Машинное обучение

День 2. Классификация.

ПЛАН

День 2. Классификация.

01

ОСНОВЫ
10 МИН

Формальная постановка задачи ML.
Основные понятия и проблемы.

МЕТОДЫ
15 МИН

02

Методы классификации: линейные методы, решающие деревья.

03

ПЛАН РЕШЕНИЯ ML-ЗАДАЧ 5 МИН

Как организован процесс решения ML-задачи.
В первом приближении.

ПРАКТИКА
10 МИН

04

Решим 1 задачу по плану методами пакета python -- sklearn.

01 ОСНОВЫ 10 МИН

Постановка задачи и признаки

X – множество объектов / features
Y – целевое значение / target
f – решающая функция или алгоритм ML

$$f(X) = Y$$

X =

F1	F2	F3	F4	F5
14.3	4	type1	127	2.8
13.35	5	type2	100	2.65
12	1	type1	101	2.8
17.1	4	type1	113	3.85

Y =

class
1
2
3
2

Типы признаков:

- количественные
- категориальные

Кодирование категориальных признаков:

- OneHotEncoding

Sex
male
female

→

F1	F2
0	1
1	0

- LabelEncoding

Sex
male
female

→

F1
0
1

01 ОСНОВЫ 10 МИН Выборки и метрики

Этапы:

- обучение (train)
- тестирование (test)

Разделение данных:

- обучение (train) – 80%
- тестирование (test) – 20%

Метрика

Метрика – это число.

- Это показатель того, насколько хорошо работает наш алгоритм и какая у него обобщающая способность.
- Метрика считается только на тестовой выборке

Accuracy

Метрика доли верно угаданных ответов (accuracy).

Y_true	Y_pred
1	1
1	0
0	0
1	0
1	1

$$Accuracy = \frac{\sum_{i=1}^{N}[Y_{true}(i) == Y_{pred}(i)]}{N}$$

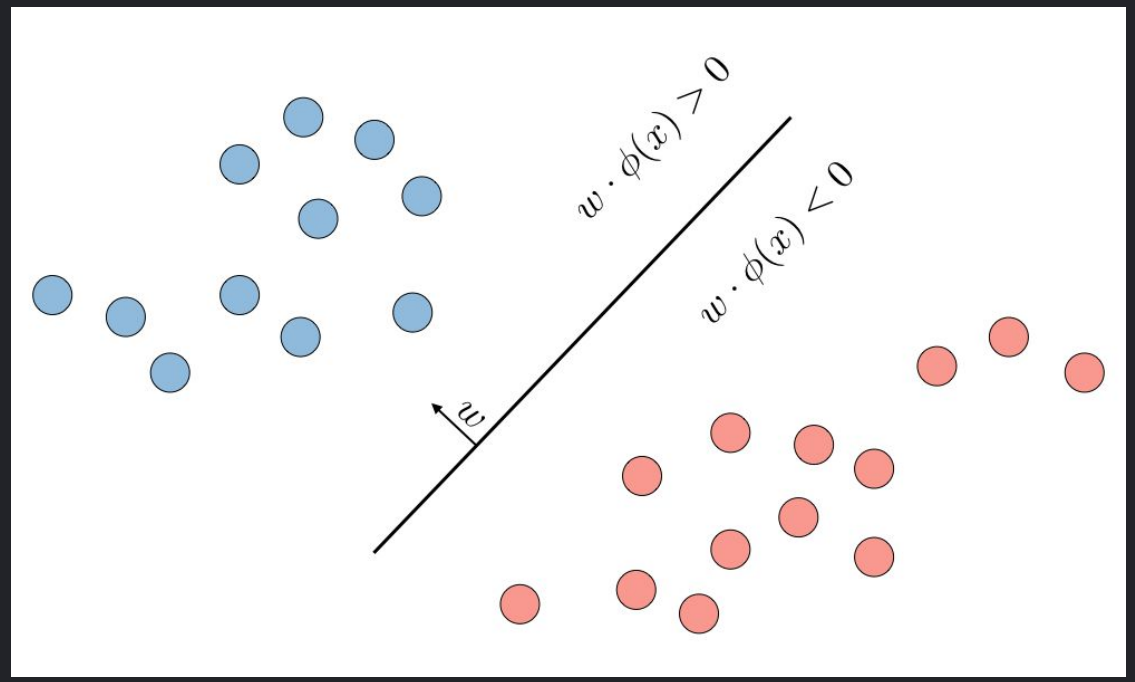
$$= 0.6$$

02 МЕТОДЫ
15 МИН

Типы алгоритмов классификации*

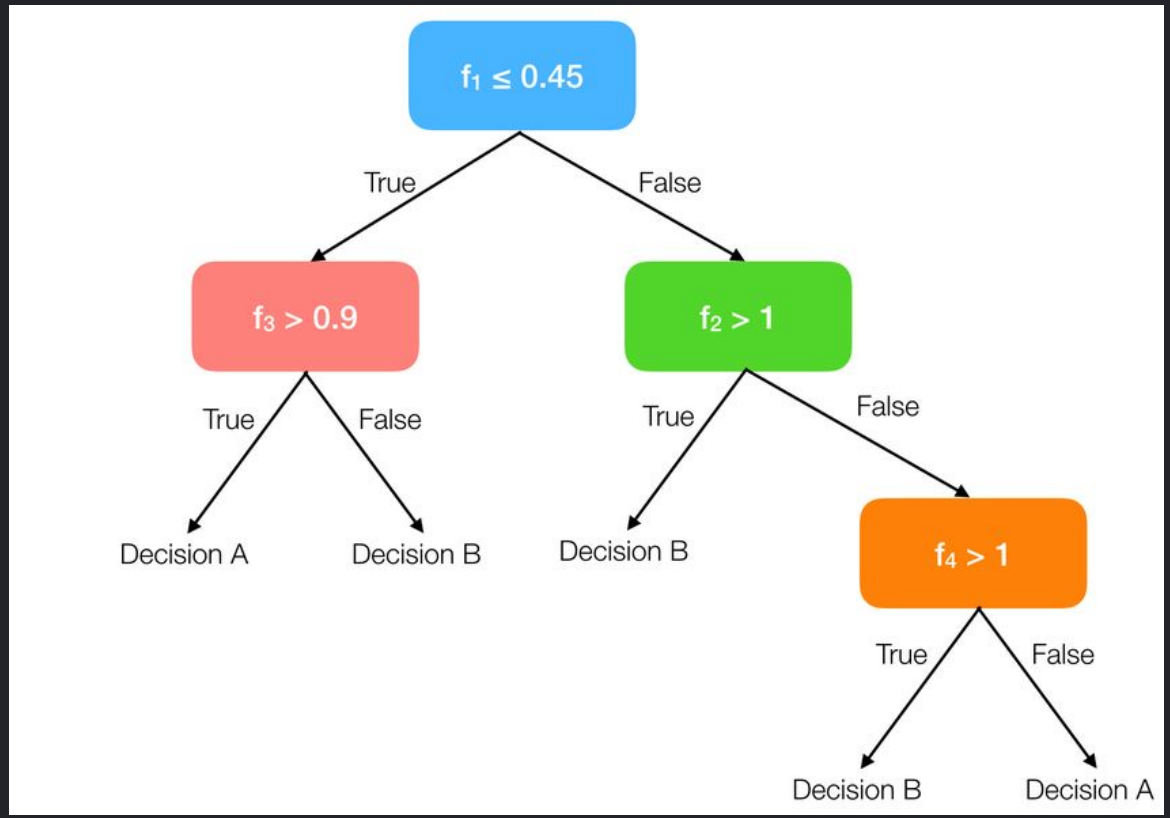
Линейные методы

$$Y = \langle X \cdot W \rangle$$



SGD, SVM, Logistic regression, etc.

Деревья



Decision tree classifier, Decision tree regressor, Random Forest, etc.

*существуют и другие типы, но мы остановимся только на этих

Линейные методы

Дано:
X – матрица объектов-признаков, $x_i \in R^n$
Y – вектор ответов, $y_i \in \{+1, -1\}$

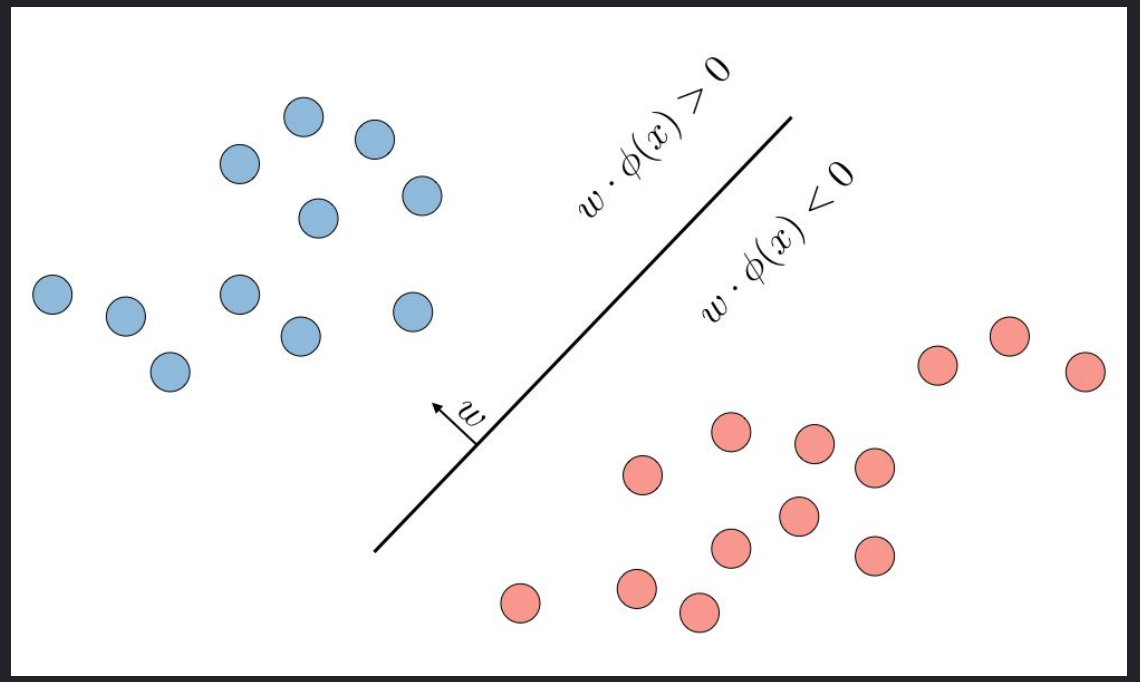
Постановка задачи бинарной классиф.:
$Y = f(X \cdot w) = sign(\langle X \cdot w \rangle)$

Найти:
w – ?

w – вектор весов, $w_i \in R^n$
$\langle X \cdot w \rangle$ – скалярное произведение

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots \\ x_{2,1} & x_{2,2} & \cdots \\ \vdots & \vdots & \ddots \\ x_{n,1} & x_{n,2} & \cdots \end{pmatrix}$$

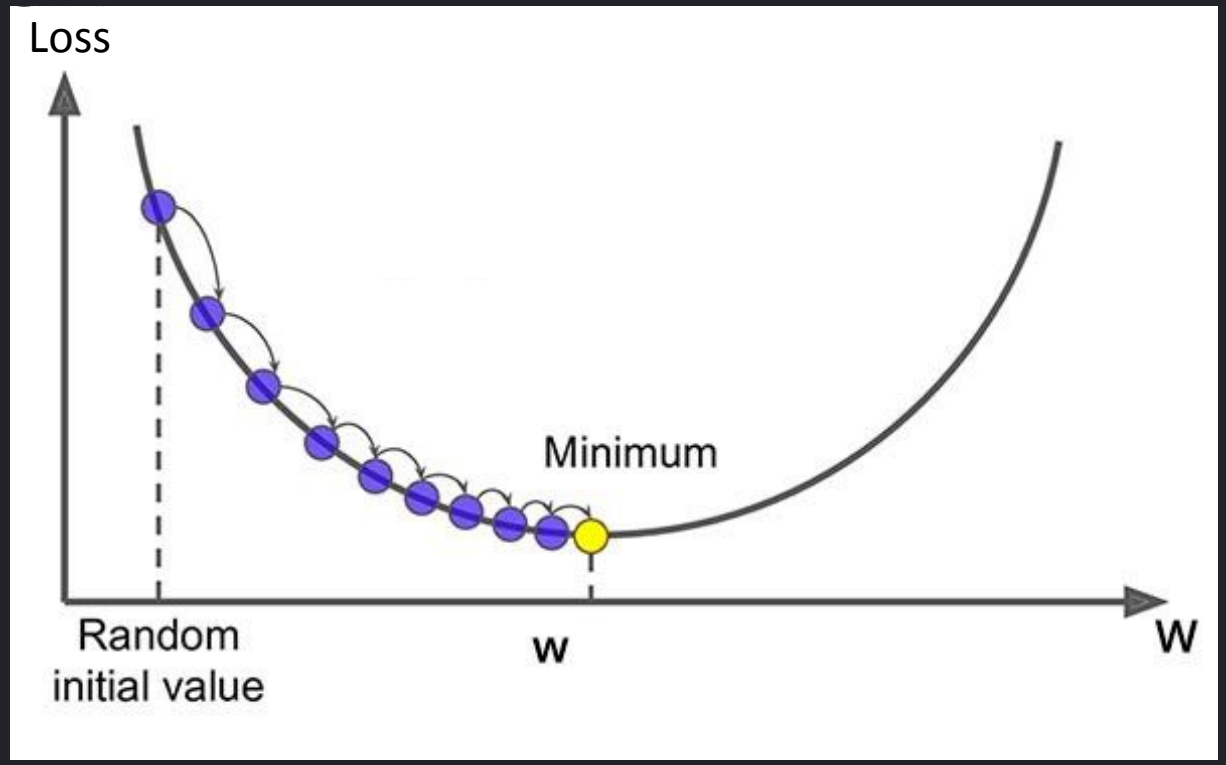
X w Y



02 МЕТОДЫ
15 МИН

Линейные методы

Решение задачи – поиск вектора весов.
Алгоритм решения – метод градиентного спуска.



Loss (loss-function, функция потерь) может задаваться по-разному.

Все линейные методы устроены одинаковы.
У них только отличается функция потерь (loss).

Популярные линейные методы:

- **SVM –** метод опорных вектором
- **Logistic Regression –** логистическая регрессия
- **AdaBoost –** метод адаптивного бустинга

02 МЕТОДЫ
15 МИН

Деревья

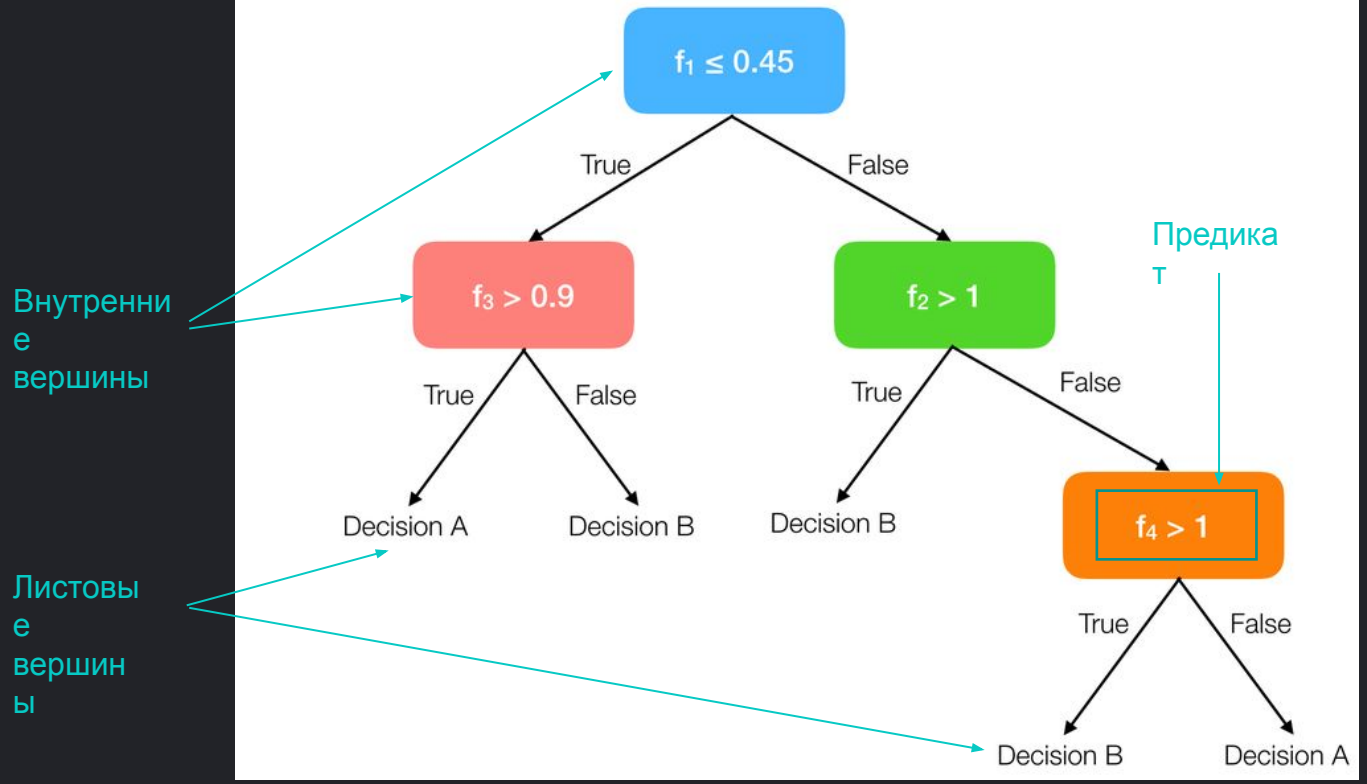
Дано:
X – матрица объектов-признаков, $x_i \in R^n$
Y – вектор ответов, $y_i \in \{+1, -1\}$

Постановка задачи бинарной классиф.:
$Y = f(X)$

Найти:
f – ?

Бинарное решающее дерево

- Ациклический граф
- Если вершина соединена с 2 дочерними – внутренняя вершина
- Если нет – листовая (терминальная) вершина
- На внутренних вершинах сидят предикаты



02 МЕТОДЫ
15 МИН

Деревья

Критерий разбиения (ветвления):

- **GINI**

  Показывает, сколько есть пар объектов одного и того же класса, которые вместе идут в левую либо в правую дочернюю вершину.

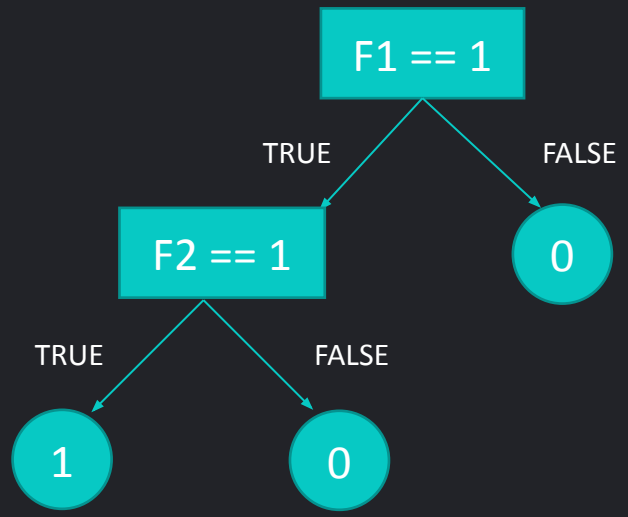
- **ENTROPY**

  Критерий разбиения из теории информации. Суть примерно такая же как у gini.

$$Gini(F1) = 1 - p_1^2 - \mathrm{p}_0^2$$
$$F_{split} = \min_F\{F1, F2, F3\}$$

$$Entropy(F1) = -p_1 \log_2 p_1 - p_0 \log_2 \mathrm{p}_0$$
$$IG(F1) = Entropy(F1) - \text{average}\,[Entropy(F2), Entropy(F2)]$$

F1	F2	F3	Y
1	1	0	1
1	0	0	0
0	1	0	0
1	1	1	1

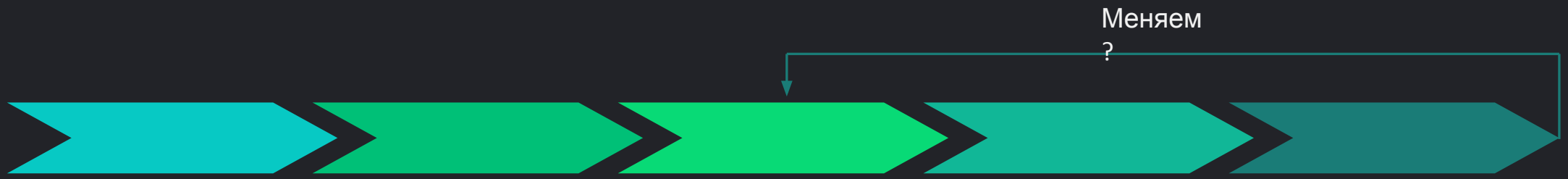


И это все?

- **TreeGrowth**
- **Overfitting**
- **Prunning**
- **NonBinary**
- **MaxDepth**
- **MinSamplesLeaf**
- **RandomForest**

03 ПЛАН РЕШЕНИЯ ML-ЗАДАЧ 5 МИН

Первое приближение



step 1

Загрузка данных

Предобработка данных: заполняем пропуски, удаляем дубликаты, кодируем категориальные переменные.

step 2

Разделение

Делим на обучающую и тестовую выборки.

step 3

Выбор методов

Выбираем алгоритм, метрику качества.

step 4

Обучаем

Обучаем модель.

step 5

Контроль

Анализируем результаты.

04 ПРАКТИКА
10 МИН

Практика 2

Первая ML-задача.

Откройте учебный Notebook из архива с материалами к занятию.

Метрики классификации

Статус	N
0	990
1	10

$$Accuracy = \frac{990}{990 + 10} = 0.99$$

$$BaseRate = \frac{\text{len}(class_{\max})}{\text{len}(total)} = \frac{990}{1000} = 0.99$$

Полнота (recall)

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

На сколько хорошо алгоритм определяет класс №1
Контролируем ошибку 2 рода

Матрица ошибок (Confusion matrix):

	Y = 1	Y = 0
A = 1	TP истинно положительные	FP ложно положительные
A = 0	FN ложно негативные	TN Истинно негативные

	Y = 1	Y = 0
A = 1	0	10
A = 0	0	990

$$Recall = \frac{0}{0 + 0} = 0$$

Когда использовать?
Когда нужно минимизировать **ложные пропуски**.

	Y = 1	Y = 0
A = 1	480	490
A = 0	10	20

$$Acc = \frac{480 + 20}{1000} = 0.5$$

$$Recall = \frac{480}{480 + 10} = 0.98$$

Точность (precision)

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

Доля объектов, названных классом №1

Контролируем ошибку 1 рода

Матрица ошибок (Confusion matrix):

	Y = 1	Y = 0
A = 1	TP истинно положительные	FP ложно положительные
A = 0	FN ложно негативные	TN Истинно негативные

	Y = 1	Y = 0
A = 1	0	10
A = 0	0	990

$$Precision = \frac{0}{0 + 10} = 0$$

Когда использовать?

Когда нужно минимизировать **ложные попадания**.

	Y = 1	Y = 0
A = 1	480	20
A = 0	490	10

$$Acc = \frac{480 + 10}{1000} = 0.49$$

$$Precision = \frac{480}{480 + 10} = 0.98$$

F2 мера

$$F2 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Среднее гармоническое между точностью и полностью.
Контролируем ошибки 1 и 2 рода.

Матрица ошибок (Confusion matrix):

	Y = 1	Y = 0
A = 1	TP истинно положительные	FP ложно положительные
A = 0	FN ложно негативные	TN Истинно негативные

	Y = 1	Y = 0
A = 1	100	0
A = 0	900	0

$$F2\_score = \frac{2 \cdot 1 \cdot 0.1}{1 + 0.1} = 0.18$$

Когда использовать?
Когда нужно минимизировать обе ошибки.

	Y = 1	Y = 0
A = 1	700	150
A = 0	150	0

$$Acc = \frac{700}{1000} = 0.7$$

$$F2\_score = \frac{2 \cdot 0.82^2}{2 \cdot 0.82} = 0.82$$