Многоклассовая классификация.

Кантонистова Е.О.

ВШЭ, 2023

ПЛАН ЛЕКЦИИ

1. Переобучение и регуляризация
2. Задачи многоклассовой классификации

ПЕРЕОБУЧЕНИЕ И РЕГУЛЯРИЗАЦИЯ

МЕТОД БОРЬБЫ С ПЕРЕОБУЧЕНИЕМ: РЕГУЛЯРИЗАЦИЯ

Большие значения параметров (весов) модели w – признак переобучения.

P.S. Если в данных есть линейно-зависимые признаки, они тоже приводят к переобучению (и к большим весам).

МЕТОД БОРЬБЫ С ПЕРЕОБУЧЕНИЕМ: РЕГУЛЯРИЗАЦИЯ

Большие значения параметров (весов) модели w – признак переобучения.

Решение проблемы – ***регуляризация***.

Будем минимизировать регуляризованный функционал ошибки:

$$\boldsymbol{Q_{alpha}(w) = Q(w) + \alpha \cdot R(w) \rightarrow \min_{w}} \quad ,$$

где $R(w)$ - регуляризатор.

РЕГУЛЯРИЗАЦИЯ

- **Регуляризация штрафует за слишком большие веса.**

Наиболее используемые регуляризаторы:

- L_2-регуляризатор: $R(w) = \left|\left|w\right|\right|_2 = \sum_{i=1}^{d} w_i^2$
- L_1-регуляризатор: $R(w) = \left|\left|w\right|\right|_1 = \sum_{i=1}^{d} |w_i|$

РЕГУЛЯРИЗАЦИЯ

- **Регуляризация штрафует за слишком большие веса.**

Наиболее используемые регуляризаторы:

- L_2-регуляризатор: $R(w) = \left||w|\right|_2 = \sum_{i=1}^{d} w_i^2$
- L_1-регуляризатор: $R(w) = \left||w|\right|_1 = \sum_{i=1}^{d} |w_i|$

Пример регуляризованного функционала:

$$Q(a(w), X) = \frac{1}{l}\sum_{i=1}^{l}((w, x_i) - y_i)^2 + \alpha \sum_{i=1}^{d} w_i^2 ,$$

где α – коэффициент регуляризации.

ПОЛЕЗНОЕ СВОЙСТВО L1-РЕГУЛЯРИЗАЦИИ

Все ли признаки в задаче нужны?

- Некоторые признаки могут не иметь отношения к задаче, т.е. они не нужны.
- Если есть ограничения на скорость получения предсказаний, то чем меньше признаков, тем быстрее
- Если признаков больше, чем объектов, то решение задачи будет неоднозначным.

Поэтому в таких случаях надо делать отбор признаков, то есть убирать некоторые признаки.

L_1-РЕГУЛЯРИЗАЦИЯ

Утверждение. В результате обучения модели с L_1-регуляризатором происходит зануление некоторых весов, т.е. отбор признаков.

Можно показать, что задачи

$$(1) \quad Q(w) + \alpha \left|\left|w\right|\right|_1 \to \min_w$$

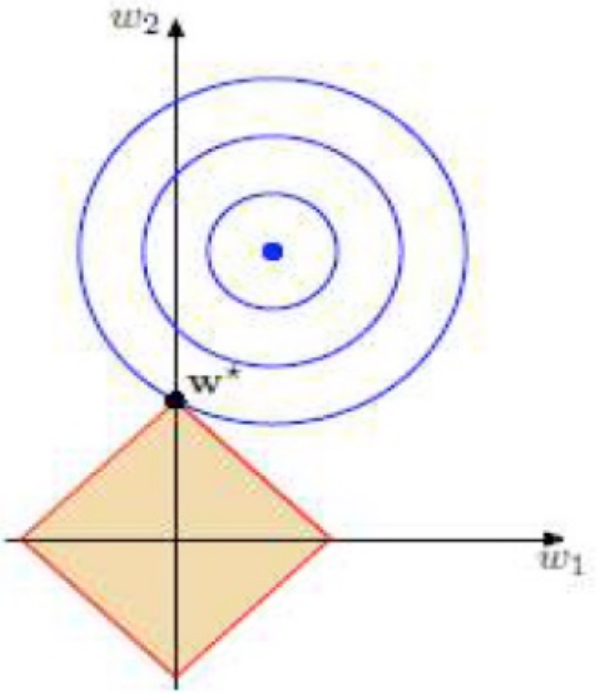
и

$$(2) \quad \begin{cases} Q(w) \to \min\limits_w \\ \left|\left|w\right|\right|_1 \leq C \end{cases}$$

эквивалентны.

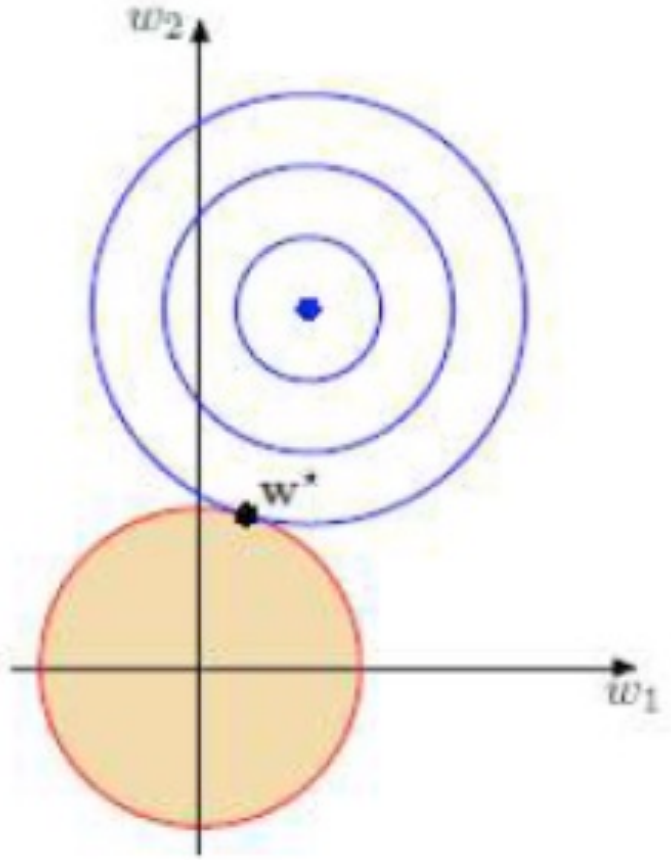
ОТБОР ПРИЗНАКОВ ПО L1-РЕГУЛЯРИЗАЦИИ

Нарисуем линии уровня $Q(w)$ и область $\left|\left|w\right|\right|_1 \leq C$:



Если признак незначимый, то соответствующий вес близок к 0. Отсюда получим, что в большинстве случаев решение нашей задачи попадает в вершину ромба, т.е. обнуляет незначимый признак.

L2-РЕГУЛЯРИЗАЦИЯ НЕ ОБНУЛЯЕТ ПРИЗНАКИ



РАЗРЕЖЕННЫЕ МОДЕЛИ

Модели, в которых часть весов равна 0, называются ***разреженными моделями***.

- L1-регуляризация зануляет часть весов, то есть делает модель разреженной.

МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ

ПОДХОД ONE-VS-ALL

Решаем задачу классификации на K классов.

- Обучим K бинарных классификаторов $b_1(x), \dots, b_K(x)$, каждый из которых решает задачу: ***принадлежит объект x к классу k_i или не принадлежит?***

Например, линейные классификаторы будут иметь вид

$$b_k(x) = sign((w_k, x))$$

ПОДХОД ONE-VS-ALL

Решаем задачу классификации на K классов.

- Обучим K бинарных классификаторов $b_1(x), \dots, b_K(x)$, каждый из которых решает задачу: *принадлежит объект x к классу k_i или не принадлежит?*

Например, линейные классификаторы будут иметь вид

$$b_k(x) = sign((w_k, x))$$

- Тогда в качестве итогового предсказания будем выдавать **класс самого уверенного классификатора:**

$$a(x) = \underset{k \in \{1,\dots,K\}}{arg\max}((w_k, x))$$

ПОДХОД ONE-VS-ALL

Решаем задачу классификации на K классов.

- Обучим K бинарных классификаторов $b_1(x), \dots, b_k(x)$, каждый из которых решает задачу: *принадлежит объект x к классу k_i или не принадлежит?*

Например, линейные классификаторы будут иметь вид

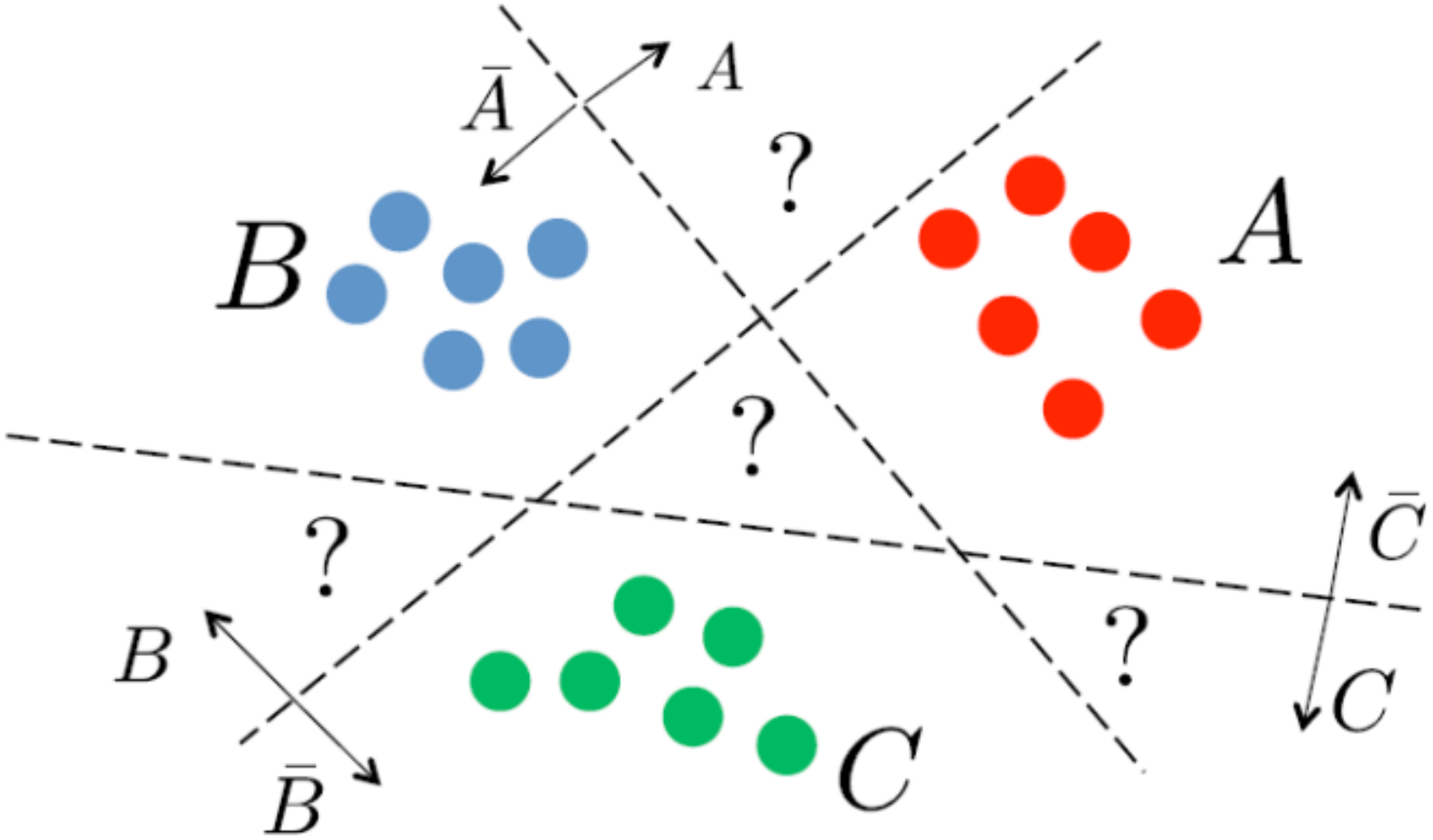
$$b_k(x) = sign((w_k, x))$$

- Тогда в качестве итогового предсказания будем выдавать класс самого уверенного классификатора:

$$a(x) = \underset{k \in \{1,\dots,K\}}{arg\max}((w_k, x))$$

- Предсказания классификаторов могут иметь разные масштабы, поэтому сравнивать их некорректно.

ПОДХОД ONE-VS-ALL



ПОДХОД ALL-VS-ALL

- **Для каждой пары классов i и j обучим бинарный классификатор $a_{ij}(x)$, который будет предсказывать класс i или j**

(если всего K классов, то получим C_K^2 классификаторов).

Каждый такой классификатор будем обучать только на объектах классов i и j.

ПОДХОД ALL-VS-ALL

- Для каждой пары классов i и j обучим бинарный классификатор $a_{ij}(x)$, который будет предсказывать класс i или j

(если всего K классов, то получим C_K^2 классификаторов).

Каждый такой классификатор будем обучать только на объектах классов i и j.

- **В качестве итогового предсказания выдадим класс, который предсказало наибольшее число алгоритмов:**

$$a(x) = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \sum_{i=1}^{K} \sum_{j \neq i} [a_{ij}(x) = k]$$

ПОДХОД ALL-VS-ALL

- Для каждой пары классов i и j обучим бинарный классификатор $a_{ij}(x)$, который будет предсказывать класс i или j

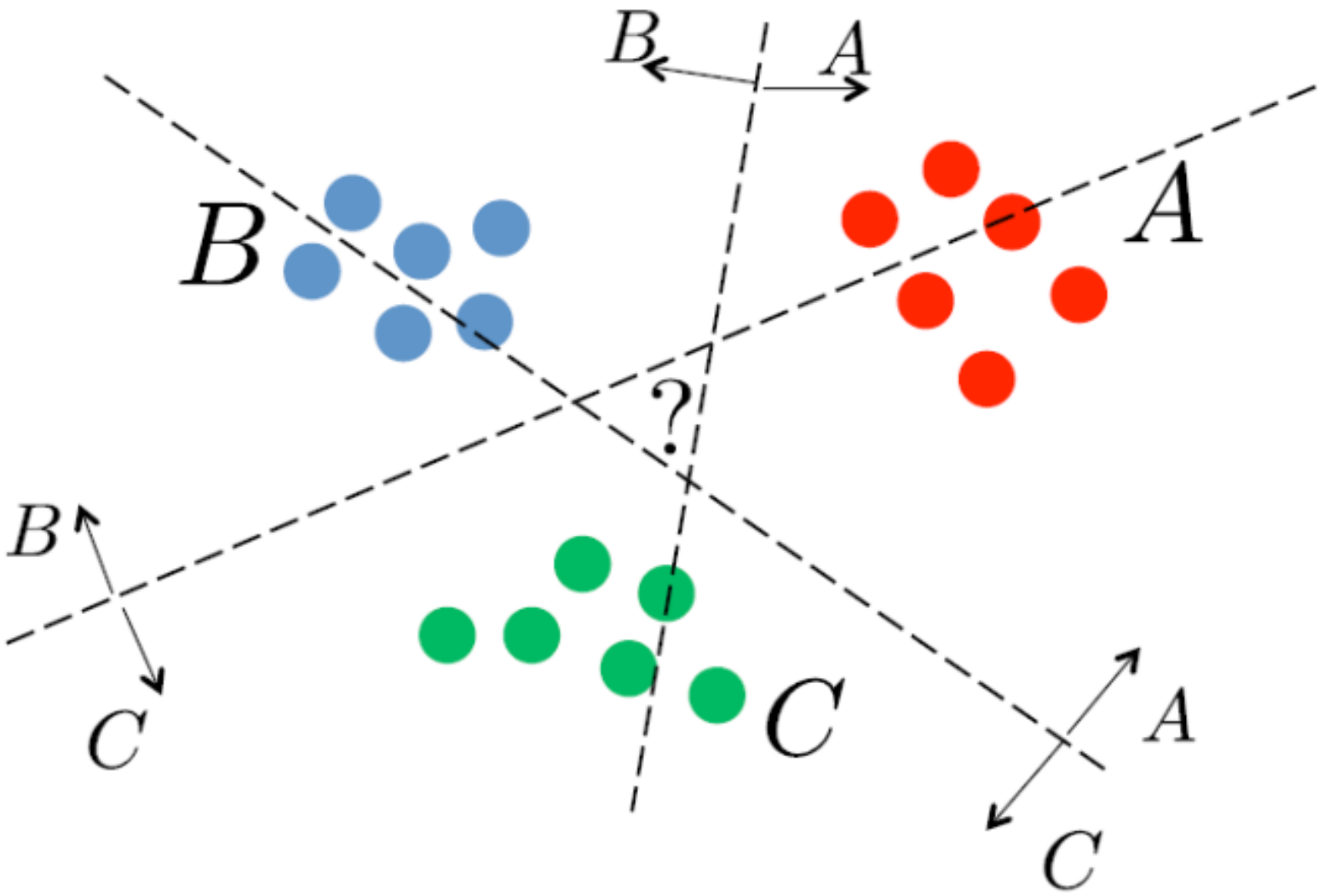
(если всего K классов, то получим **C_K^2 классификаторов**).

Каждый такой классификатор будем обучать только на объектах классов i и j.

- В качестве итогового предсказания выдадим класс, который предсказало наибольшее число алгоритмов:

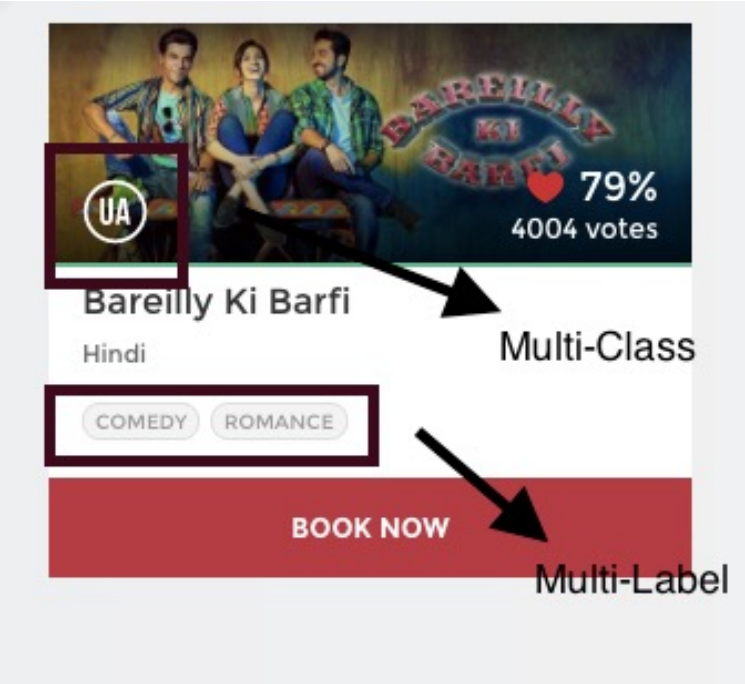
$$a(x) = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \sum_{i=1}^{K} \sum_{j \neq i} [a_{ij}(x) = k]$$

ПОДХОД ALL-VS-ALL



MULTICLASS AND MULTI-LABEL CLASSIFICATION

- Если каждый объект может принадлежать только одному классу, то решаем задачу multiclass классификации
- Если каждый объект может принадлежать нескольким классам (задача классификации с пересекающимися классами), то решаем задачу multi-label классификации.



МЕТРИКИ КАЧЕСТВА

Для бинарной классификации мы использовали такие метрики как accuracy, precision, recall, f1-score, confusion matrix, roc-auc, pr-auc. Многие из них обобщаются на многоклассовый случай. Давайте поговорим как.

- Метрика *accuracy* - это доля правильных ответов модели, она без изменений в формуле может применяться для любого количества классов.

МЕТРИКИ КАЧЕСТВА

		True/Actual	
		Positive ()	Negative
Predicted	Positive ()	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

Чему равны precision и recall?

МЕТРИКИ КАЧЕСТВА

		True/Actual		
		Cat ()	Fish ()	Hen ()
Predicted	Cat ()	4	6	3
	Fish ()	1	2	0
	Hen ()	1	2	6

Для вычисления точности и полноты в этом случае существует несколько подходов:

- Микроусреднение (micro-average)
- Макроусреднение (macro-average)
- Взвешенное усреднение (weighted-average)

МАКРОУСРЕДНЕНИЕ

В этом подходе мы вычисляем значение выбранной метрики для каждой бинарной ситуации (кошка/не кошка, рыба/не рыба, курица/не курица), а затем усредняем полученные числа.

- Например, посчитаем точность и полноту для ситуации кошка/не кошка:

		True/Actual		
		Cat	Fish	Hen
Predicted	Cat	4	6	3
	Fish	1	2	0
	Hen	1	2	6

МАКРОУСРЕДНЕНИЕ

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

- $precision(cat) = \frac{TP}{TP+FP} = \frac{4}{4+6+3} = \frac{4}{13}$

То есть false positive - это все объекты, которые модель ошибочно назвала кошкой (их $6 + 3$)

- $recall(cat) = \frac{TP}{TP+FN} = \frac{4}{4+1+1}$

Здесь false negative - это все кошки, которых модель не нашла (кошки, названные моделью не кошками).

Тогда macro-average

$$precision = \frac{precision(cat) + precision(fish) + precision(hen)}{3}$$

ВЗВЕШЕННОЕ УСРЕДНЕНИЕ

		True/Actual		
		Cat ()	Fish ()	Hen ()
Predicted	Cat ()	4	6	3
	Fish ()	1	2	0
	Hen ()	1	2	6

Взвешенное усреднение (weighted-average)

В этом подходе мы усредняем посчитанные для каждого класса метрики с весами, пропорциональными количеству объектов класса.

То есть weighted average

$$precision = \frac{6}{25} \cdot precision(cat) + \frac{10}{25} \cdot precision(fish) + \frac{9}{25} \cdot precision(hen),$$

так как всего 25 объектов, и из них 6 кошек, 10 рыб и 9 куриц.

МИКРОУСРЕДНЕНИЕ

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

Микроусреднение (micro-average)

В этом подходе мы вычисляем значения TP, TN, FP, FN по всей матрице ошибок сразу, исходя из их определения. Затем по полученным числам вычисляем выбранные метрики.

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

TP - это количество верно угаданных объектов положительного класса. В нашем случае $TP = 4 + 2 + 6 = 12$

FP - это суммарное количество false positive-предсказаний. Например, если cat предсказана как fish, то это false positive для fish. Таким образом, FP - это сумма всех неверных предсказаний, то есть $FP = 6 + 3 + 1 + 0 + 1 + 2 = 13$

Получаем micro-average

$$precision = \frac{12}{12 + 13} = \frac{12}{25}$$

МИКРОУСРЕДНЕНИЕ

		True/Actual		
		Cat ()	Fish ()	Hen ()
Predicted	Cat ()	4	6	3
	Fish ()	1	2	0
	Hen ()	1	2	6

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

FN - это сумма false negative-предсказаний. Например, если cat предсказана как fish, то это false negative для cat. Таким образом, FN - это опять же сумма всех неверных предсказаний, то есть $FN = 6 + 3 + 1 + 0 + 1 + 2 = 13$

Получается, что *в случае микроусреднения*

$$precision = recall$$

И так как f1-score - это среднее гармоническое точности и полноты, то *при микроусреднении*

$$precision = recall = f1$$