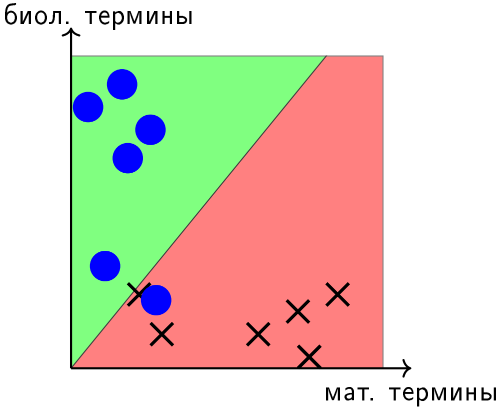
Введение

Алексей Сорокин

МГУ им. М. В. Ломоносова
весенний семестр 2022–2023 учебного года
Межфакультетский курс “Введение в компьютерную лингвистику” 22 февраля, занятие 2
Линейные классификаторы.

Машинное обучение

Упрощённая тематическая классификация на 2 класса:



Машинное обучение

- Базовая модель – линейный классификатор:

$$
\begin{aligned}
f(x) &= \operatorname{sgn}\left(\langle w, x\rangle - w_0\right), && \text{(случай двух классов)}\\
f(x) &= \arg\max_i(\langle w_i, x\rangle - w_{i0}) && \text{(случай нескольких классов)}
\end{aligned}
$$

- Входным данным сопоставляется вектор признаков $[x_1, \ldots, x_n]$.

Машинное обучение

- Базовая модель – линейный классификатор:

$$
\begin{aligned}
f(x) &= \operatorname{sgn}\left(\langle w, x\rangle - w_0\right), && \text{(случай двух классов)}\\
f(x) &= \arg\max_i(\langle w_i, x\rangle - w_{i0}) && \text{(случай нескольких классов)}
\end{aligned}
$$

- Входным данным сопоставляется вектор признаков $[x_1, \ldots, x_n]$.
- Простейший способ векторизации предложений:
  - Есть словарь из слов $v_1, \ldots, v_n$.
  - $x_i(T)$ – число вхождений слова v_i в текст T.

Линейный классификатор (случай 2 классов)

- Базовая модель – линейный классификатор (для двух классов):

$$f(x) = \operatorname{sgn}(\langle w, x\rangle - w_0),$$

- Каждый признак x_i голосуем за свой класс с весом $|w_i|$,
- $w_i > 0$ – голосует за положительный класс,
- $w_i < 0$ – голосует за отрицательный класс.

Линейный классификатор (случай 2 классов)

- Решающая функция:

$$\begin{aligned} h(x) = \langle w, x\rangle - w_0 \quad &- \quad \text{решающая функция,} \\ f(x) = \operatorname{sgn} h(x) \quad &- \quad \text{предсказанная метка класса,} \\ |h(x)| \quad &- \quad \text{расстояние от разделяющей поверхности} \end{aligned}$$

- Сравнение с эталоном:

$$\begin{aligned} y(x) \in -1, 1 \quad &- \quad \text{метка класса,} \\ (y(x)f(x) > 0) \quad &- \quad \text{условие правильности классификации,} \\ \max(-y(x)f(x), 0) \quad &- \quad \text{ошибка классификации} \end{aligned}$$

- Введём $M_w(x, y(x)) = y(x)f(x) = y(x)(\langle w, x\rangle - w_0)$ – отступ объекта x.

Машинное обучение (случай 2 классов)

- Обучение модели – подбор коэффициентов $w_0, w_1, \dots, w_n$ по обучающей выборке.
- Задача – минимизировать суммарную ошибку на ней.
- То есть надо ввести функцию штрафа $Q_w(x, y)$ и решать задачу

$$\sum_i Q_w(x_i, y_i) \to \min$$

- Например, можно минимизировать сумму модулей отрицательных отступов

$$Q_w(x_i, y_i) = \max(-M_w(x_i, y_i), 0) = \max(-y(x_i)(\langle w, x_i \rangle - w_0), 0)$$

Алгоритм персептрона Розенблатта

- Пусть $w^{(0)}$ — начальный вектор весов, $\eta > 0$ — темп обучения.

Алгоритм персептрона Розенблатта

- Пусть $w^{(0)}$ — начальный вектор весов, $\eta > 0$ — темп обучения.
- Если для всех x_i $M(x_i) > 0$, то уже построена правильная классификация.
- Иначе найдём i, такое что $M(x_i) \leqslant 0$.

Алгоритм персептрона Розенблатта

- Пусть $w^{(0)}$ — начальный вектор весов, $\eta > 0$ — темп обучения.
- Если для всех x_i $M(x_i) > 0$, то уже построена правильная классификация.
- Иначе найдём i, такое что $M(x_i) \leqslant 0$.
- Обновим вектор весов $w^{(1)} = w^{(0)} + \eta x_i y_i$.

Алгоритм персептрона Розенблатта

- Пусть $w^{(0)}$ — начальный вектор весов, $\eta > 0$ — темп обучения.
- Если для всех x_i $M(x_i) > 0$, то уже построена правильная классификация.
- Иначе найдём i, такое что $M(x_i) \leqslant 0$.
- Обновим вектор весов $w^{(1)} = w^{(0)} + \eta x_i y_i$.

Как изменился отступ:

$$\begin{aligned} M'(x_i) &= (w^{(1)}, x_i)y_i = (w^{(0)} + \eta x_i y_i, x_i)y_i = \\ &= (w^{(0)}, x_i)y_i + \eta(x_i, x_i)y_i^2 > M(x_i) \end{aligned}$$

Алгоритм персептрона Розенблатта

- Пусть $w^{(0)}$ — начальный вектор весов, $\eta > 0$ — темп обучения.
- Если для всех x_i $M(x_i) > 0$, то уже построена правильная классификация.
- Иначе найдём i, такое что $M(x_i) \leqslant 0$.
- Обновим вектор весов $w^{(1)} = w^{(0)} + \eta x_i y_i$.

Как изменился отступ:

$$
\begin{aligned}
M'(x_i) &= (w^{(1)}, x_i)y_i = (w^{(0)} + \eta x_i y_i, x_i)y_i = \\
&= (w^{(0)}, x_i)y_i + \eta(x_i, x_i)y_i^2 > M(x_i)
\end{aligned}
$$

Отступ для данного объекта улучшился, а для остальных?

Сходимость персептрона

Линейная разделимость

Выборка $\langle X_T, Y_T \rangle$ называется линейно разделимой с порогом δ, если найдётся такая линейная функция $f(x) = \sum\limits_{j=1}^{n} w_j x_j - w_0$, что для всех i $f(x_i) y_i > \delta$.

Сходимость персептрона

Линейная разделимость

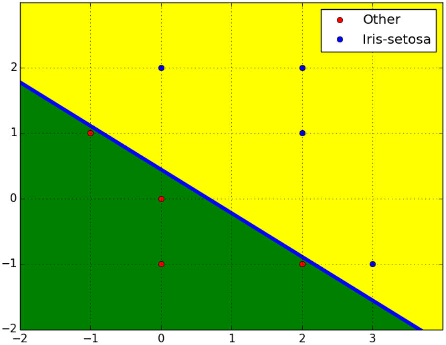
Выборка $\langle X_T, Y_T \rangle$ называется линейно разделимой с порогом δ, если найдётся такая линейная функция $f(x) = \sum_{j=1}^{n} w_j x_j - w_0$, что для всех i $f(x_i) y_i > \delta$.

Теорема

Для всякой линейно разделимой выборки персептрон верно находит разделяющую гиперплоскость за конечное число шагов.

Тестовый пример

Класс	Объекты
-1	$[0,0], [0,-1], [2,-1], [-1,1]$
1	$[2,1], [2,2], [0,2], [3,-1]$

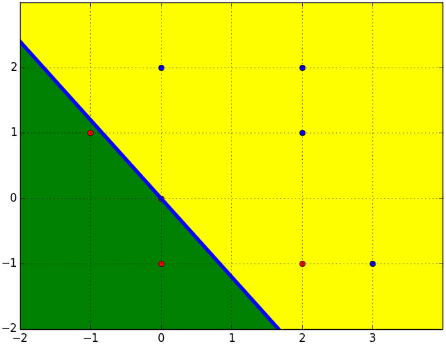


Пошаговая классификация

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250

Пошаговая классификация

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250



Пошаговая классификация

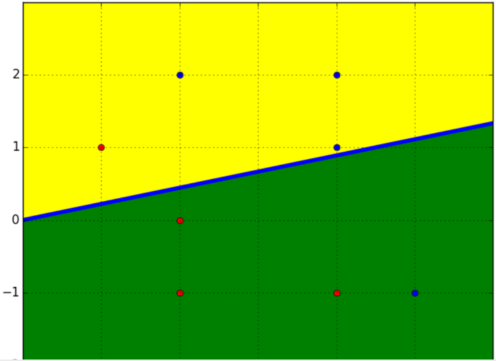
	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250

$$w^{(1)} = w^{(0)} - [-1, 2, -1] = [1, -0.5, 2.25]$$

Пошаговая классификация

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250

$$w^{(1)} = w^{(0)} - [-1, 2, -1] = [1, -0.5, 2.25]$$



Пошаговая классификация

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250

$$w^{(1)} = w^{(0)} - [-1, 2, -1] = [1, -0.5, 2.25]$$

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w_1 = [1, -0.5, 2.25]$	1.000	3.250	4.250	-1.750	0.250	2.500	3.500	-4.750

Пошаговая классификация

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250

$$w^{(1)} = w^{(0)} - [-1, 2, -1] = [1, -0.5, 2.25]$$

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w_1 = [1, -0.5, 2.25]$	1.000	3.250	4.250	-1.750	0.250	2.500	3.500	-4.750

$$w^{(2)} = w_1 + [-1, 3, -1] = [0, 2.5, 1.25] \text{ и т. д.}$$

Пошаговая классификация

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250

$$w^{(1)} = w^{(0)} - [-1, 2, -1] = [1, -0.5, 2.25]$$

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w_1 = [1, -0.5, 2.25]$	1.000	3.250	4.250	-1.750	0.250	2.500	3.500	-4.750

$$w^{(2)} = w_1 + [-1, 3, -1] = [0, 2.5, 1.25] \text{ и т. д.}$$

В конце концов $w^{(5)} = [1, 1.5, 2.25]$

Пошаговая классификация

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w^{(0)} = [0, 1.5, 1.25]$	-0.000	1.250	-1.750	0.250	4.250	5.500	2.500	3.250

$$w^{(1)} = w^{(0)} - [-1, 2, -1] = [1, -0.5, 2.25]$$

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w_1 = [1, -0.5, 2.25]$	1.000	3.250	4.250	-1.750	0.250	2.500	3.500	-4.750

$$w^{(2)} = w_1 + [-1, 3, -1] = [0, 2.5, 1.25] \text{ и т. д.}$$

В конце концов $w^{(5)} = [1, 1.5, 2.25]$

	[0,0]	[0, -1]	[2,-1]	[-1, 1]	[2, 1]	[2, 2]	[0, 2]	[3, -1]
	-1	-1	-1	-1	1	1	1	1
$w_5 = [1, 1.5, 2.25]$	1.000	3.250	0.250	0.250	4.250	6.500	3.500	1.250

Линейный классификатор для нескольких классов

- Линейный классификатор для двух классов:

$$f(x) = \operatorname{sgn}(\langle w, x \rangle - w_0)$$

- В случае нескольких классов решающее правило выглядит как

$$f(x) = \arg\max_i \langle w_i, x \rangle - w_{i0}$$

- Т. е. есть несколько решающих функций, и нужно выбрать тот класс, значение для которого больше.

Персептрон для нескольких классов

- Двуклассовый персептрон стремится увеличить значение функции $g(x) = \langle w, x\rangle - w_0$ на объектах положительного класса и уменьшить на объектах отрицательного.
- Для нескольких классов нужно увеличить $g_i(x)$ для $i = y(x)$ и уменьшить для остальных классов.
- Элементарный шаг алгоритма на входе $(x, i = y(x))$:
  - Найти $\hat{\imath} = \arg\max_j(\langle w_j, x\rangle - w_{j0})$.
  - Если $i = \hat{\imath}$, ничего не делать (ответ правильный).

Персептрон для нескольких классов

- Двуклассовый персептрон стремится увеличить значение функции $g(x) = \langle w, x\rangle - w_0$ на объектах положительного класса и уменьшить на объектах отрицательного.
- Для нескольких классов нужно увеличить $g_i(x)$ для $i = y(x)$ и уменьшить для остальных классов.
- Элементарный шаг алгоритма на входе $(x, i = y(x))$:
  - Найти $\hat{\imath} = \arg\max\limits_j(\langle w_j, x\rangle - w_{j0})$.
  - Если $i = \hat{\imath}$, ничего не делать (ответ правильный).
  - Иначе

$$\begin{aligned} w_i &\leftarrow w_i + \eta x, \\ w_{\hat{\imath}} &\leftarrow w_{\hat{\imath}} - \eta x \end{aligned}$$

- Это увеличит $\langle w_i, x\rangle$ и уменьшит $\langle w_{\hat{\imath}}, x\rangle$.

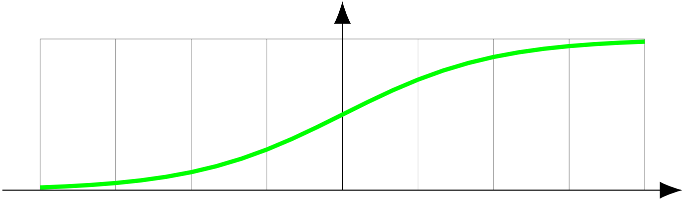
Вероятность классов в линейном классификаторе

- Пусть линейный классификатор имеет вид

$$\begin{aligned} h(x) &= \operatorname{sgn} a(x) \\ a(x) &= \langle w, x\rangle - w_0 \end{aligned}$$

- Тогда вероятность положительного класса равна:

$$p(x=1) = \sigma(a(x)) = \frac{e^{a(x)}}{e^{a(x)}+1} = \frac{1}{e^{-a(x)}+1}$$



- Сигмоидная функция переводит $(-\infty; +\infty)$ в $(0; 1)$ (вероятность).

Вероятностная постановка задачи классификации

- Пусть решается задача классификации для двух классов $Y = [0, 1]$.
- Линейный классификатор предсказывает вероятность

$$p_w(x) = \sigma(\exp(\langle w, x \rangle - w_0)).$$

- Если $y(x) = 1$, то логично стремиться сделать $p_w(x)$ близкой к 1.
- В качестве функции штрафа для случая $y(x) = 1$ можно взять

$$L(y, p) = -\log p_w(x) = -\log \frac{1}{e^{-a(x)} + 1} = \log\left(1 + e^{-a(x)}\right)$$

Вероятностная постановка задачи классификации

- В качестве функции штрафа для случая $y(x) = 1$ можно взять

$$L(y, p) = -\log p_w(x) = -\log \frac{1}{e^{-a(x)} + 1} = \log\left(1 + e^{-a(x)}\right)$$

- Для $y(x) = 0$ аналогично

$$L(y, p) = -\log(1 - p_w(x)) = -\log \frac{e^{-a(x)}}{e^{-a(x)} + 1} = \log\left(1 + e^{a(x)}\right)$$

Вероятностная постановка задачи классификации

- В качестве функции штрафа для случая $y(x) = 1$ можно взять

$$L(y, p) = -\log p_w(x) = -\log \frac{1}{e^{-a(x)} + 1} = \log\left(1 + e^{-a(x)}\right)$$

- Для $y(x) = 0$ аналогично

$$L(y, p) = -\log(1 - p_w(x)) = -\log \frac{e^{-a(x)}}{e^{-a(x)} + 1} = \log\left(1 + e^{a(x)}\right)$$

- Эти формулы можно объединить

$$L(y, p) = -y \log p - (1 - y) \log(1 - p)$$

Стохастическая минимизация штрафа

- Мы получили функцию штрафа

$$L(y, p_w(x)) = -y \log p_w(x) - (1-y)\log(1-p_w(x))$$

- Её нужно сделать как можно меньше

$$L(X, Y, w) = \sum_{x,y} L(y, p_w(x)) \to \min_w$$

- Минимизировать сразу по всей выборке – долго, лучше пытаться уменьшить штраф для отдельного объекта (как в персептроне).

Стохастический градиентный спуск

- Функция быстрее всего растёт в направлении своего вектора частных производных.
- Этот вектор называется градиентом:

$$\frac{\partial L}{\partial w} = [\frac{\partial L}{\partial w_1}, \ldots, \frac{\partial L}{\partial w_n}]$$

- Тогда уменьшать её надо в противоположном направлении:

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

Стохастический градиентный спуск для логистической регрессии

- Логистическая регрессия:

$$\begin{aligned} Q_i &= -y_i \log p_i - (1 - y_i) \log (1 - p_i) \\ p_i &= \sigma(a_i) = \frac{1}{1+\exp(-a_i)} \\ a_i &= \langle w, x_i \rangle + b \end{aligned}$$

- Вычислим шаг градиентного спуска для логистической регрессии.

$$\begin{aligned} \frac{\partial Q_i}{\partial w} &= \frac{\partial Q_i}{\partial p_i} \frac{\partial p_i}{\partial a_i} \frac{\partial a_i}{\partial w} \\ \frac{\partial Q_i}{\partial p_i} &= -\frac{y_i}{p_i} + \frac{1-y_i}{1-p_i}, \\ \frac{\partial p_i}{\partial a_i} &= p_i(1 - p_i), \\ \frac{\partial a_i}{\partial w} &= x_i. \end{aligned}$$

- Для $y_i = 1$ имеем $\frac{\partial Q_i}{\partial w} = (1 - p_i)x_i$.

Стохастический градиентный спуск для логистической регрессии

- Шаг логистической регрессии:

$$\begin{aligned} w &\leftarrow w + \eta(1 - p_i)x_i \qquad &(\text{для } y_i = 1)), \\ w &\leftarrow w - \eta p_i x_i \qquad &(\text{для } y_i = 0)) \end{aligned}$$

- Это аналогично исправлению весов в персептроне, но с учётом текущей ошибки.
- Логистическая регрессия сильнее исправляет веса для тех объектов, на которых сильнее ошибка.

Вероятность классов в линейном классификаторе

- Многоклассовый классификатор:

$$
\begin{aligned}
h(x) &= \arg\max_k a_k(x) \\
a_k(x) &= \langle w_k, x \rangle - w_{k0}, \quad k = 1, \dots, K.
\end{aligned}
$$

- Тогда вероятность k-го класса равна:

$$
p(h(x) = k) = \mathrm{softmax}(\mathbf{a})_k = \frac{e^{a_k(x)}}{\sum_j e^{a_j(x)}}
$$

Функция штрафа для нескольких классов

- Функция штрафа (для $y(x) = k$):

$$L(y, \mathbf{p}) = -\log p_k = -\log \frac{e^{a_k(x)}}{\sum_j e^{a_j(x)}} = \log \sum_j e^{a_j(x)} - a_k(x)$$

- Получаем шаг градиентного спуска для каждого из w_j:

$$\begin{aligned} w_k &\leftarrow w_k + \eta(1 - p_k)x_k && \text{(для правильного класса)}, \\ w &\leftarrow w - \eta p_j x_j && \text{(для остальных классов, } j \neq k\text{)} \end{aligned}$$

Функция штрафа для нескольких классов

- Функция штрафа (для $y(x) = k$):

$$L(y, \mathbf{p}) = -\log p_k = -\log \frac{e^{a_k(x)}}{\sum_j e^{a_j(x)}} = \log \sum_j e^{a_j(x)} - a_k(x)$$

- Получаем шаг градиентного спуска для каждого из w_j:

$$\begin{aligned} w_k &\leftarrow w_k + \eta(1 - p_k)x_k && \text{(для правильного класса),} \\ w &\leftarrow w - \eta p_j x_j && \text{(для остальных классов, } j \neq k) \end{aligned}$$

- Увеличиваем рейтинг положительного класса и уменьшаем остальных.
- Изменение тем сильнее, чем больше ошибка.

Выборки

- Обучение ведётся на обучающей выборке, качество мерится на контрольной.
- Они не пересекаются и должны по возможности отличаться.

Выборки

- Обучение ведётся на обучающей выборке, качество мерится на контрольной.
- Они не пересекаются и должны по возможности отличаться.
- Ещё часто используется валидационная выборка:
  - Как и тестовая, она не используется для настройки весов модели.
  - Она нужна, чтобы остановить обучение, когда качество на обучающей выборке ещё растёт, а на валидационной уже нет.
  - Это называется переобучение (основная проблема нейронных сетей).

Бинарные метрики качества

- Таблица ошибок (confusion matrix):

Правильно \ Предсказано	1	0
1	TP (TruePositive)	FN(FalseNegative)
0	FP (FalsePositive)	TN(TrueNegative)

Бинарные метрики качества

- Таблица ошибок (confusion matrix):

Правильно \ Предсказано	1	0
1	TP (TruePositive)	FN(FalseNegative)
0	FP (FalsePositive)	TN(TrueNegative)

- Меры качества:
  - Корректность (доля правильных ответов):

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

  - Точность:

$$\mathrm{P(recision)} = \frac{TP}{TP+FP}$$

  - Полнота:

$$\mathrm{R(ecall)} = \frac{TP}{TP+FN}$$

  - F-мера:

$$\mathrm{F1} = \frac{2PR}{P+R} = \frac{TP}{TP+0.5FN+0.5FP}$$

Многоклассовые метрики качества

- Корректность (процент правильных ответов).
- Усреднённые значения бинарных метрик:
  - По объектам: взвешенное усреднение (не учитывает несбалансированность классов).
  - По классам: макроусреднение (считает все классы равнозначными).

Многоклассовые метрики качества

No	Actual	Predicted	Match
1	Airplane	Airplane	✓
2	Car	Boat	✗
3	Car	Car	✓
4	Car	Car	✓
5	Car	Boat	✗
6	Airplane	Boat	✗
7	Boat	Boat	✓
8	Car	Airplane	✗
9	Airplane	Airplane	✓
10	Car	Car	✓

	precision	recall	f1-score	support
Aeroplane	0.67	0.67	0.67	3
Boat	0.25	1.00	0.40	1
Car	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

Per-Class F1 scores

Average F1 scores