://www.dropbox.com/s/65rdiv42ixe2eac/roc-lift.xlsx>
C - actual class of the training example

9. ROC-крива. Неправильний алгоритм побудови ROC-кривої

Алгоритм построения ROC-кривой

Следующий алгоритм строит ROC-кривую за m обращений к дискриминантной функции.

Входные данные:

- Выборка X^m
- Функция $f(x, w)$ при фиксированном векторе параметров w .

Результат:

- $\{(FPR_i, TPR_i)\}_{i=0}^m$ — последовательность из $(m+1)$ точек ROC-кривой;
- AUC — площадь под ROC-кривой;

1. вычислить количество представителей классов $+1$ и -1 в выборке:
 $m_- := \sum_{i=1}^m [y_i = -1], m_+ := \sum_{i=1}^m [y_i = +1];$
2. упорядочить выборку X^m по убыванию значений $f(x_i, w)$;
3. установить начальную точку ROC-кривой:
 $(FPR_0, TPR_0) := (0,0);$
 $AUC := 0;$
4. для всех $i := 1..m$
если $(y_i = -1)$, то сместиться на один шаг вправо:
 $FPR_i := FPR_{i-1} + \frac{1}{m_-}; TPR_i := TPR_{i-1};$
 $AUC := AUC + \frac{1}{m_-} TPR_i;$
5. иначе сместиться на один шаг вверх:
 $FPR_i := FPR_{i-1}; TPR_i := TPR_{i-1} + \frac{1}{m_+};$

<http://www.machinelearning.ru/wiki/index.php?title=ROC-кривая>

Канонический алгоритм построения ROC-кривой

Входы: L — множество примеров $f[i]$ — рейтинг, полученный моделью, или вероятность того, что i -й пример имеет положительный исход; min и max — минимальное и максимальное значения, возвращаемые f ; d_x — шаг; P и N — количество положительных и отрицательных примеров соответственно.

1. $t = min$
2. повторять
3. $FP = TP = 0$
4. для всех примеров i принадлежит L {
5. если $f[i] \geq t$ тогда // этот пример находится за порогом
6. если i положительный пример тогда
7. $\{TP = TP + 1\}$
8. иначе // это отрицательный пример
9. $\{FP = FP + 1\}$
10. }
11. $Se = TP / P * 100$
12. $point = FP / N //$ расчет $(100 \text{ минус } Sp)$
13. Добавить точку $(point, Se)$ в ROC-кривую
14. $t = t + d_x$
15. пока $(t > max)$

<https://loginom.ru/blog/logistic-regression-roc-auc>

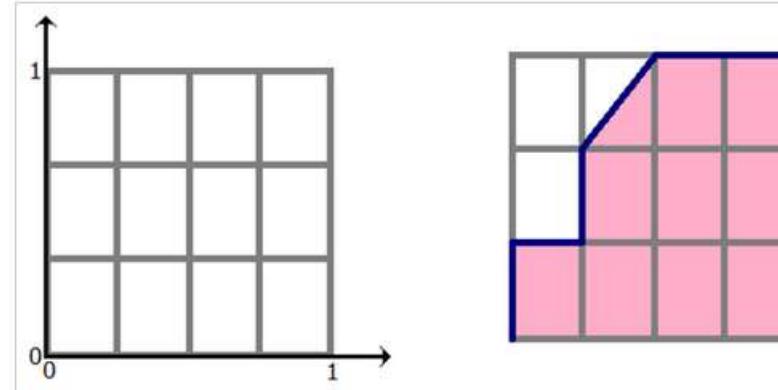
9. ROC-крива. Алгоритм побудови (1)

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2



Кількість рядків = кількості об'єктів класу 1
Кількість колонок = кількості об'єктів класу 0

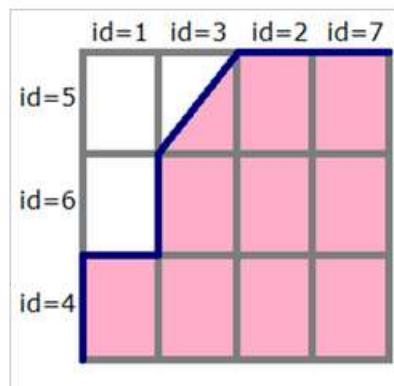
9. ROC-крива. Алгоритм побудови (2)

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2



- Тепер будемо переглядати рядки табл. 2 зверху вниз і промальовувати на сітці лінії, переходячи із одного вузла в інший
- Стартуємо з точки $(0, 0)$
- Якщо значення мітки класу в переглядуємо рядку дорівнює 1, то робимо крок вгору; якщо 0, то робимо крок праворуч. Ясно, що у підсумку ми потрапимо в точку $(1, 1)$

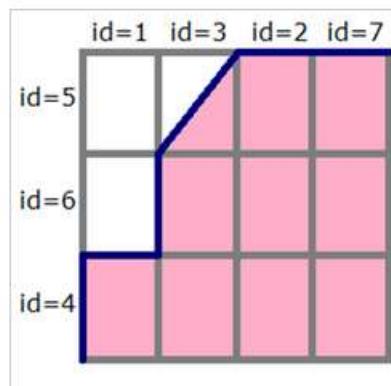
9. ROC-крива. Алгоритм побудови (3)

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2



- Якщо у декількох об'єктів значення оцінок рівні, то ми робимо крок в точку, яка на a блоків вище і b блоків правіше, де a - число одиниць в групі об'єктів з одним значенням мітки, b - число нулів в ній
- Зокрема, якщо всі об'єкти мають однакову мітку, то ми відразу крокуємо з точки $(0, 0)$ в точку $(1, 1)$

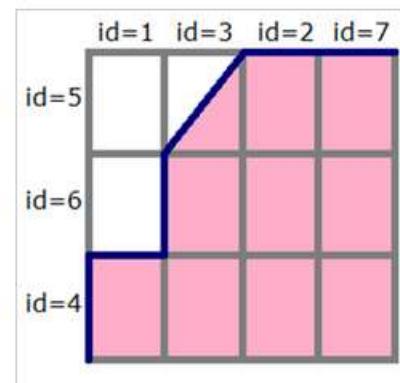
9. ROC-крива. Алгоритм побудови (4)

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2



$$\text{AUC ROC} = 9.5 / 12 \approx 0.79$$

AUC ROC дорівнює частці пар об'єктів виду (об'єкт класу 1, об'єкт класу 0), які алгоритм вірно упорядкував, тобто перший об'єкт йде в упорядкованому списку раніше

9. ROC-крива. Аналітично

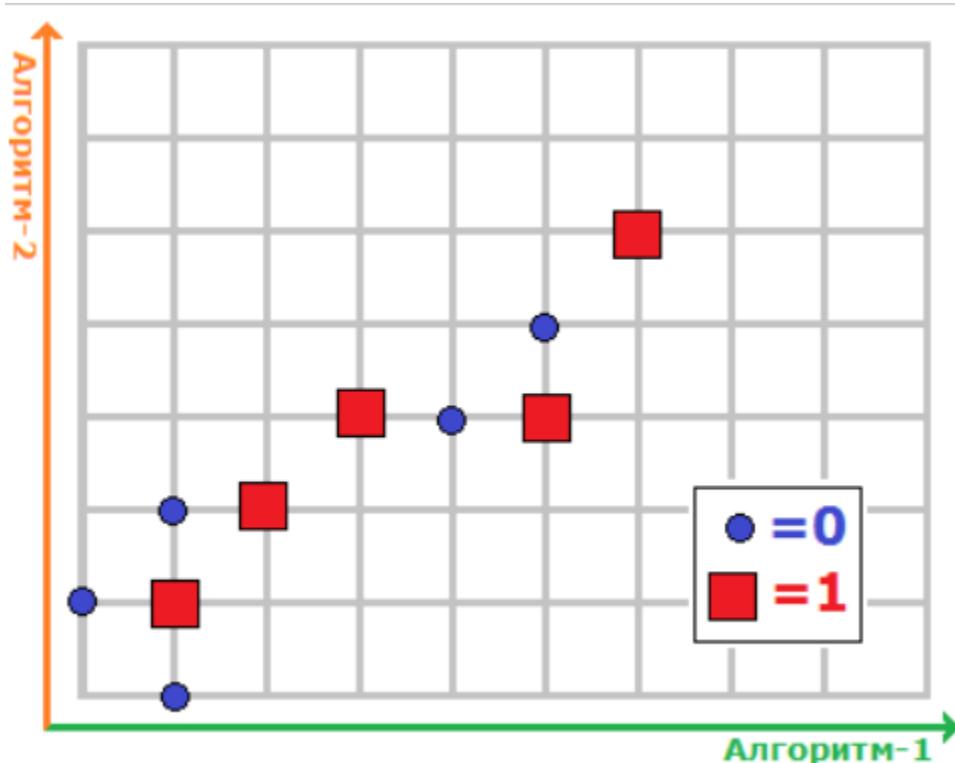
AUC ROC дорівнює частці пар об'єктів виду (об'єкт класу 1, об'єкт класу 0), які алгоритм вірно упорядкував, тобто перший об'єкт йде в упорядкованому списку раніше

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

$$I'[a_i < a_j] = \begin{cases} 0, & a_i > a_j, \\ 0.5 & a_i = a_j, \\ 1, & a_i < a_j, \end{cases} \quad I[y_i < y_j] = \begin{cases} 0, & y_i \geq y_j, \\ 1, & y_i < y_j, \end{cases}$$

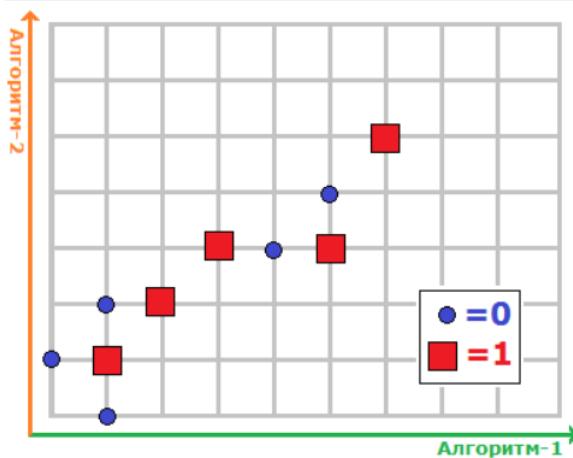
a_i – ответ алгоритма на i -м объекте, y_i – его метка (класс), q – число объектов в тесте.

9. ROC-крива. Приклад як на контрольний

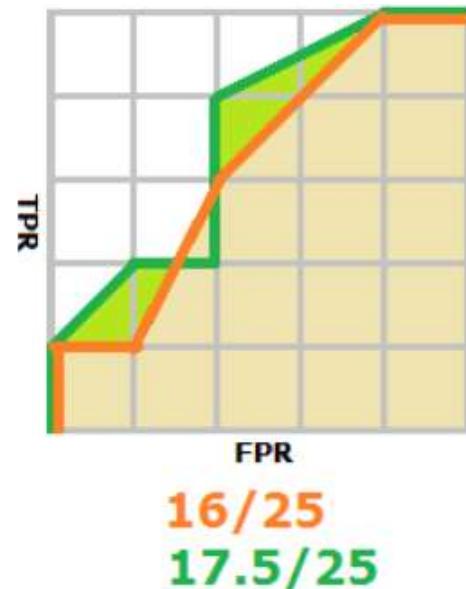


- Задача бінарної класифікації.
- На рис. 1 показані об'єкти у просторі оцінок (=відповідей) двох алгоритмів (оцінки – дійсні числа).
- Потрібно побудувати ROC-криву і обчислити AUC (ROC) для кожного з алгоритмів

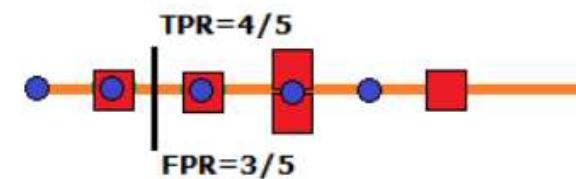
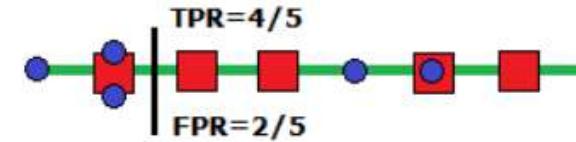
9. ROC-крива. Приклад як на екзамені



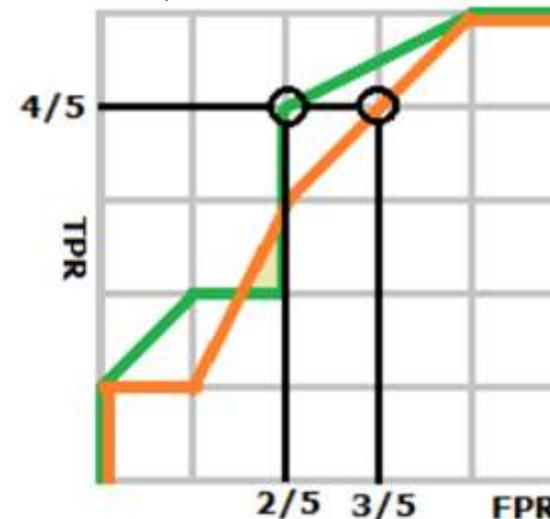
Проектуємо
на осі



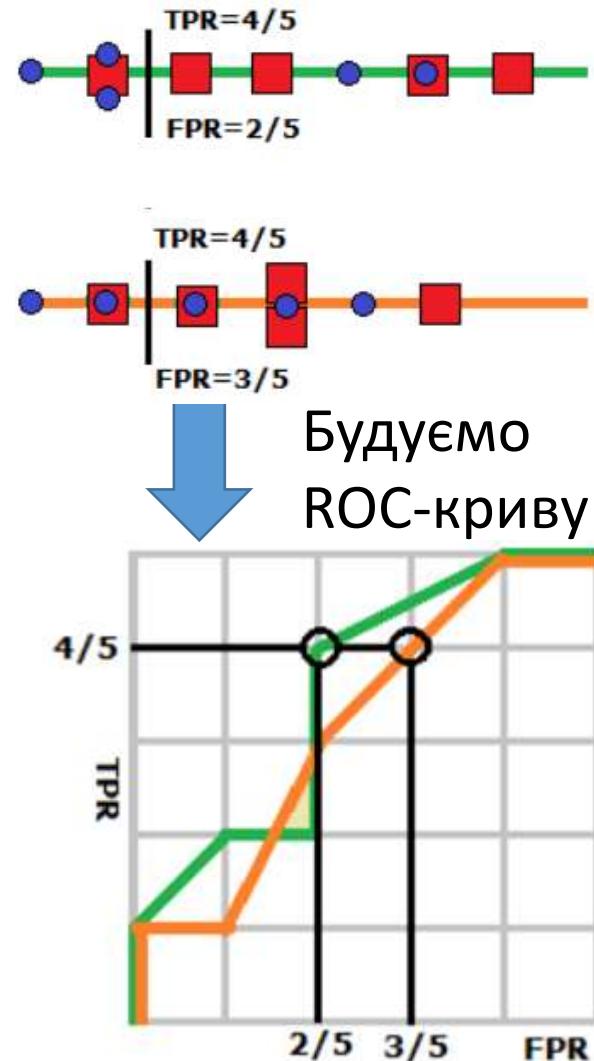
Рахуємо
площу
AUC ROC



Будуємо
ROC-криву



9. ROC-крива. Сенс порога класифікації



$$FPR = 2/5$$

це процент (доля) точок класу 0 («здорових людей»), які неправильно класифіковані нашим алгоритмом («як хворі»)

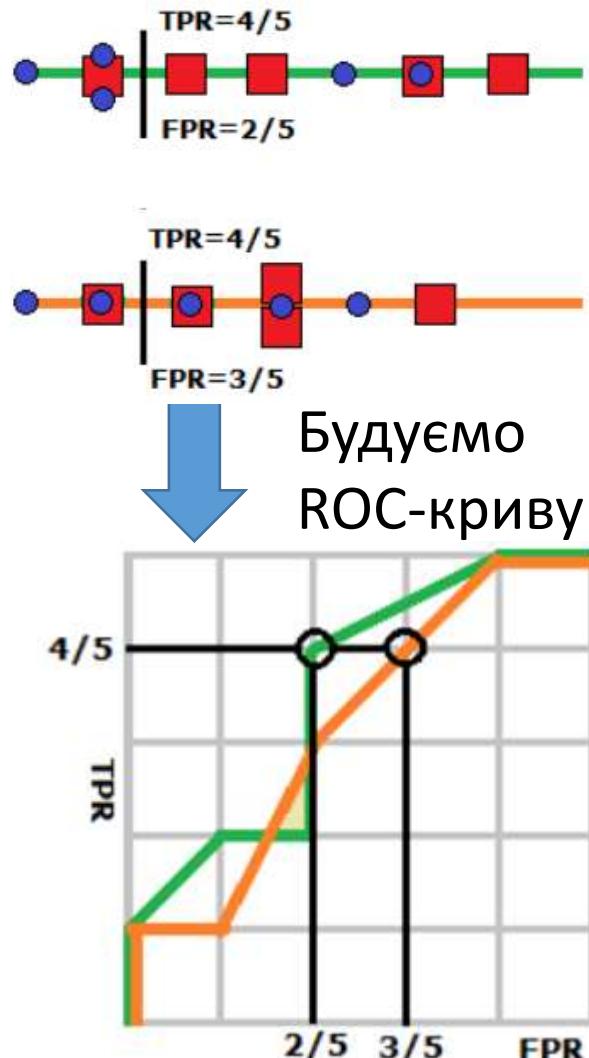
$$FPR = \frac{FP}{FP + TN}.$$

$$TPR = 4/5$$

це процент (доля) точок класу 1 («хворих людей»), які правильно класифіковані нашим алгоритмом («як хворі»)

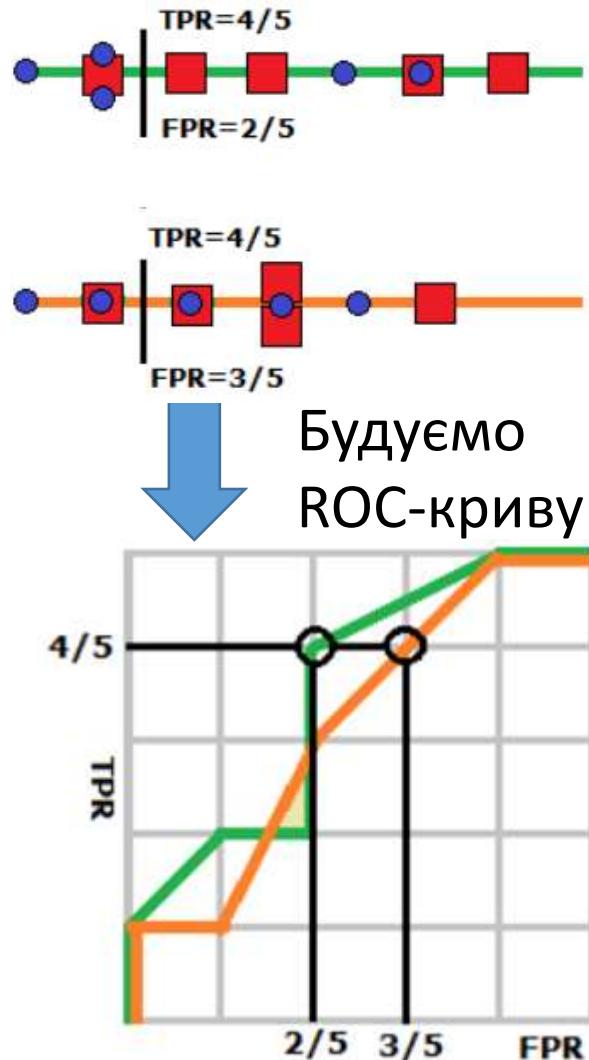
$$TPR = \frac{TP}{TP + FN}$$

9. ROC-крива. Який поріг класифікації вибрати?



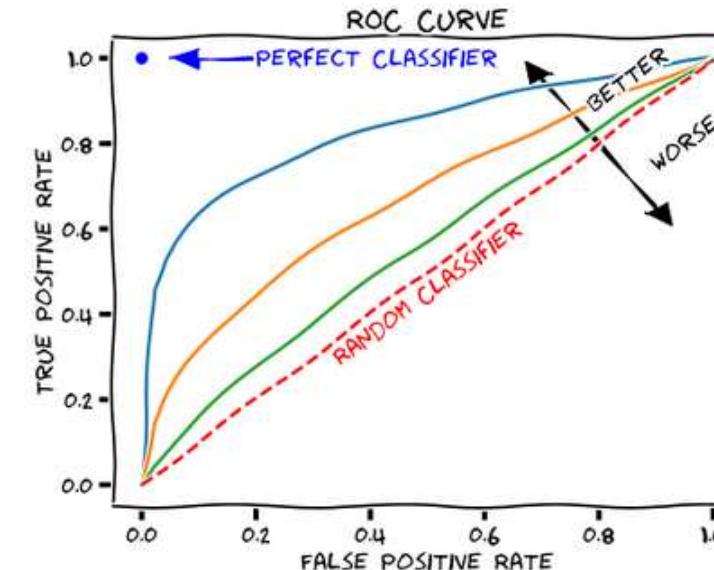
**А як по ROC-кривій
вибрати поріг
класифікації?**

9. ROC-крива. Який поріг класифікації вибрати?



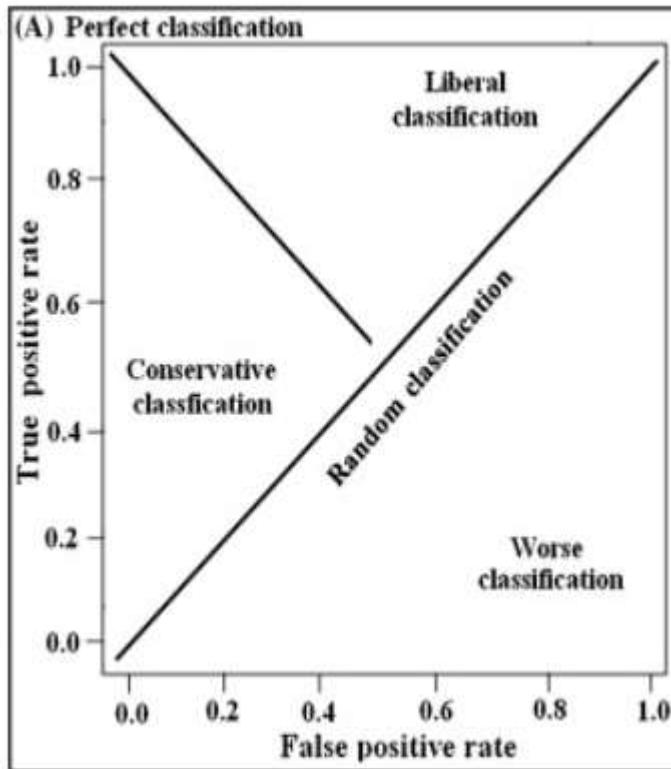
А як по ROC-кривій вибрати поріг класифікації?

Підказка:



9. ROC-крива. Який поріг класифікації вибрati?

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

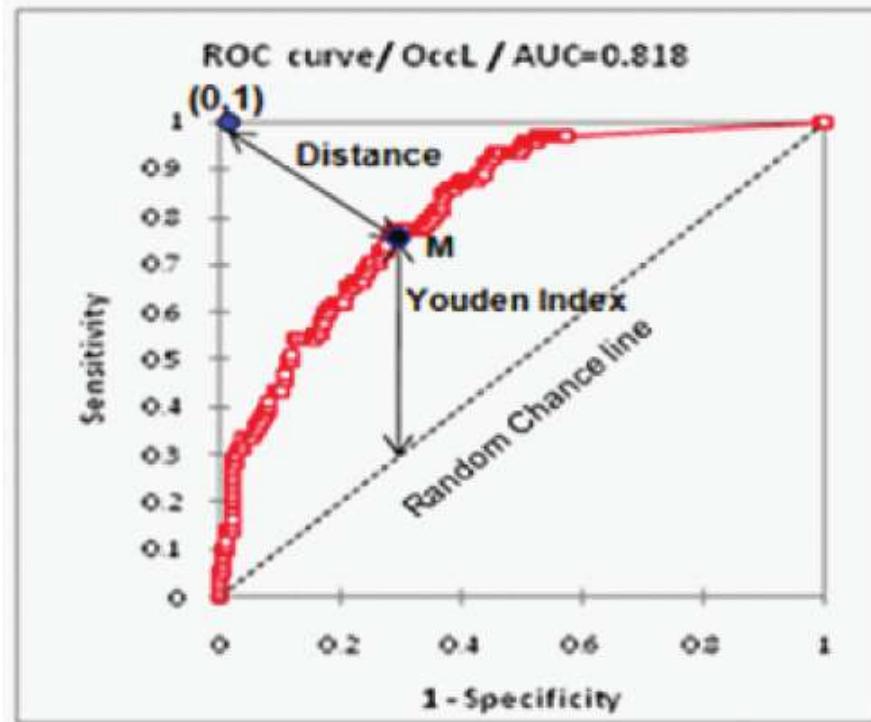


[Kunal Roy, Supratik Kar and Rudra Narayan Das. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, 484 p. (2015) Elsevier Inc.]

- Залежить від того, який результат Ви хочете досягти
- A perfect classification correctly classifies all positive cases and has no false positives
- Most classifiers vary from “conservative” to “liberal”
- A conservative classification (lower-left region of the ROC space) requires strong evidence to classify a point as positive, while a liberal classification (upper-right region of the ROC space) does not require much evidence to classify an event as positive

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

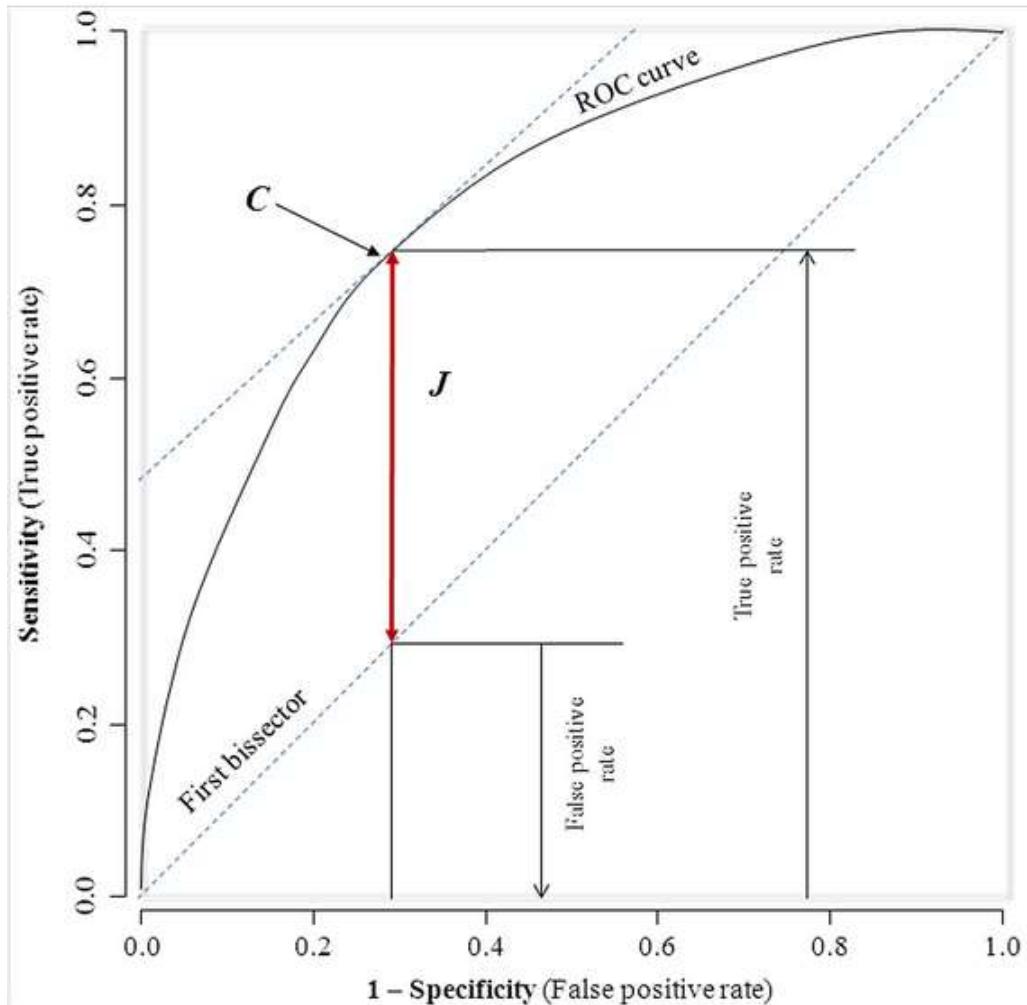
9. ROC-крива. Який поріг вибрати (optimal cut-off value)?



1. Points on Curve Closest to the (0, 1)
2. The goal is to maximize the vertical distance from line of random chance to the point M

[Fkih F, Omri MN (2012) Information retrieval from unstructured web text document based on automatic learning of the threshold. IJIRR 2(4):12'–30]

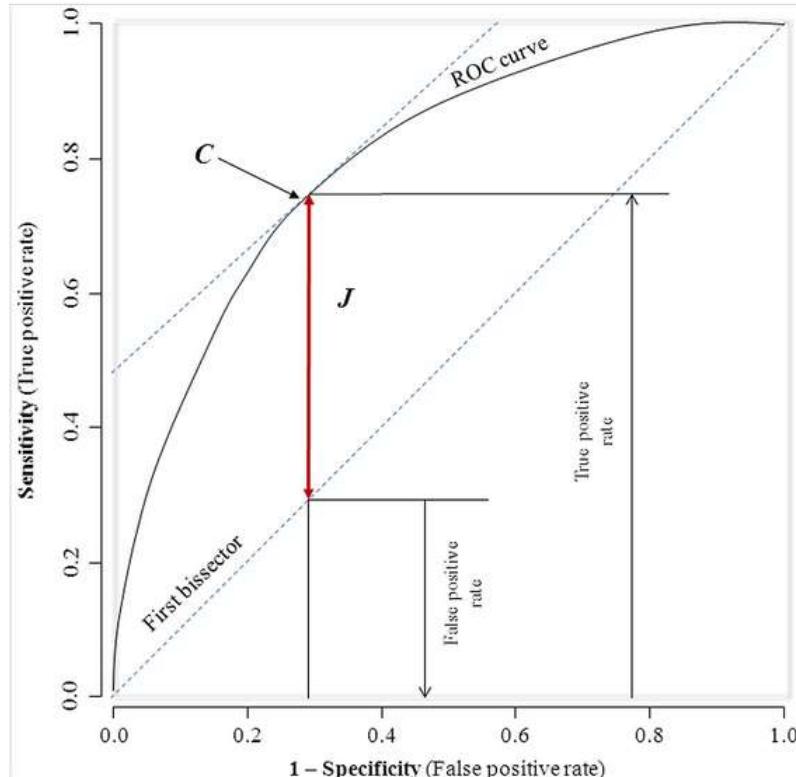
10. Youden's J statistic (Youden's index, індекс Юдена)



- $J = \max(\text{TPR} - \text{FPR})$
- $J = \max(\text{Sensitivity} + \text{Specificity} - 1) = \max\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1\right)$
- The index was suggested by Youden [Youden, W.J. (1950). "Index for rating diagnostic tests". *Cancer*. 3: 32–35] as a way of summarising the performance of a diagnostic test

[https://figshare.com/articles/_Hypothetical_receiver_operating_characteristic_depicting_the_Youden_index_J_and_the_optimal_cut_point_C_/771392]

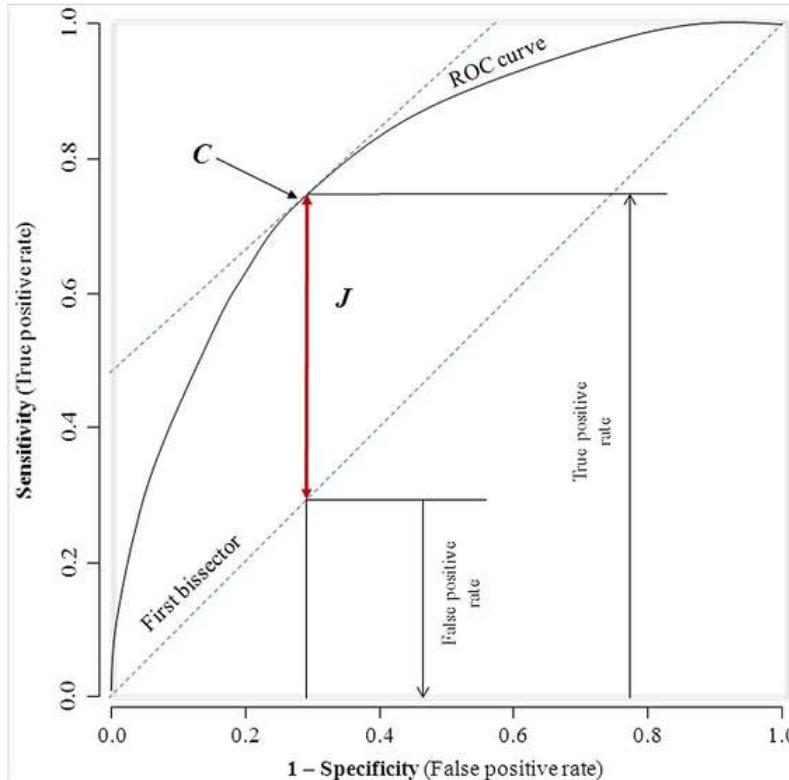
10. Youden's J statistic (Youden's index, індекс Юдена)



$$J = \max\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1\right)$$

- ❑ Its value ranges from 0 through 1, and has 0 when a diagnostic test gives the same proportion of positive results for groups with and without the disease, i.e. the test is useless
- A value of 1 indicates that there are no false positives or false negatives, i.e. the test is perfect
- While it is technically possible to obtain a value of J less than 0, a value of less than zero just indicates that the positive and negative labels have been switched

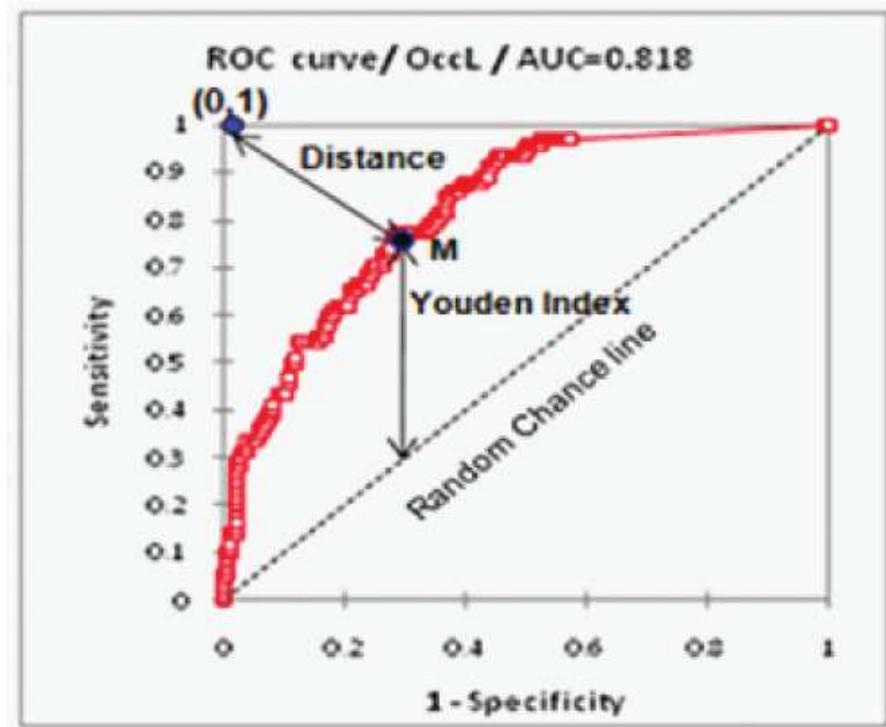
10. Youden's J statistic (Youden's index, індекс Юдена)



$$J = \max\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1\right)$$

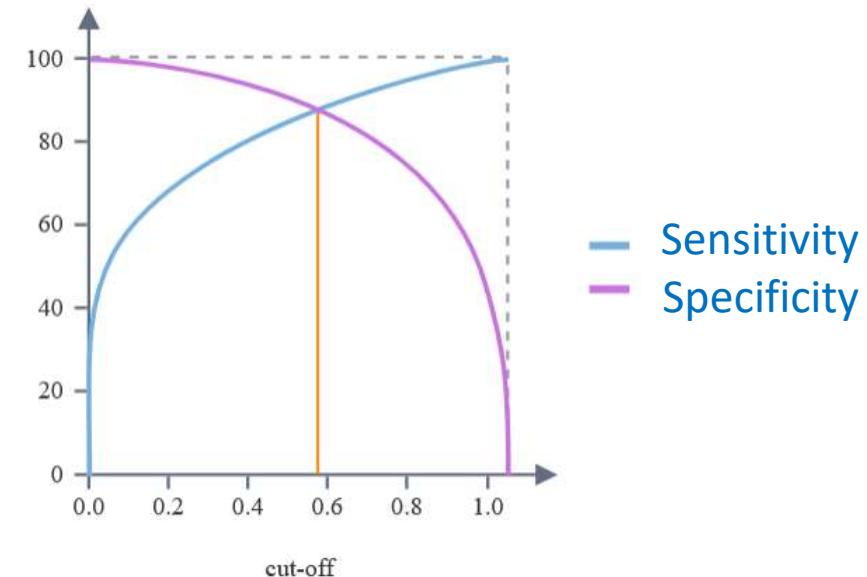
- The index gives equal weight to false positive and false negative values, so all tests with the same value of the index give the same proportion of total misclassified results
- $J = \max (\text{Sensitivity} + \text{Specificity} - 1)$
- $BA = (\text{Sensitivity} + \text{Specificity}) / 2$
- It is now clear why BA works well on unbalanced datasets!

9. ROC-крива. Який поріг вибрати (optimal cut-off value)?

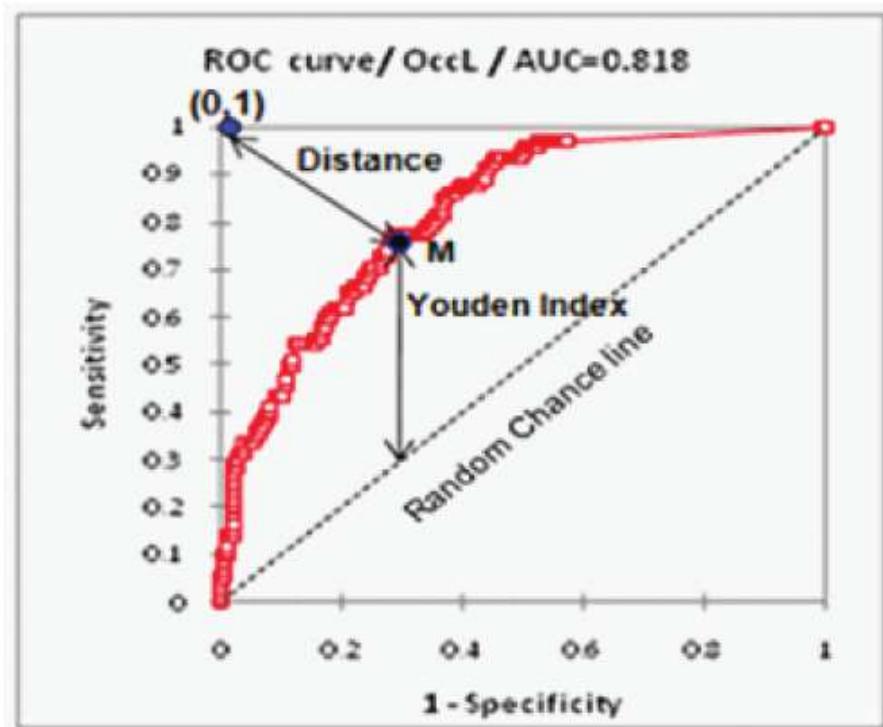


[Fkih F, Omri MN (2012) Information retrieval from unstructured web text document based on automatic learning of the threshold. IJIRR 2(4):12'–30]

3. Точка співпадання Sensitivity та Specificity, тобто критерій:

$$\min |Sensitivity - Specificity|$$


9. ROC-крива. Який поріг вибрати (optimal cut-off value)?



[Fkih F, Omri MN (2012) Information retrieval from unstructured web text document based on automatic learning of the threshold. IJIRR 2(4):12'–30]

4. Вимога мінімально дозволеної величини чутливості / специфічності моделі

- Наприклад, потрібно забезпечити специфічність тесту не менше 90%
- У цьому випадку оптимальному порогу відповідає максимальна чутливість, яка досягається при 90% (чи більшій) специфічності

9. ROC-крива. Який поріг вибрати (optimal cut-off value)?

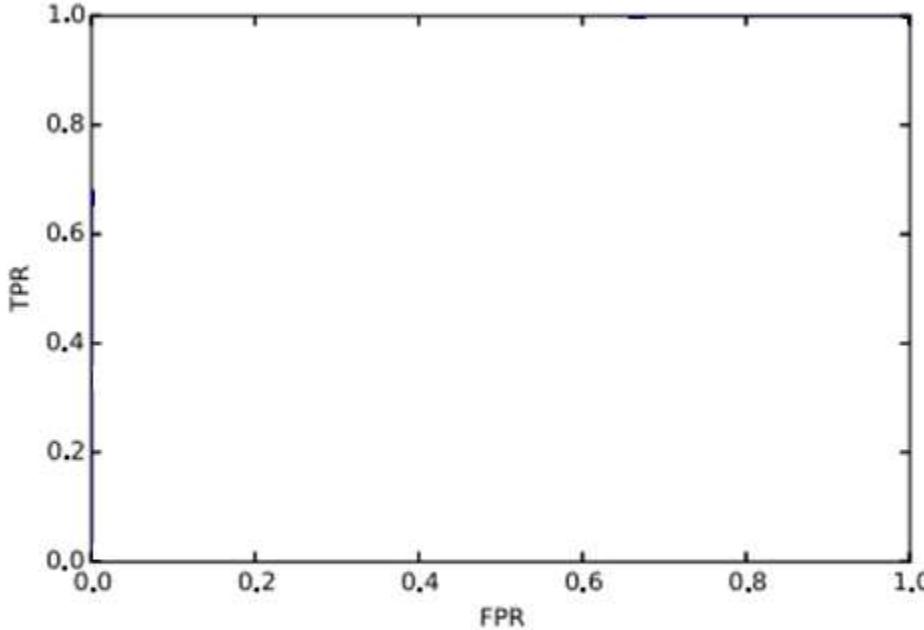
4. Специфіка задачі (знання проблемної області)!

- “Often, the optimal classification threshold is defined as the cut point with the maximum difference between the TPR and FPR (e.g. the Youden Index ...).
- This definition may not be the optimal threshold, depending on the clinical context.
- For example, for a biomarker to be accepted in clinical practice, it must have a better classification performance than the existing test, which has a FPR of 10%. Thus, the optimal threshold in this scenario would be defined as the maximum TPR for an FPR of at the most 10%“

[Chirag R. Parikh, Heather Thiessen Philbrook, Biomarkers of Kidney Disease, (Second edition), Academic Press, 2017]

9. ROC-крива

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

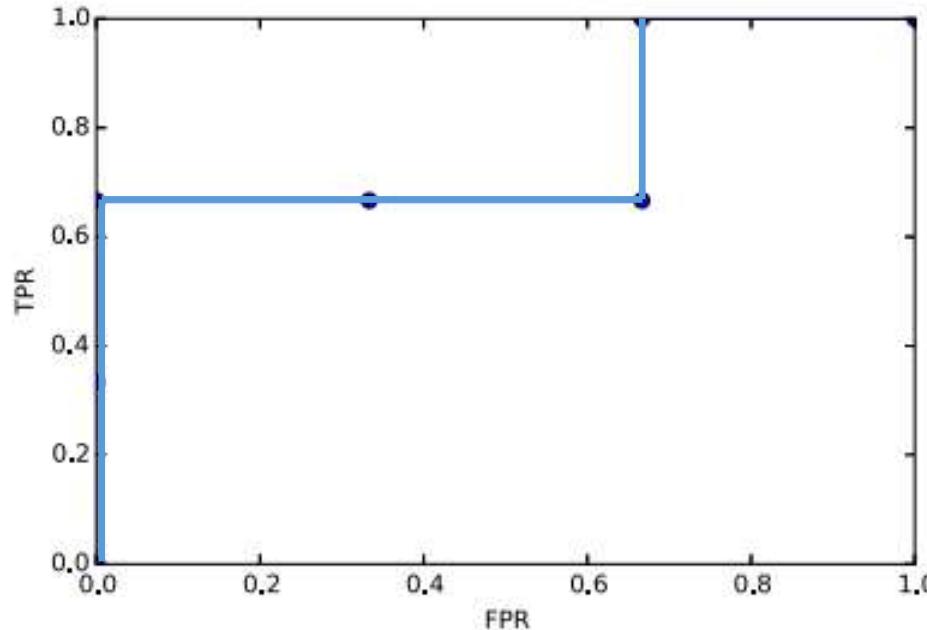


$$FPR = \frac{FP}{FP + TN}$$

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

9. ROC-крива

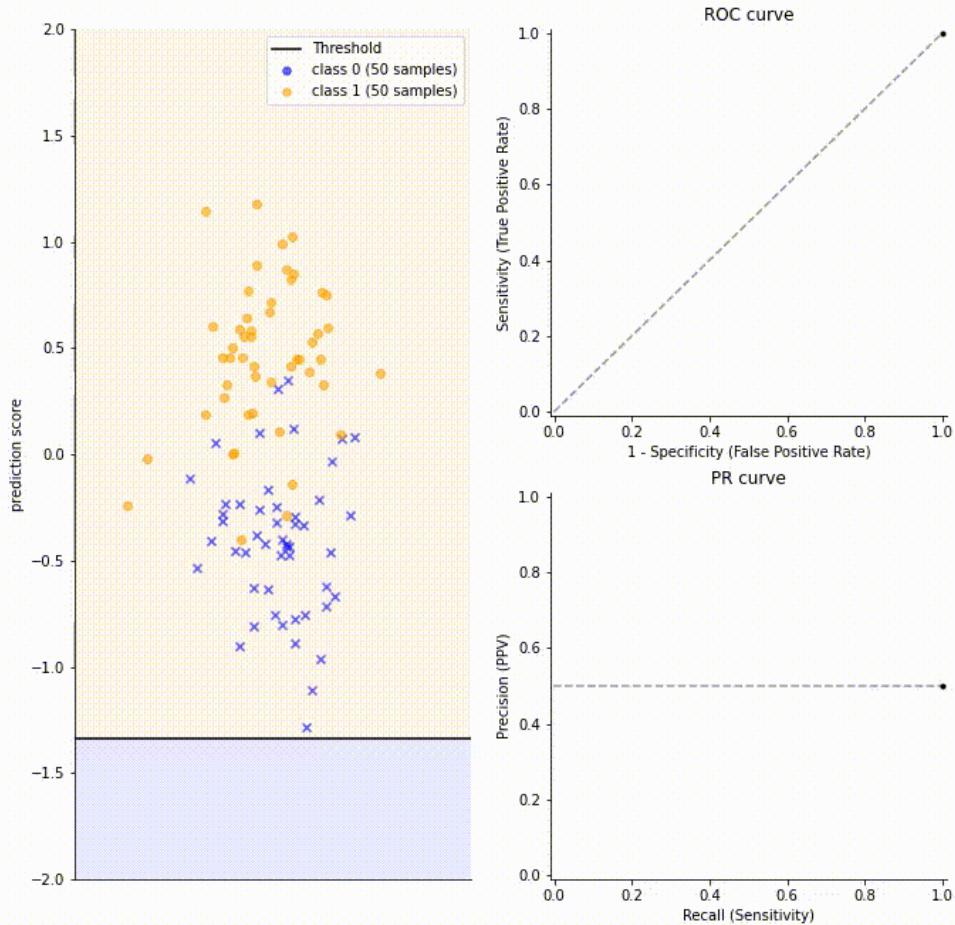
$$TPR = \frac{TP}{TP + FN}$$



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

$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

The Relationship Between Precision-Recall and ROC Curves (1)



- Дано анімація показує динаміку побудови ROC- і PR-кривої в залежності від зміни порогу
- На відміну від розглянутих нами раніше способів, тут криві будуються від меншого значення до більшого
- Для гарного класифікатора ROC-крива більш опукла до верхнього лівого кута, а PR-крива – до верхнього правого кута

<https://ichi.pro/ru/na-krivyh-roc-i-precision-recall-213574150750732>

The Relationship Between Precision-Recall and ROC Curves (2)

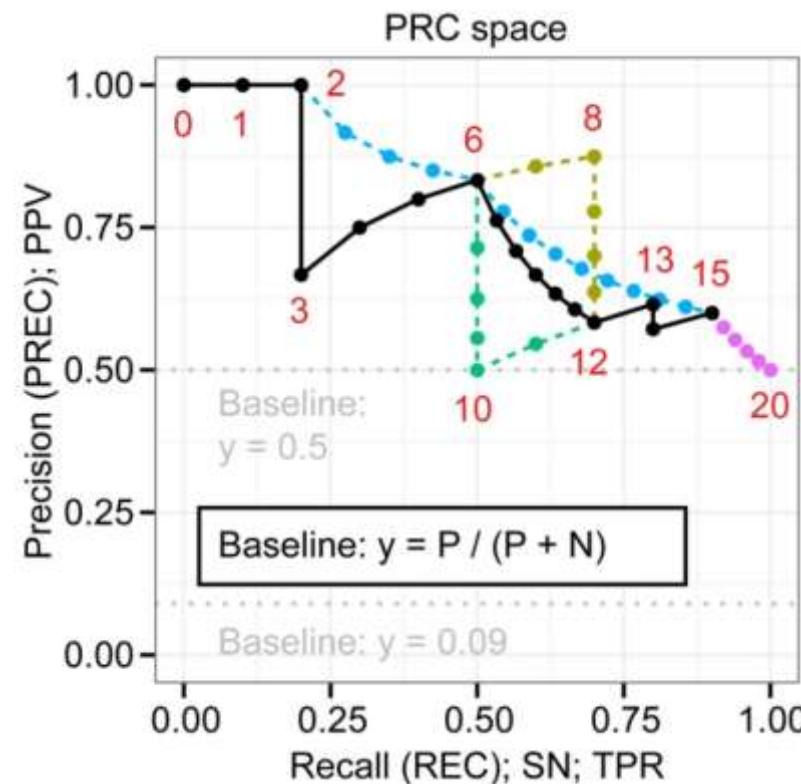
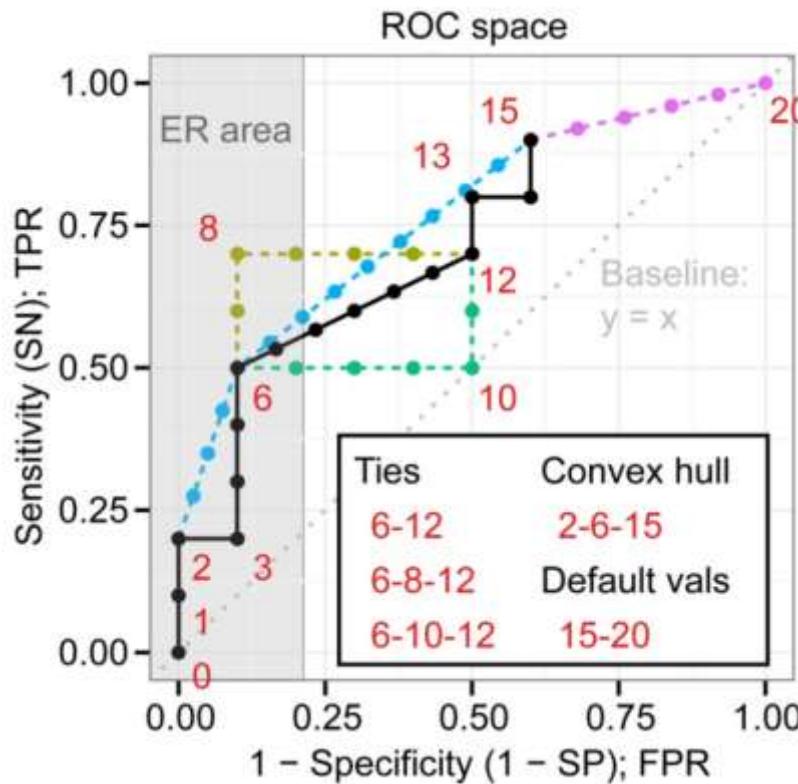
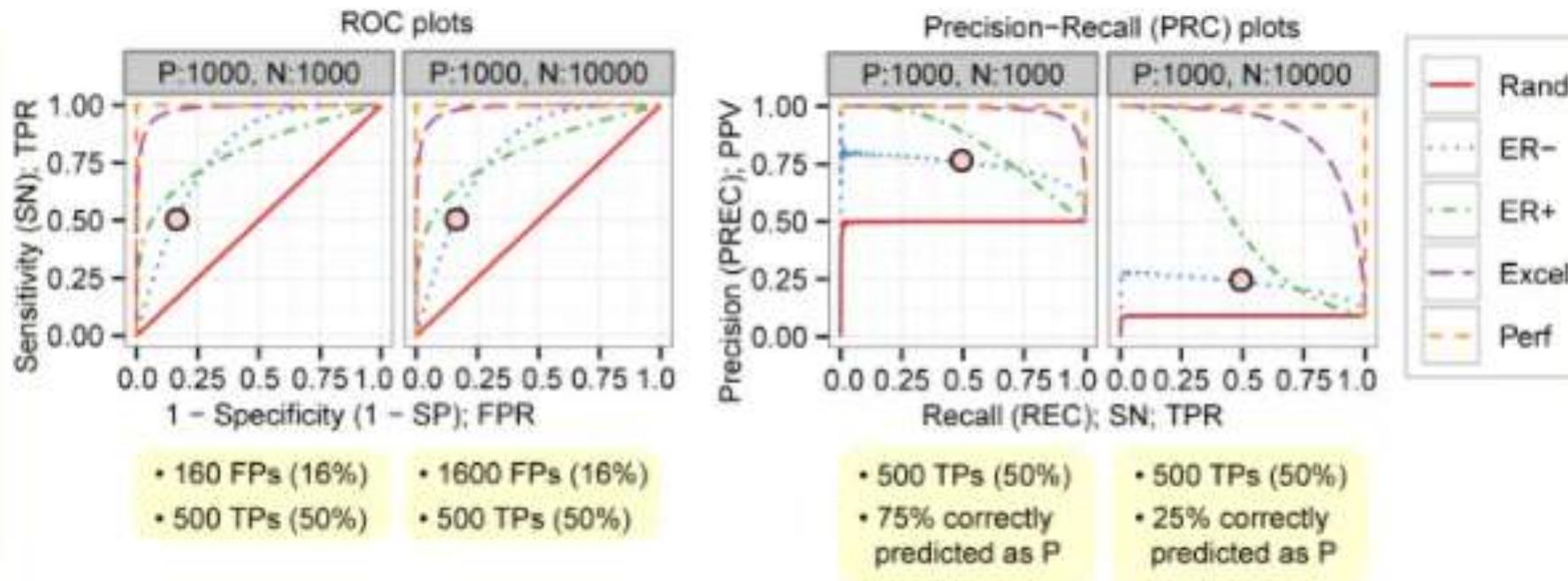


Fig. from [Saito T., Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015; 10:e0118432]

□ Davis J., Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves. // Proceedings of the 23rd Intern. Conf. on ML, Pittsburgh, PA:

- For any dataset, the ROC curve and PR curve for a given algorithm contain the same points
- This equivalence, leads to that a curve dominates in ROC space if and only if it dominates in PR space
- An algorithm that optimizes the area under the ROC curve is not guaranteed to optimize the area under the PR curve

The Relationship Between Precision-Recall and ROC Curves (3)



- ✓ Each panel contains two plots with balanced (left) and imbalanced (right) datasets
- ✓ Five curves represent five different performance levels: Random (Rand; red), Poor early retrieval (ER-; blue), Good early retrieval (ER+; green), Excellent (Excel; purple), and Perfect (Perf; orange)

- Saito T., Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015; 10:e0118432;
- PR curve is changed but ROC curve are unchanged between balanced and imbalanced data (for some cases)

The Relationship Between Precision-Recall and ROC Curves (4)

(розрахунки для ROC-кривої):

Перший алгоритм: вважає 100 листів спамом і в 90 з них не помиляється, тобто
 $TP=90$, $FN=10$, $FP=10$, $TN=999890$

$$\text{Recall} (=TPR) = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0,9$$

$$FPR = \frac{FP}{FP+TN} = \frac{10}{10+999890} = 0,00001$$

Другий алгоритм: вважає 2000 листів спамом і в 90 з них не помиляється, тобто
 $TP=90$, $FN=10$, $FP=1910$, $TN=997990$

$$\text{Recall} (=TPR) = \frac{TP}{TP + FN} = \frac{90}{90 + 10} = 0,9$$

$$FPR = \frac{FP}{FP+TN} = \frac{1910}{1910+997990} = 0,00191$$

Різниця: **0,0019** в FPR, тобто алгоритми майже не відрізняються

(розрахунки для PR-кривої):

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{90}{90 + 10} = 0,9$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{90}{90 + 1910} = 0,045$$

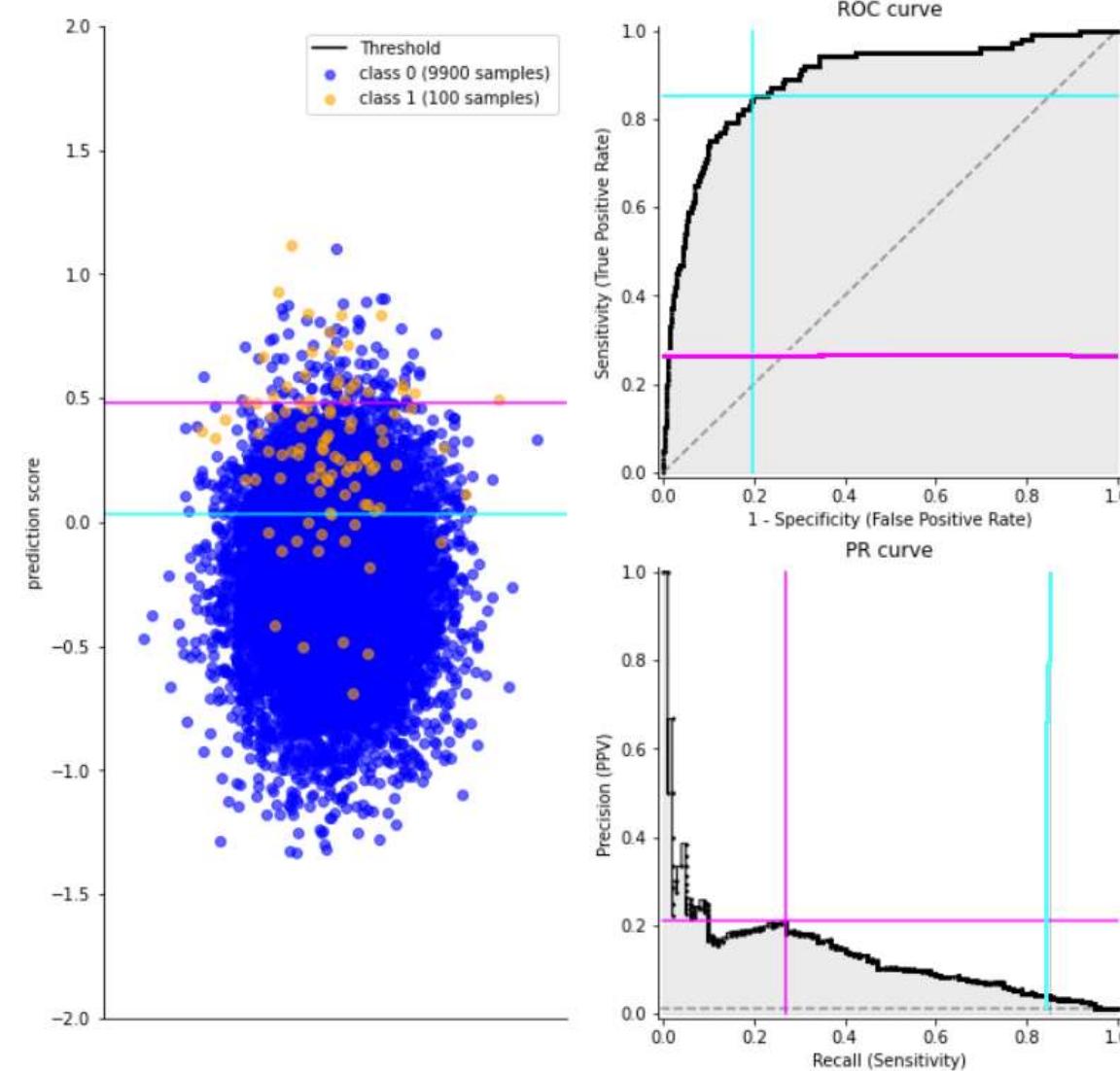
Різниця: **0,855** в точності (precision), тобто перший алгоритм явно краще

□ Якщо набір даних незбалансований, то ROC-крива є малочутливою, бо $TN >> FP$

□ Приклад: необхідно знайти 100 спам-листів з 1 мільйона всіх листів

➤ **Нехай ми відібрали два алгоритми. Який з них вибрати?**

The Relationship Between Precision-Recall and ROC Curves (5)



- Saito T., Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015; 10:e0118432:
 - якщо набір даних незбалансований, то ROC-крива є занадто оптимістичною
- Яку криву використовувати: ROC чи PR – залежить від розв'язуваної задачі:
 - якщо ми маємо справу з виявленням шахрайства, то краще застосовувати криву PR, бо вона дає менше хибних спрацьовувань (не будемо перевантажувати службу перевірки)
 - якщо ми маємо справу з виявленням раку, то краще застосовувати криву ROC, бо краще більше перевіряти, але бути впевненим, що не пропустили смертельну хворобу

ROC & PR curves for multiple class classification problems

- David J. Hand, Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Language*, Volume 45, Issue 2, November 2001, pp. 171–186

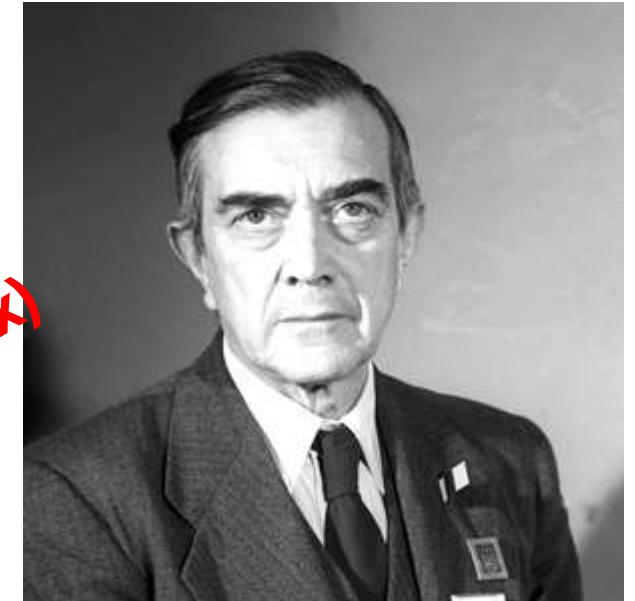
<https://doi.org/10.1023/A:1010920819831>

- “We extend the definition to the case of more than two classes by averaging pairwise comparisons. This measure reduces to the standard form in the two class case.
- We compare its properties with the standard measure of proportion correct and an alternative definition of proportion correct based on pairwise comparison of classes for a simple artificial case and illustrate its application on eight data sets.
- On the data sets we examined, the measures produced similar, but not identical results, reflecting the different aspects of performance that they were measuring.
- Like the area under the ROC curve, the measure we propose is useful in those many situations where it is impossible to give costs for the different kinds of misclassification”

11. Gini coefficient (Gini index)

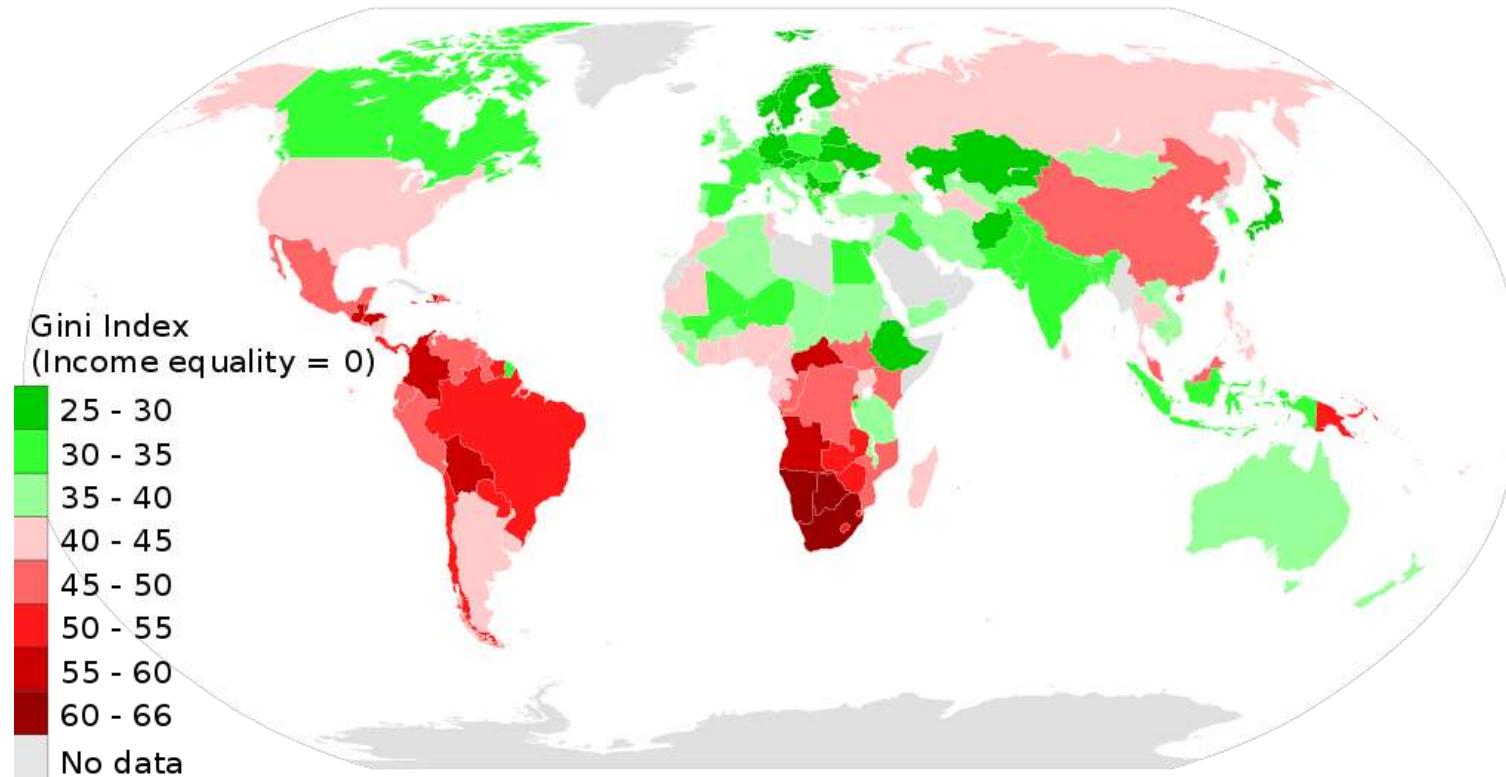
- Коефіцієнт Джині – показник нерівності розподілу деякої величини, що приймає значення між 0 і 1, де 0 означає абсолютну рівність (величина приймає лише одне значення), а 1 позначає повну нерівність [Джіні. Варіативність та мінливість ознаки (1912)]
- Найбільш відомим коефіцієнтом є як міра нерівності доходів домогосподарств деякої країни чи регіону
- Коефіцієнт Джині для доходів домогосподарств є найпопулярнішим показником економічної нерівності в країні

НЕ ПЛУТАТИ
з неоднорідністю
(індексом) Джині,
Gini impurity (index)



Corrado Gini (1884 – 1965), італійський статистик і соціолог, автор статті «Наукові основи фашизму»

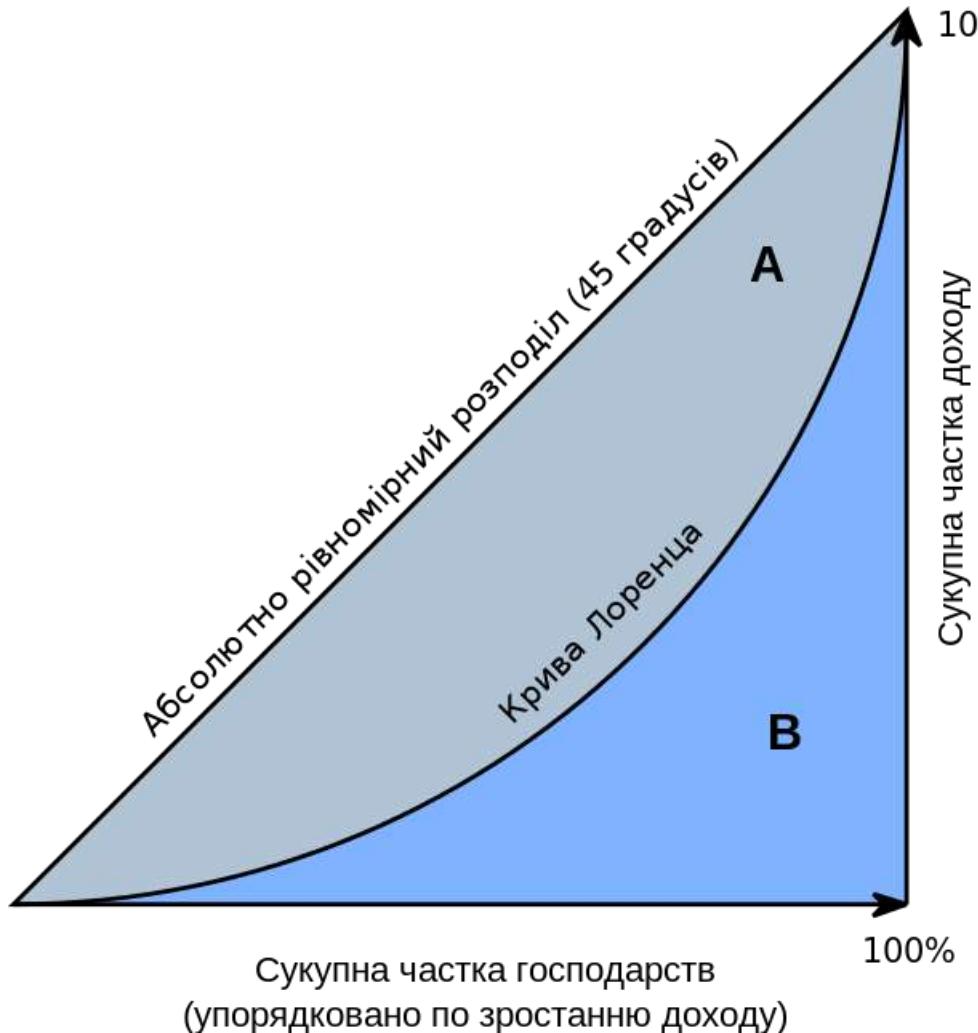
11. Gini coefficient (Gini index)



Індекс Джині розподілу доходу для країн світу (згідно з даними 2014 року)

Індекс Джині = коефіцієнт Джині у відсотках

11. Gini coefficient (Gini index). Графічне представлення



Індекс Джині найпростіше визначити за допомогою **кривої Лоренца** [Макс Лоренц. Методи вимірювання концентрації багатства (1905)], що зображує частку величини u , що зосереджується на $x\%$ популяції з найменшим значенням цієї величини

Наприклад для розподілу доходів точка (20%, 10%) буде лежати на кривій Лоренца, якщо сукупний дохід 20% найбідніших домогосподарств рівний 10% сукупного доходу усіх домогосподарств

11. Gini coefficient (Gini index). Графічне представлення

А де крива абсолютно нерівномірного розподілу?

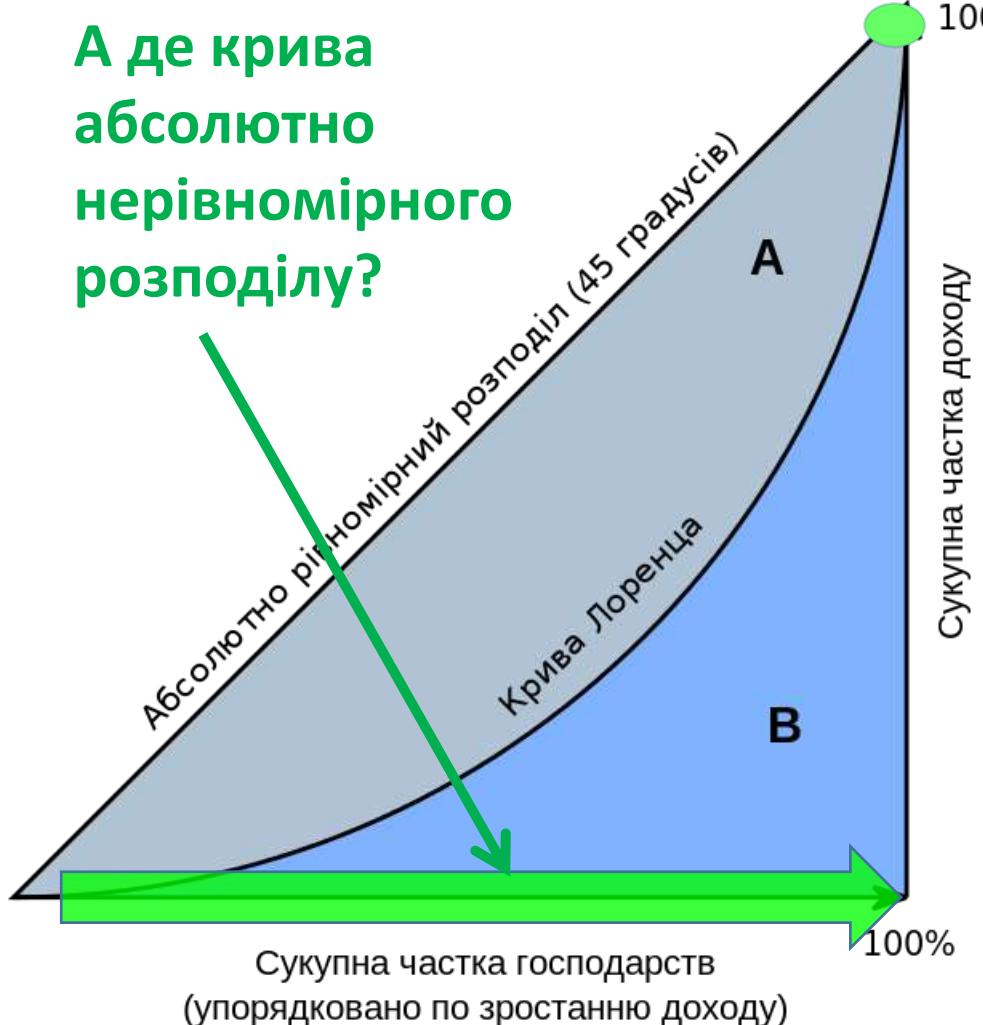


Індекс Джині найпростіше визначити за допомогою **кривої Лоренца** [Макс Лоренц. Методи вимірювання концентрації багатства (1905)], що зображує частку величини u , що зосереджується на $x\%$ популяції з найменшим значенням цієї величини

Наприклад для розподілу доходів точка $(20\%, 10\%)$ буде лежати на кривій Лоренца, якщо сукупний дохід 20% найбідніших домогосподарств рівний 10% сукупного доходу усіх домогосподарств

11. Gini coefficient (Gini index). Графічне представлення

А де крива абсолютно нерівномірного розподілу?



Індекс Джині найпростіше визначити за допомогою **кривої Лоренца** [Макс Лоренц. Методи вимірювання концентрації багатства (1905)], що зображує частку величини u , що зосереджується на $x\%$ популяції з найменшим значенням цієї величини

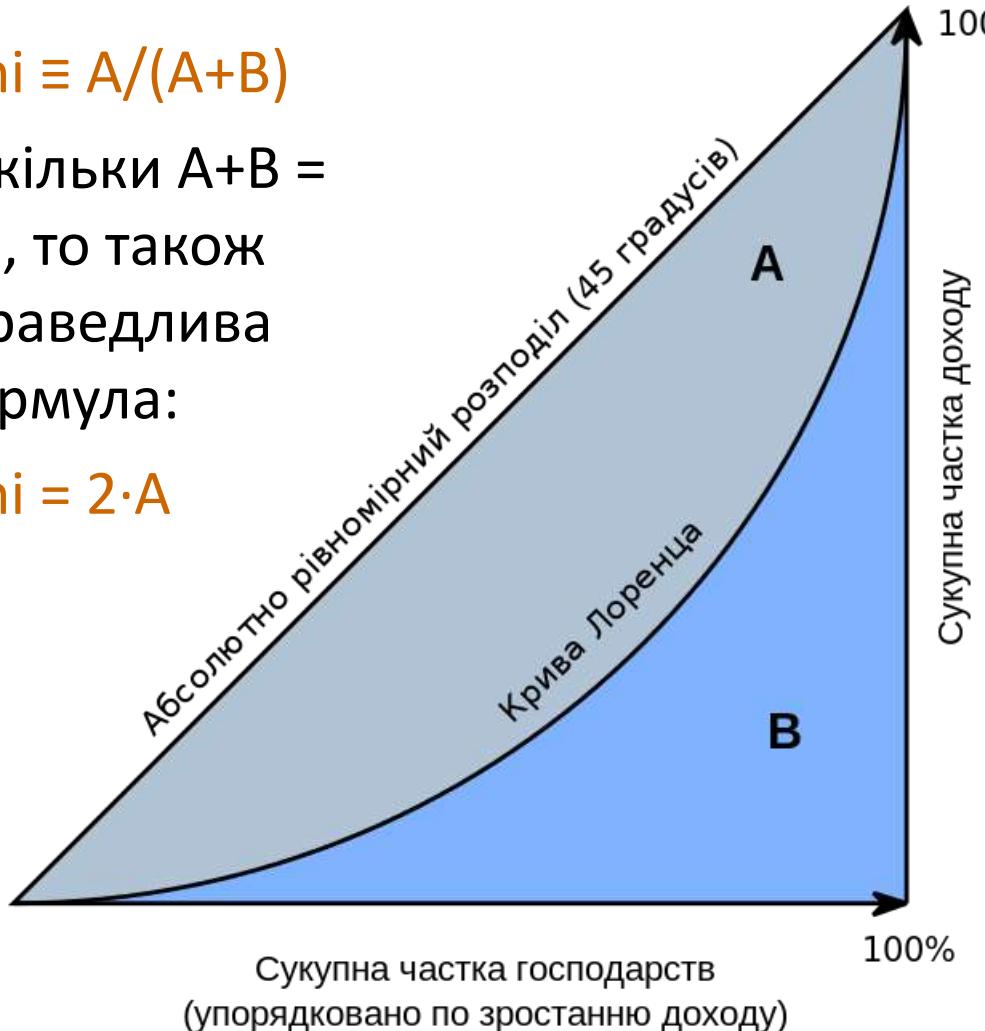
Наприклад для розподілу доходів точка (20%, 10%) буде лежати на кривій Лоренца, якщо сукупний дохід 20% найбідніших домогосподарств рівний 10% сукупного доходу усіх домогосподарств

11. Gini coefficient (Gini index). Графічне представлення

$$\text{Gini} \equiv A/(A+B)$$

Оскільки $A+B = 0.5$, то також справедлива формула:

$$\text{Gini} = 2 \cdot A$$



Індекс Джині найпростіше визначити за допомогою кривої Лоренца, що зображує частку величини у, що зосереджується на $x\%$ популяції з найменшим значенням цієї величини.

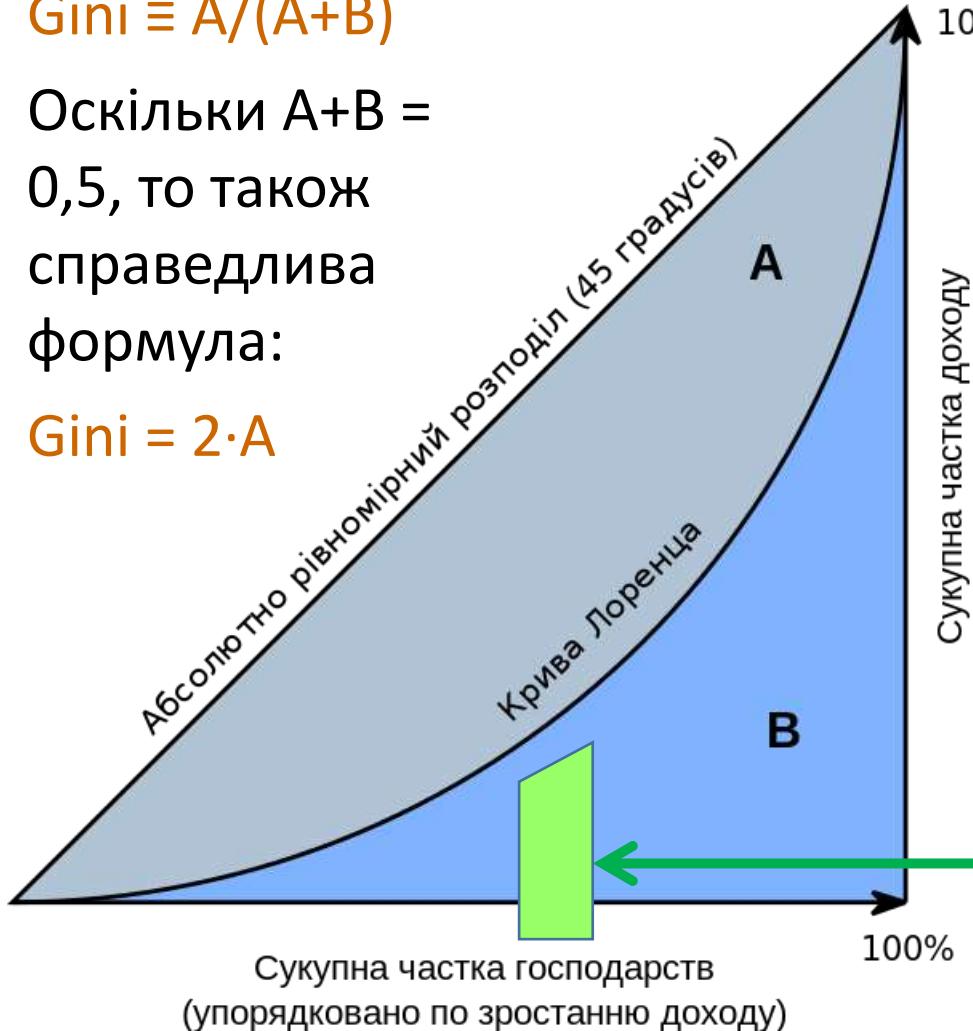
Наприклад для розподілу доходів точка (20%, 10%) буде лежати на кривій Лоренца, якщо сукупний доход 20% найбідніших домогосподарств рівний 10% сукупного доходу усіх домогосподарств

11. Gini coefficient (Gini index). Формула підрахунку Брауна

$$\text{Gini} \equiv A/(A+B)$$

Оскільки $A+B = 0,5$, то також справедлива формула:

$$\text{Gini} = 2 \cdot A$$



$$G = 1 - \sum_{k=1}^n (X_k - X_{k-1})(Y_k + Y_{k-1})$$

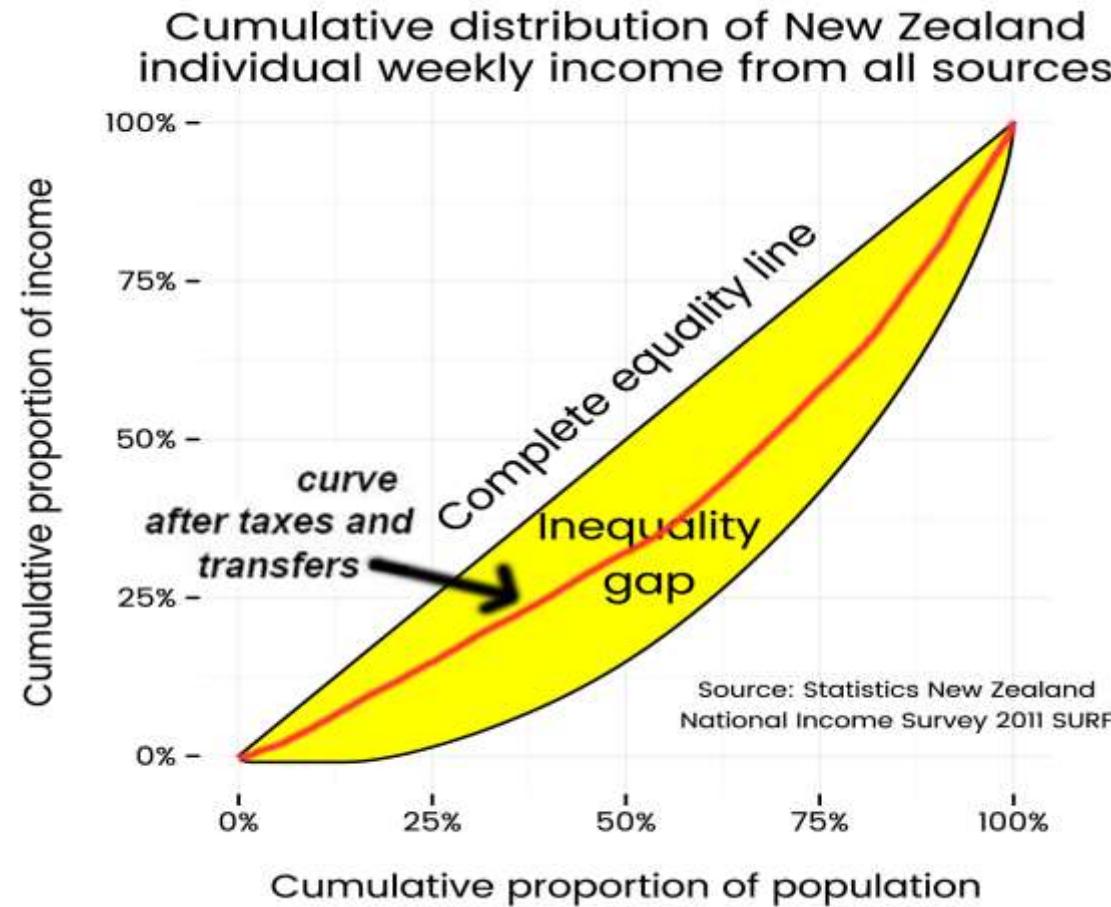
де n – кількість господарств,

X_k – кумулятивна доля господарств,

Y_k – кумулятивна доля доходу для X_k

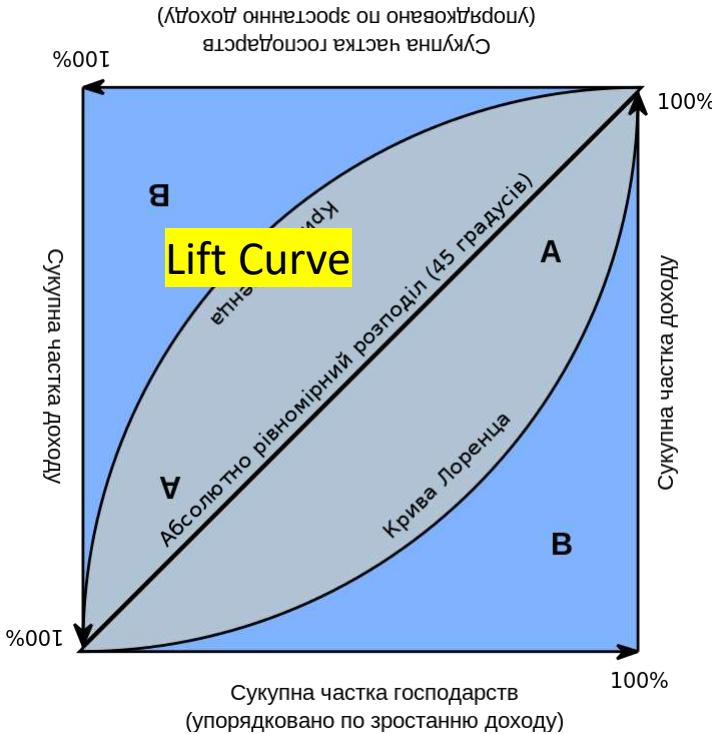
Чому приблизно дорівнює площа криволінійної трапеції?

11. Gini coefficient (Gini index)



Harel Lustiger (October, 2017). The Gini Coefficient under Binary Classification Problems

11. Gini coefficient (Gini index). Зв'язок з AUC(ROC)



The **Gini Coefficient** is $2 \cdot \text{AUC} - 1$, and its purpose is to normalize the AUC so that a random classifier scores 0, and a perfect classifier scores 1. The range of possible Gini coefficient scores is $[-1, 1]$.

$\text{Gini} \equiv A/(A+B)$. Оскільки $A+B = 0,5$, то також справедлива формула: $\text{Gini} = 2 \cdot A$

В задачі класифікації на класи 0 і 1, ці числа можна інтерпретувати як доходи \Rightarrow отримаємо криву Лоренца

Якщо ранжувати господарства не по зростанню доходів, а по спаданню, то крива Лоренца відобразиться симетрично головній діагоналі в Lift Curve (\neq ROC-кривій, але близька)

Площа під нею дорівнює:
 $\text{AUC}(\text{ROC}) = A+0,5 = (\text{Gini}+1)/2$
Тобто $\text{Gini} = 2 \cdot \text{AUC}(\text{ROC}) - 1$