Системы обработки медиаданных

Классификация. Логистическая регрессия

д.т.н., доцент Вашкевич М. И.

vashkevich@bsuir.by

Белорусский государственный университет информатики и радиоэлектроники

Кафедра электронных вычислительных средств

Задача классификации

Пример: классификация ирисов (Фишер, 1940-е)

150 растений (50 каждого сорта)

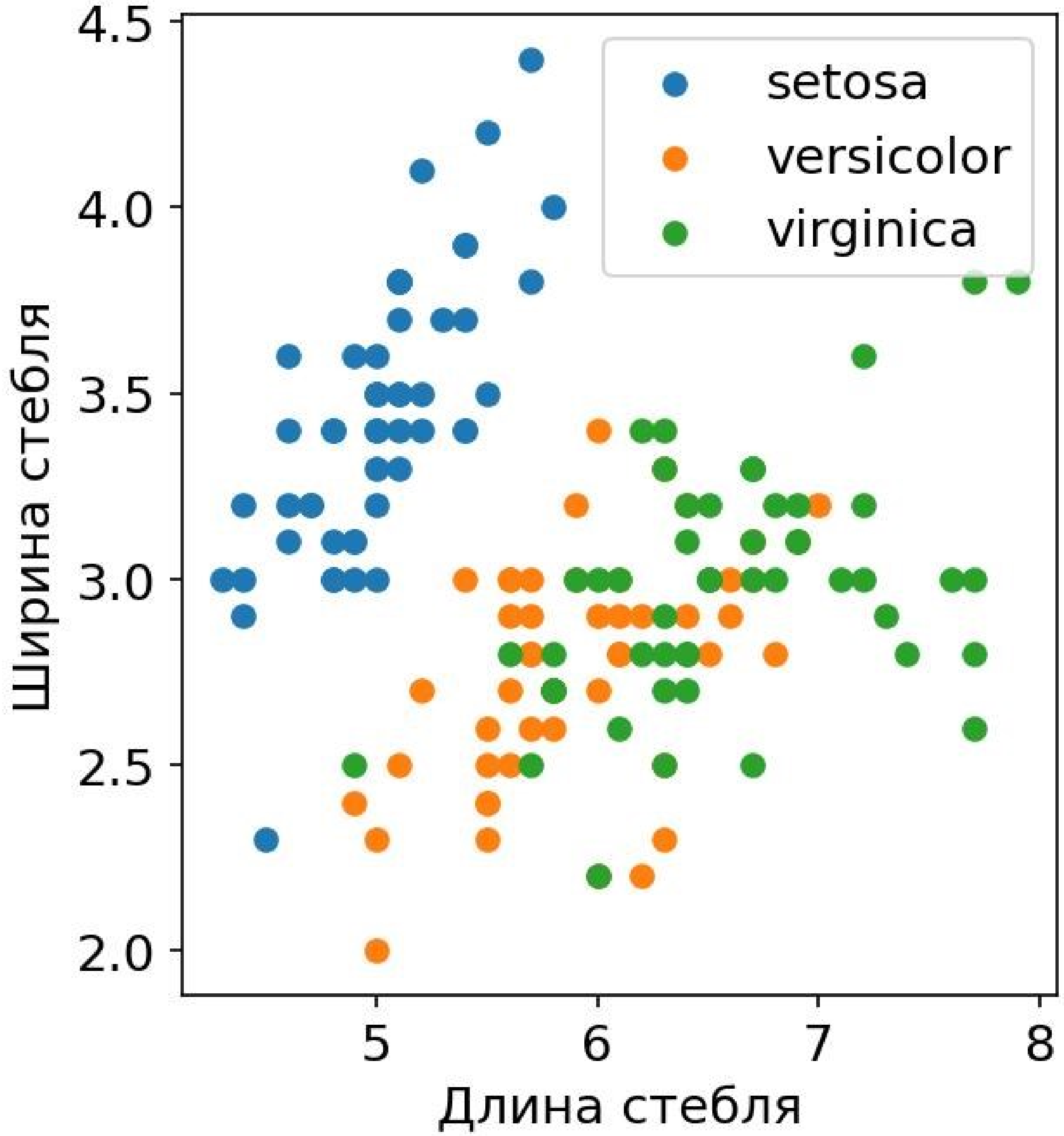
- сорта: щетинистый, разноцветный, виргинский
- длина и ширина чашелистика (в см)
- длина и ширина лепестка (в см)

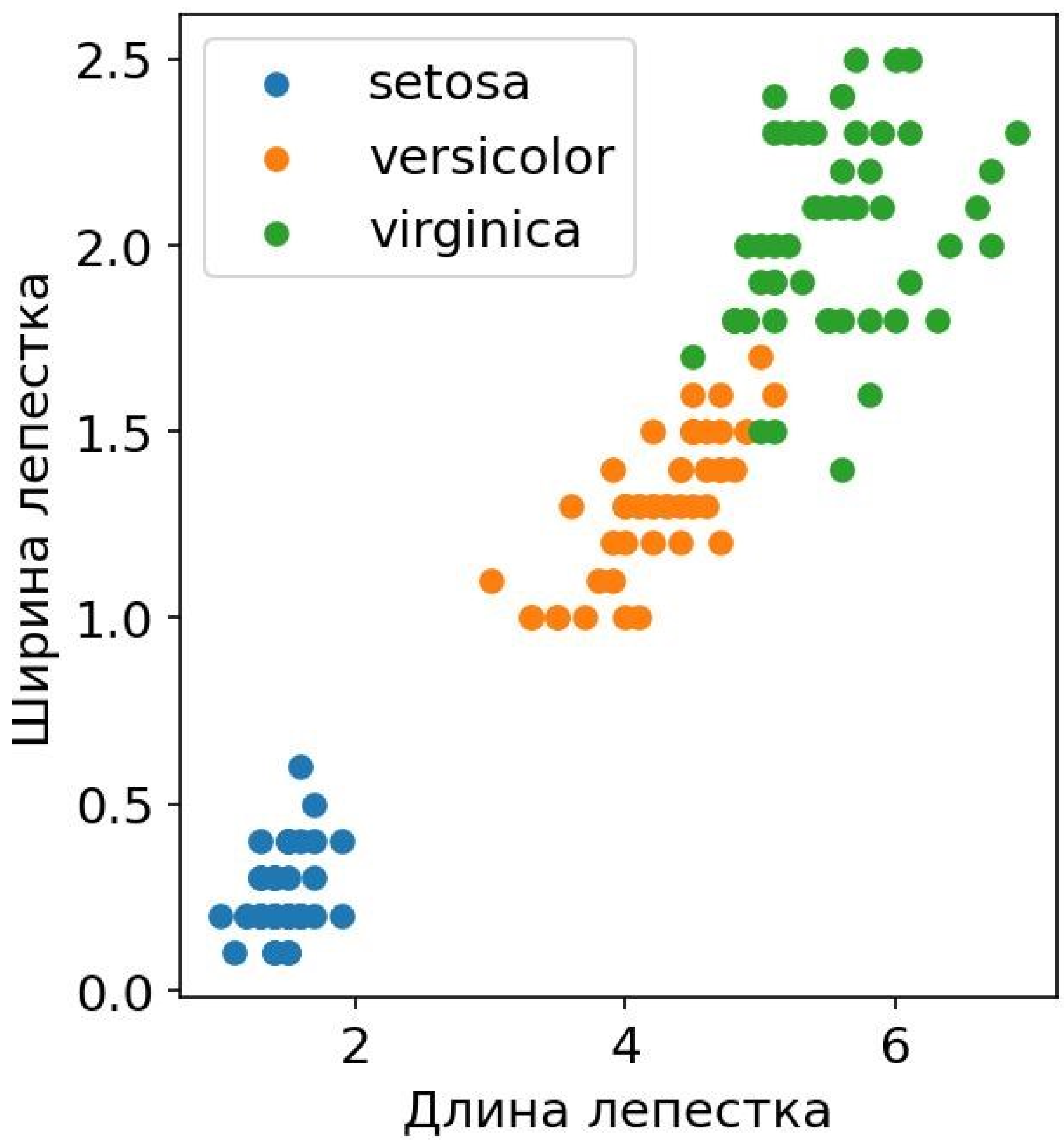
Ирис щетинистый (setosa)

Ирис разноцветный (versicolor)

Ирис виргинский (verginica)

Задача классификации





Формальная постановка задачи классификации

- $X_1, X_2, \dots, X_p$ –***предикторы*** (или ***описательные признаки***).

Все предикторы можно объединить в один вектор

$$X \coloneqq (X_1, X_2, \dots, X_p)$$

- Y – случайная величина, называемая ***целевой переменной***

Формальная постановка задачи классификации

- $X_1, X_2, \dots, X_p$ –***предикторы*** (или ***описательные признаки***).

Все предикторы можно объединить в один вектор

$$X \coloneqq (X_1, X_2, \dots, X_p)$$

- Y – случайная величина, называемая ***целевой переменной***
- Пусть $\mathcal{Y}$ – область значений Y. В задаче классификации $\mathcal{Y}$ – это конечное множество классов (пространство меток классов). Иногда говорят, что Y относится к ***категориальному*** типу данных.
- $\mathcal{D} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ – множество примеров из неизвестного совместного распределения $p(X, Y)$ входов и выходов, которое называется ***данными***.
- $\mathcal{D}$ обычно записывается списком

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

Формальная постановка задачи классификации

- Задача классификации – **предсказать** Y **на основе** X.

$$f: X \to Y$$

- $f(x)$ иногда называются ***помечающей функцией***.

Формальная постановка задачи классификации

- Задача классификации – **предсказать** Y **на основе** X.

$$f: X \to Y$$

- $f(x)$ иногда называются ***помечающей функцией***.

- Под обучением классификатора понимается построение функции $\hat{f}(x)$, которая как можно лучше аппроксимирует $f(x)$.

Формальная постановка задачи классификации

• Задача классификации – **предсказать** Y **на основе** X.

$$f: X \to Y$$

• $f(x)$ иногда называются ***помечающей функцией***.

• Под обучением классификатора понимается построение функции $\hat{f}(x)$, которая как можно лучше аппроксимирует $f(x)$.

• Другими словами, необходимо оценить функцию

$$f(x) := E\{Y|X = x\} = \int y \cdot p(y|x)dy$$

на основе имеющихся данных $\mathcal{D}$.

Бинарная классификация

В простейшем случае есть всего два класса, которые называют ***положительным*** и ***отрицательным***, ⊕ и ⊖ или $+1$ и -1.

Двухклассовую классификацию называют ***бинарной классификацией***.

Примеры

– Фильтрация спама
– Медицинская диагностика (⊕ – наличие заболевания)
– Обнаружение мошенничества с кредитными картами.

Бинарная классификация

Данные	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	…	$\mathbf{x}_N$
Метки	+	+	-	+	…	-
Предсказание модели						

Бинарная классификация

Данные	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	…	$\mathbf{x}_N$
Метки y_i	+	+	-	+	…	-
Предсказание модели $\hat{y}_i$	+	+	-	-	…	-

Результаты бинарной классификации

Метка класса	Предсказание модели	Тип результата	Название результата
$\oplus$	$\oplus$	**TP** (**T**rue **P**ositive)	Истинно-положительный

Результаты бинарной классификации

Метка класса	Предсказание модели	Тип результата	Название результата
+	+	**TP** (**T**rue **P**ositive)	Истинно-положительный
+	−	**FN** (**F**alse **N**egative)	Ложно-отрицательный

Результаты бинарной классификации

Метка класса	Предсказание модели	Тип результата	Название результата
$\oplus$	$\oplus$	**TP** (**T**rue **P**ositive)	Истинно-положительный
$\oplus$	$\ominus$	**FN** (**F**alse **N**egative)	Ложно-отрицательный
$\ominus$	$\ominus$	**TN** (**T**rue **N**egative)	Истинно-отрицательный

Результаты бинарной классификации

Метка класса	Предсказание модели	Тип результата	Название результата
+	+	**TP** (**T**rue **P**ositive)	Истинно-положительный
+	-	**FN** (**F**alse **N**egative)	Ложно-отрицательный
-	-	**TN** (**T**rue **N**egative)	Истинно-отрицательный
-	+	**FP** (**F**alse **P**ositive)	Ложно-положительный

Результаты бинарной классификации: пример

Данные	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
Метки y_i	+	+	-	+	-	-
Предсказание модели $\hat{y}_i$	+	+	-	-	+	-
Тип результата	?	?	?	?	?	?

Результаты бинарной классификации: пример

Данные	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
Метки y_i	+	+	-	+	-	-
Предсказание модели $\hat{y}_i$	+	+	-	-	+	-
Тип результата	TP	TP	?	?	?	?

Результаты бинарной классификации: пример

Данные	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
Метки y_i	+	+	-	+	-	-
Предсказание модели $\hat{y}_i$	+	+	-	-	+	-
Тип результата	TP	TP	TN	?	?	TN

Результаты бинарной классификации: пример

Данные	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
Метки y_i	+	+	-	+	-	-
Предсказание модели $\hat{y}_i$	+	+	-	-	+	-
Тип результата	TP	TP	TN	FN	?	TN

Результаты бинарной классификации: пример

Данные	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
Метки y_i	+	+	-	+	-	-
Предсказание модели $\hat{y}_i$	+	+	-	-	+	-
Тип результата	TP	TP	TN	FN	FP	TN

Матрица неточностей

Confusion matrix – матрица неточностей (или матрица спутывания).

		Предсказанные метки		
Истинные метки класса		+	−	
	+	**TP**	**FN**	$TP + FN$
	−	**FP**	**TN**	$FP + TN$
		$TP + FP$	$FN + TN$	$TP + TN + FP + FN$

В последнем столбце и в последней строке находятся ***маргиналы*** – суммы элементов в соответствующем столбце и строке.

Матрица неточностей: пример

Метки y_i	+	+	−	+	−	−
Предсказание модели $\hat{y}_i$	+	+	−	−	+	−
Тип	**TP**	**TP**	**TN**	**FN**	**FP**	**TN**

Матрица неточностей

		Предсказанные метки		
Истинные метки класса		+	−	
	+	**2**	**1**	*3*
	−	**1**	**2**	*3*
		3	*3*	*6*

Правильность

Правильность (*accuracy*) – доля правильно классифицированных тестовых примеров.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

Правильность

Правильность (*accuracy*) – доля правильно классифицированных тестовых примеров.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

		Предсказанные метки		
Истинные метки класса		+	−	
	+	TP = 2	FN = 1	3
	−	FP = 1	TN = 2	3
		3	3	6

$$Acc = ?$$

Правильность

Правильность (*accuracy*) – доля правильно классифицированных тестовых примеров.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

		Предсказанные метки		
Истинные метки класса		+	−	
	+	**TP = 2**	**FN = 1**	*3*
	−	**FP = 1**	**TN = 2**	*3*
		3	*3*	*6*

$$Acc = \frac{2 + 2}{6} = \frac{2}{3} \approx 66\%$$

Логистическая регрессия

Логистический сигмоид

Свойства логистической функции:

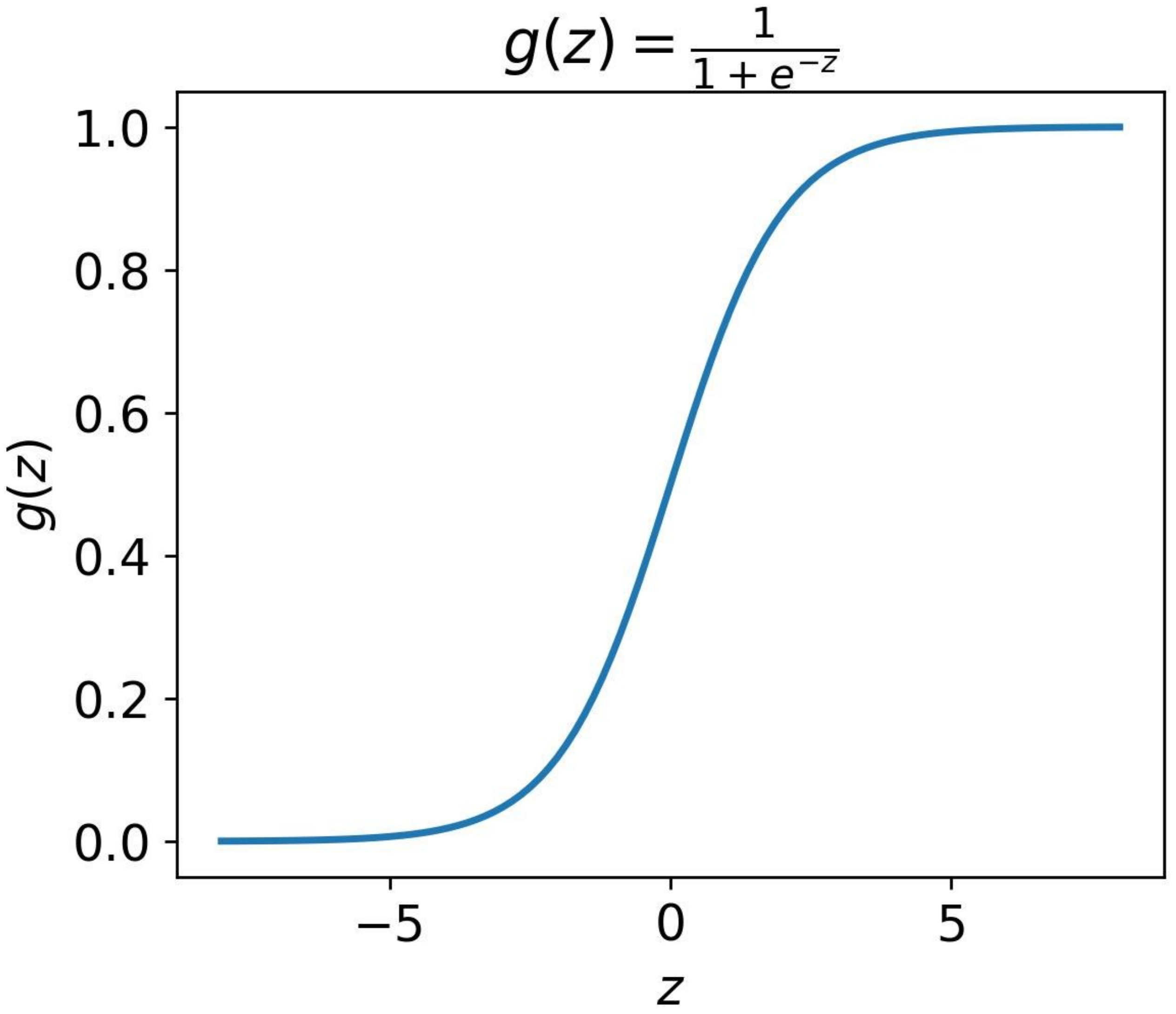
$$g(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

- значения в диапазоне от 0 до 1;
- стремится к 1 при $z \to \infty$;
- стремится к 0 при $z \to -\infty$;
- гладкая и симметричная относительно точки $(0, 0.5)$
- Симметрия:

$$g(-z) = 1 - g(z).$$

- Производная:

$$\frac{dg}{dx} = g(x)(1 - g(x)).$$



Логит-функция

Свойства логит-функции:

$$\operatorname{logit}(x) = \log\left(\frac{x}{1-x}\right)$$

• определена в диапазоне от 0 до 1;

• стремится к ∞ при $x \to 1$;

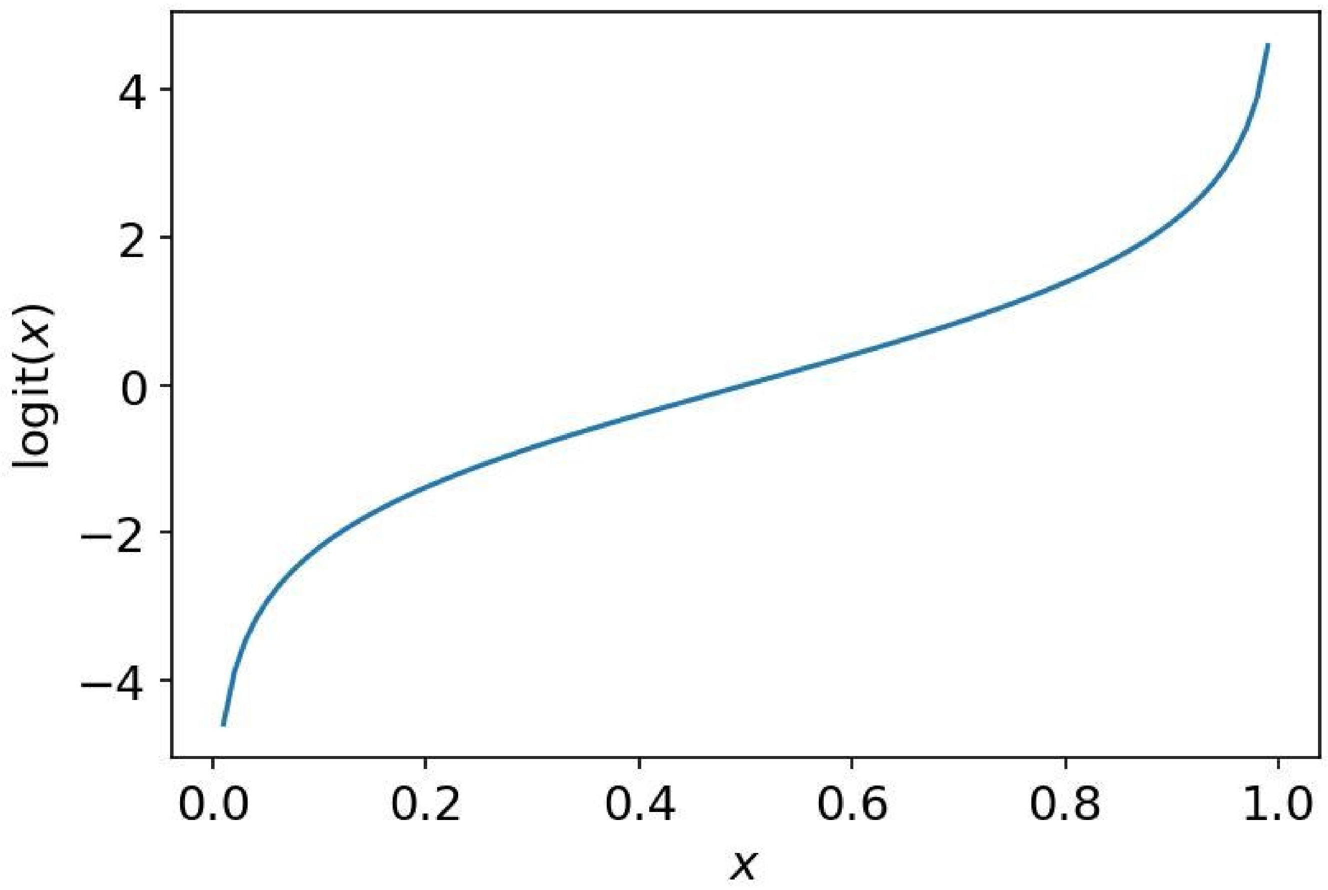
• стремится к $-\infty$ при $x \to 0$;

• гладкая и симметричная относительно точки $(0, 0.5)$

• обратная к логистической функции

$$\operatorname{logit}(g(x)) = x$$

0



Модель логистической регрессии

Целевая переменная y является бинарной, т.е. $\mathcal{Y} \coloneqq \{0,1\}$.

Можно ли использовать модель линейной регрессии

$$y = \mathbf{w}^T\mathbf{x}$$

для классификации?

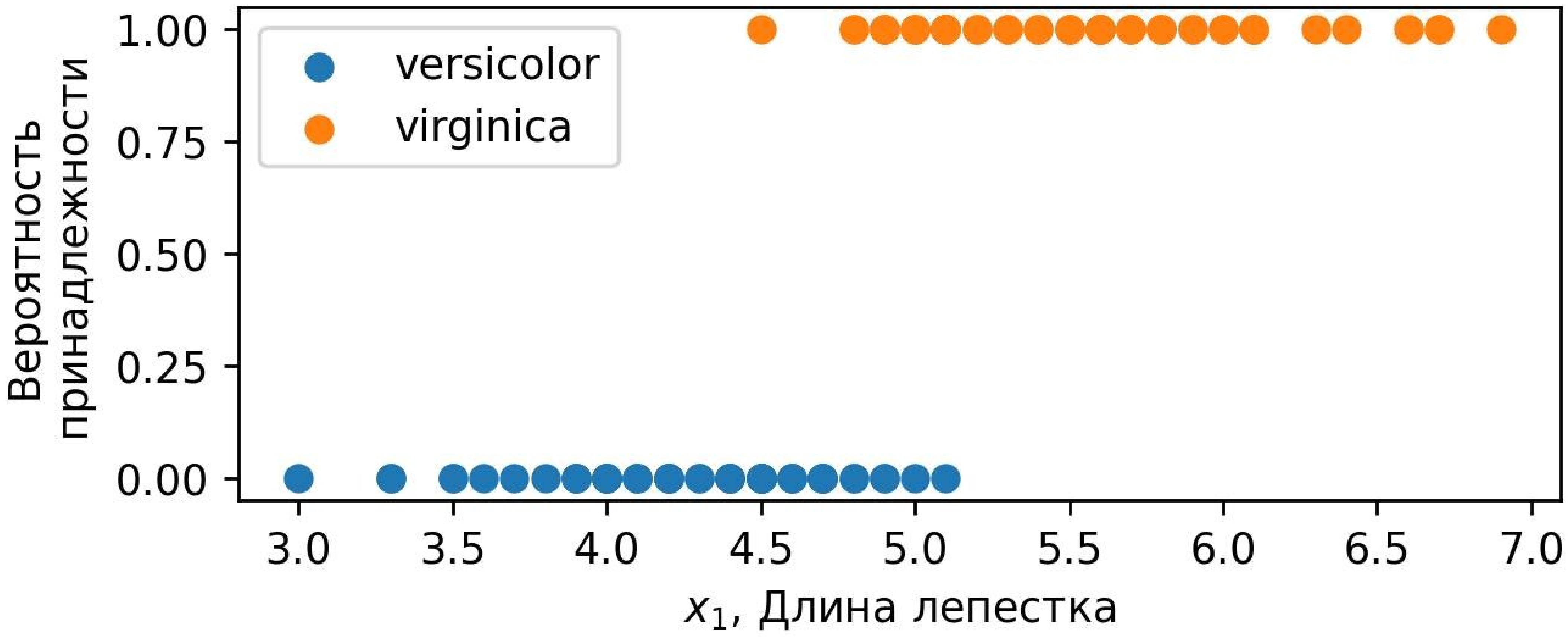
Модель логистической регрессии

Целевая переменная y является бинарной, т.е. $\mathcal{Y} \coloneqq \{0,1\}$.

Можно ли использовать модель линейной регрессии

$$y = \mathbf{w}^T \mathbf{x}$$

для классификации?



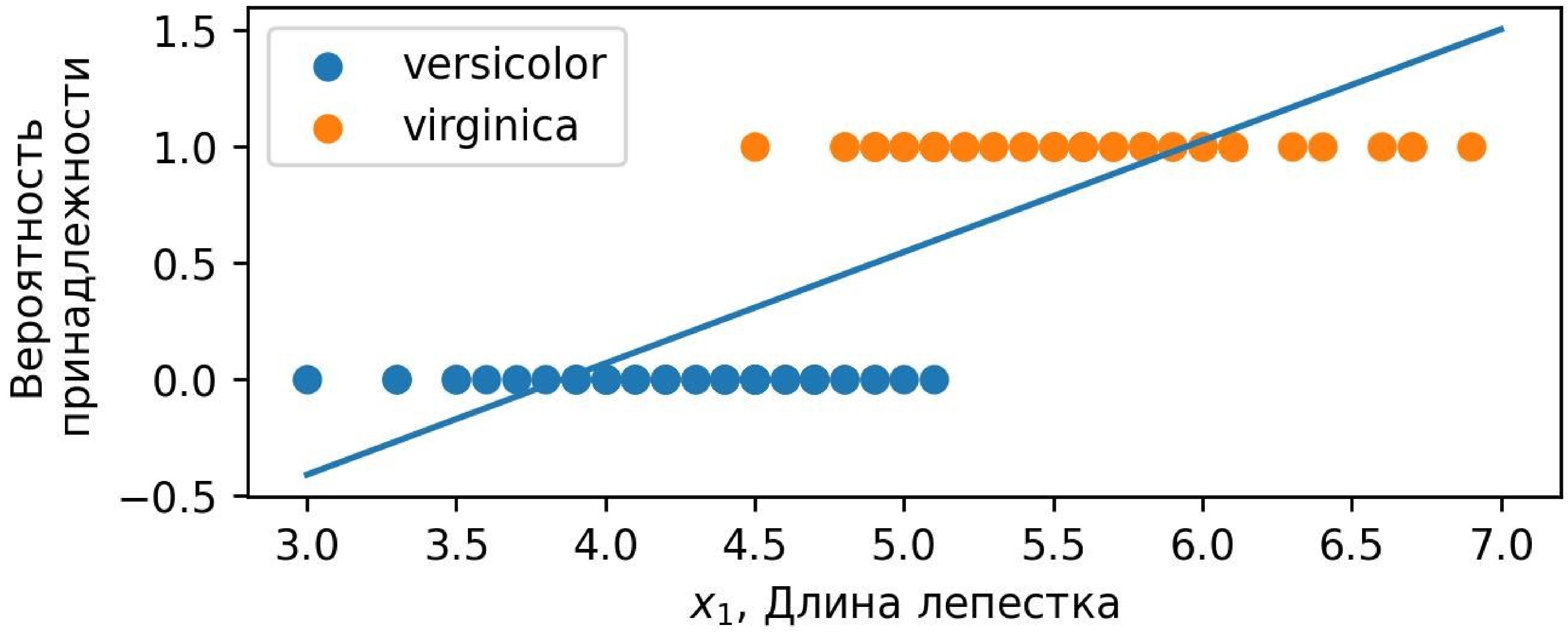
Модель логистической регрессии

Целевая переменная y является бинарной, т.е. $\mathcal{Y} \coloneqq \{0,1\}$.

Можно ли использовать модель линейной регрессии

$$y = \mathbf{w}^T\mathbf{x}$$

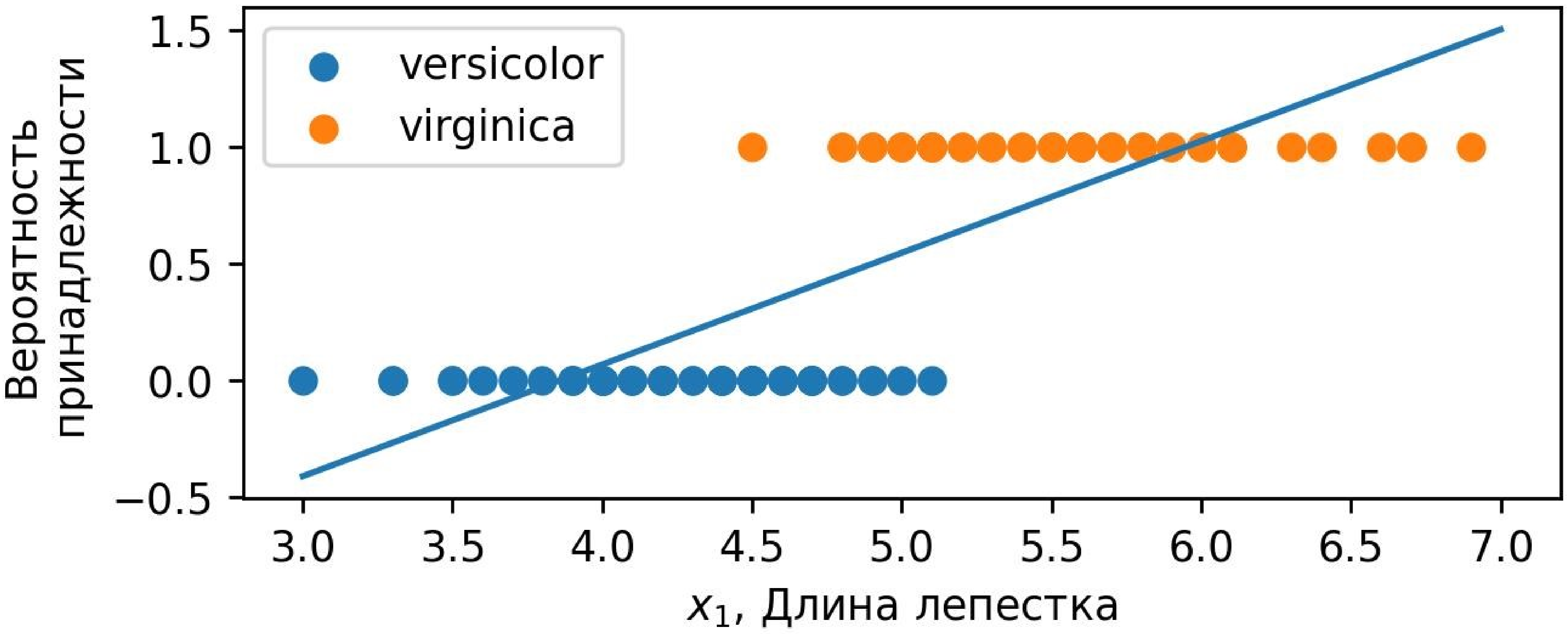
для классификации?



Модель логистической регрессии

• Линейная регрессия не подходит для предсказания вероятностей, поскольку предсказываемые ею значения принципиально не ограничены.

Применение линейной регрессии для бинарной классификации



Модель логистической регрессии

- Вместо предсказания y напрямую, можно предсказывать

$P(y = 1|\mathbf{x})$ – *вероятность, что* $y = 1$ *при условии, что наблюдается* $\mathbf{x}$.

Модель логистической регрессии

- Вместо предсказания y напрямую, можно предсказывать

$P(y = 1|\mathbf{x})$ – *вероятность, что* $y = 1$ *при условии, что наблюдается* $\mathbf{x}$.

- Для моделирования $P(y = 1|\mathbf{x})$ необходима функция, которая всегда возвращает значения из интервала от 0 до 1. В логистической регрессии для этого применяют логистическую функцию (😊):

$$P(y = 1|\mathbf{x}) = p(\mathbf{x}) = g(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}.$$

Интерпретация логистической регрессии

- Если преобразовать

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})},$$

то можно получить выражение:

$$\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \exp(-\mathbf{w}^T\mathbf{x}).$$

- Величина $p(\mathbf{x})/(1 - p(\mathbf{x}))$ называется ***риск события*** и принимает значения от 0 до ∞.

Интерпретация логистической регрессии

- Если преобразовать

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})},$$

то можно получить выражение:

$$\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \exp(-\mathbf{w}^T\mathbf{x}).$$

- Величина $p(\mathbf{x})/(1 - p(\mathbf{x}))$ называется ***риск события*** и принимает значения от 0 до ∞.

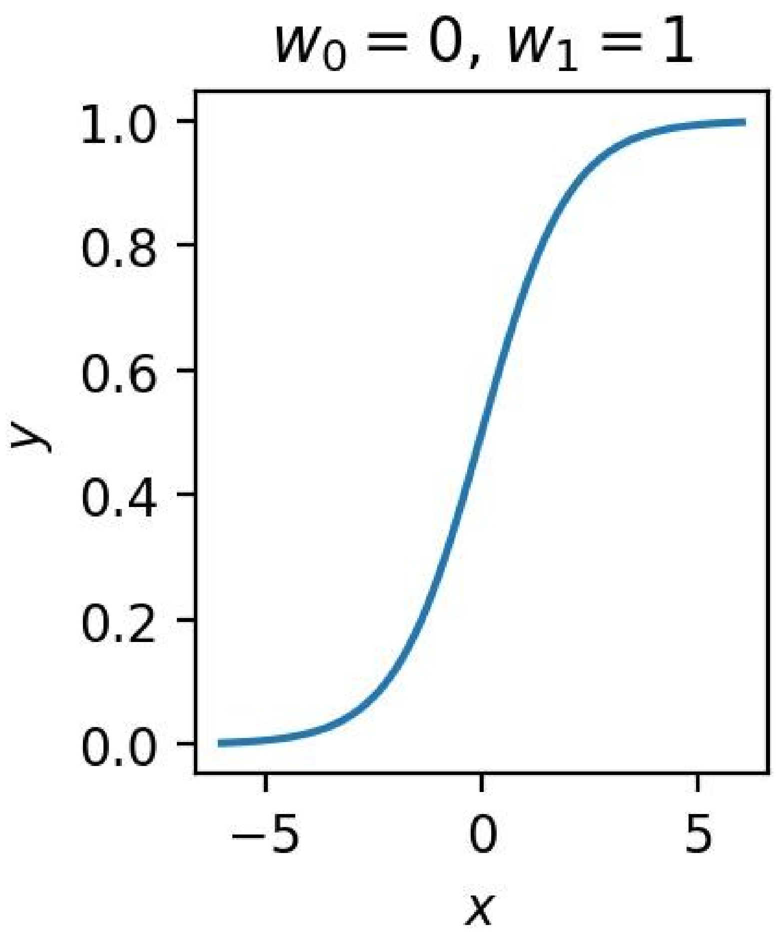
- Например, если вероятность события равна $p(\mathbf{x}) = 0{,}2$, то риск наступления этого события ¼, поскольку $\frac{0{,}2}{1-0{,}2} = 1/4$.

- Какой риск, если $p(\mathbf{x}) = 0{,}9$?

Модель логистической регрессии

- Рассмотрим, как изменение $\mathbf{w}$ влияют на вид функции
- Пример одномерных данных:

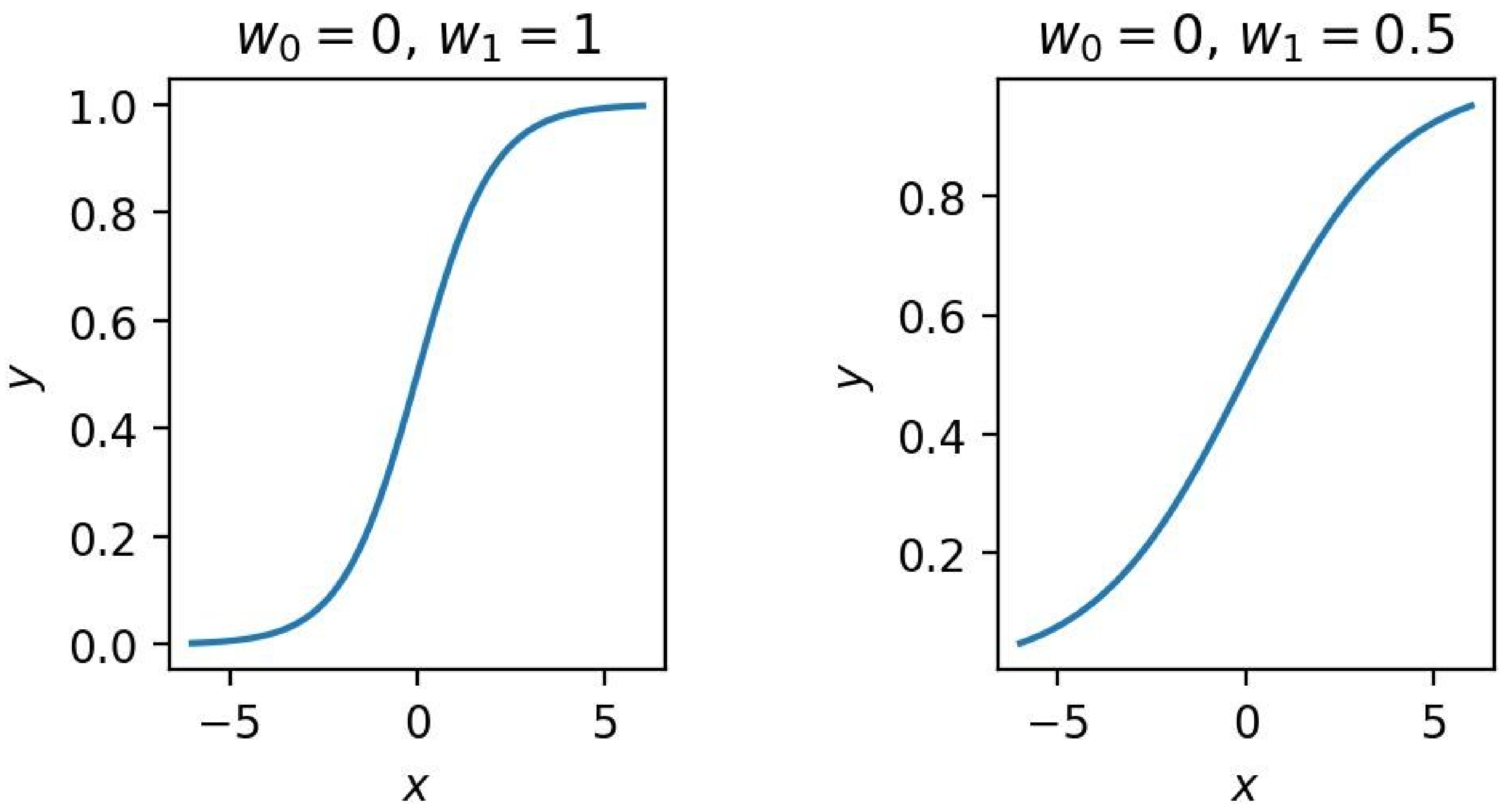
$$y = g(w_1 x + w_0) = \frac{1}{1 + \exp(-(w_1 x + w_0))}.$$



Модель логистической регрессии

- Рассмотрим, как изменение $\mathbf{w}$ влияют на вид функции
- Пример одномерных данных:

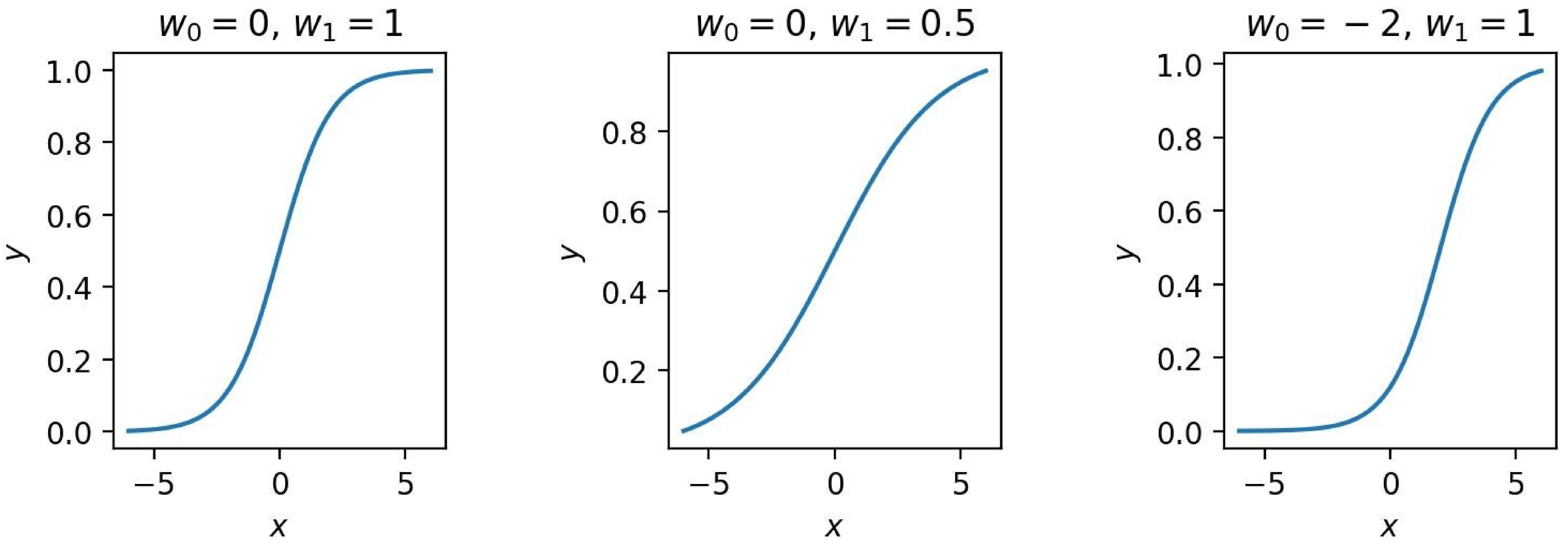
$$y = g(w_1 x + w_0) = \frac{1}{1 + \exp(-(w_1 x + w_0))}.$$



Модель логистической регрессии

- Рассмотрим, как изменение $\mathbf{w}$ влияют на вид функции
- Пример одномерных данных:

$$y = g(w_1 x + w_0) = \frac{1}{1 + \exp(-(w_1 x + w_0))}.$$



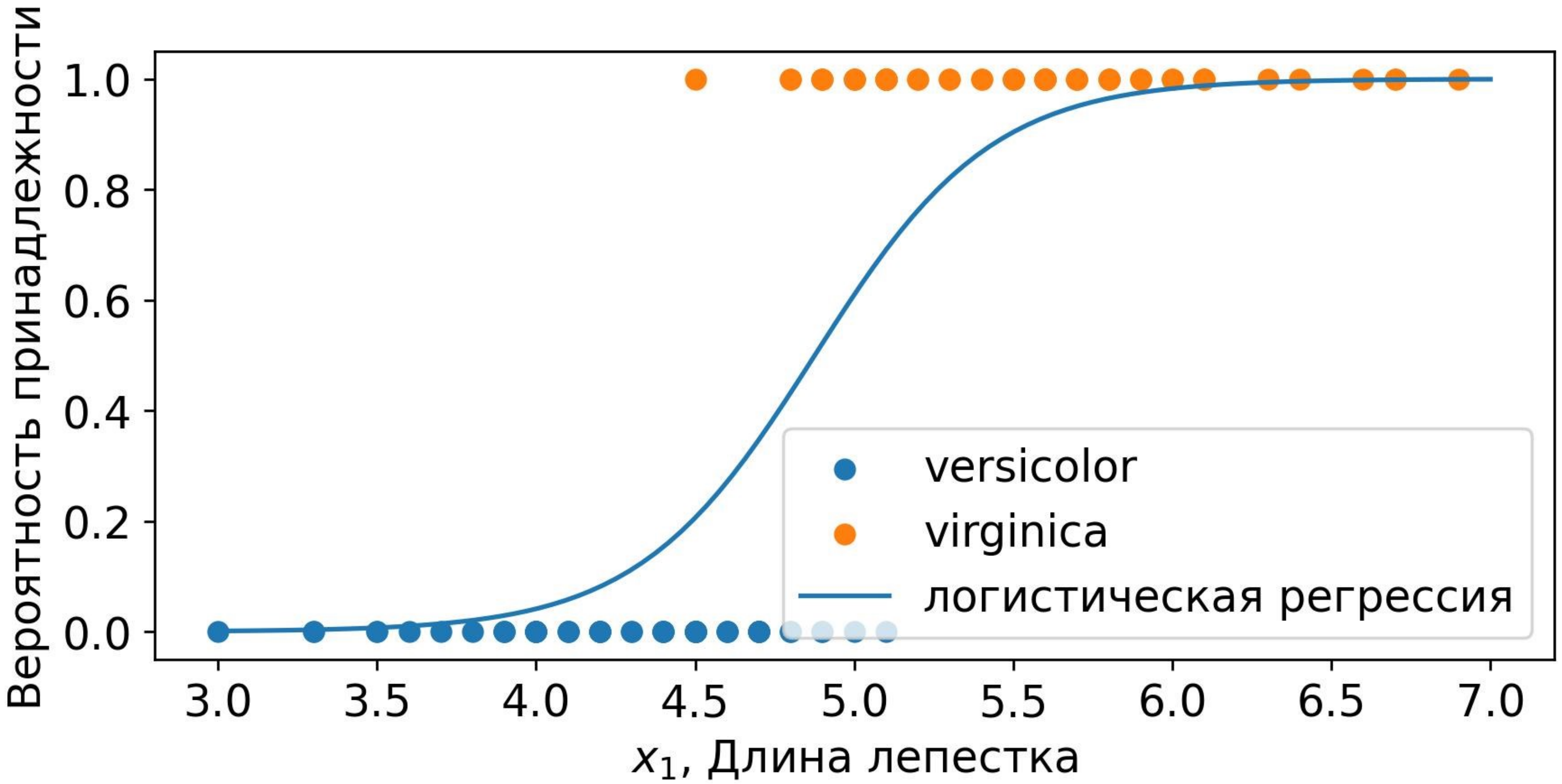
Модель логистической регрессии

- Логистическая регрессия:

$$P(y = 1|\mathbf{x}) = g(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})}.$$

- Логистическая функция всегда порождает S-образную кривую.

Применение логистической регрессии для бинарной классификации



Вероятностная интерпретация

- Поскольку мы интерпретируем $g(\mathbf{w}^T\mathbf{x})$ как вероятность $P(y = 1|\mathbf{x})$, то, используя свойства вероятностей можно записать, что

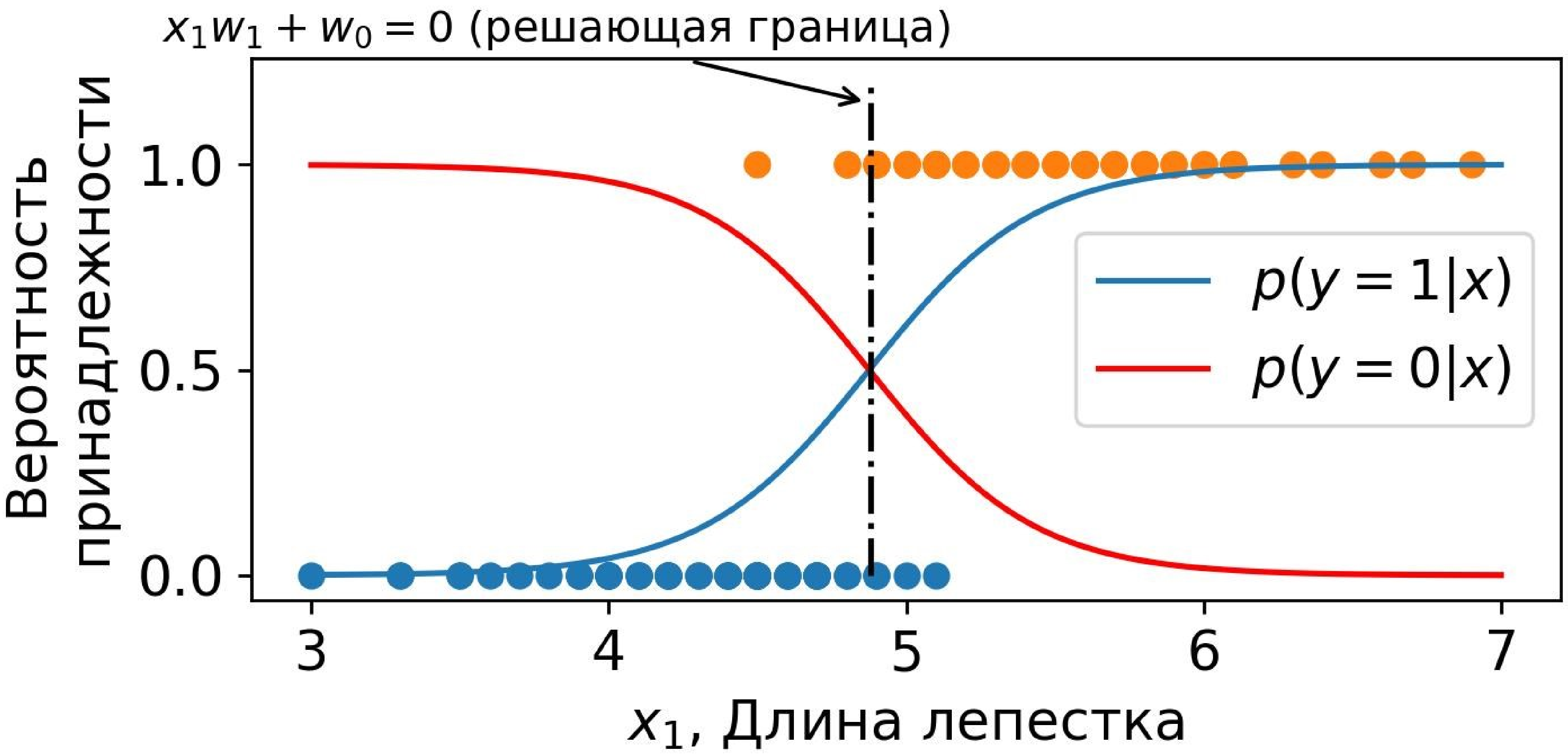
$$P(y = 1|\mathbf{x}) + P(y = 0|\mathbf{x}) = 1.$$

- Таким образом,

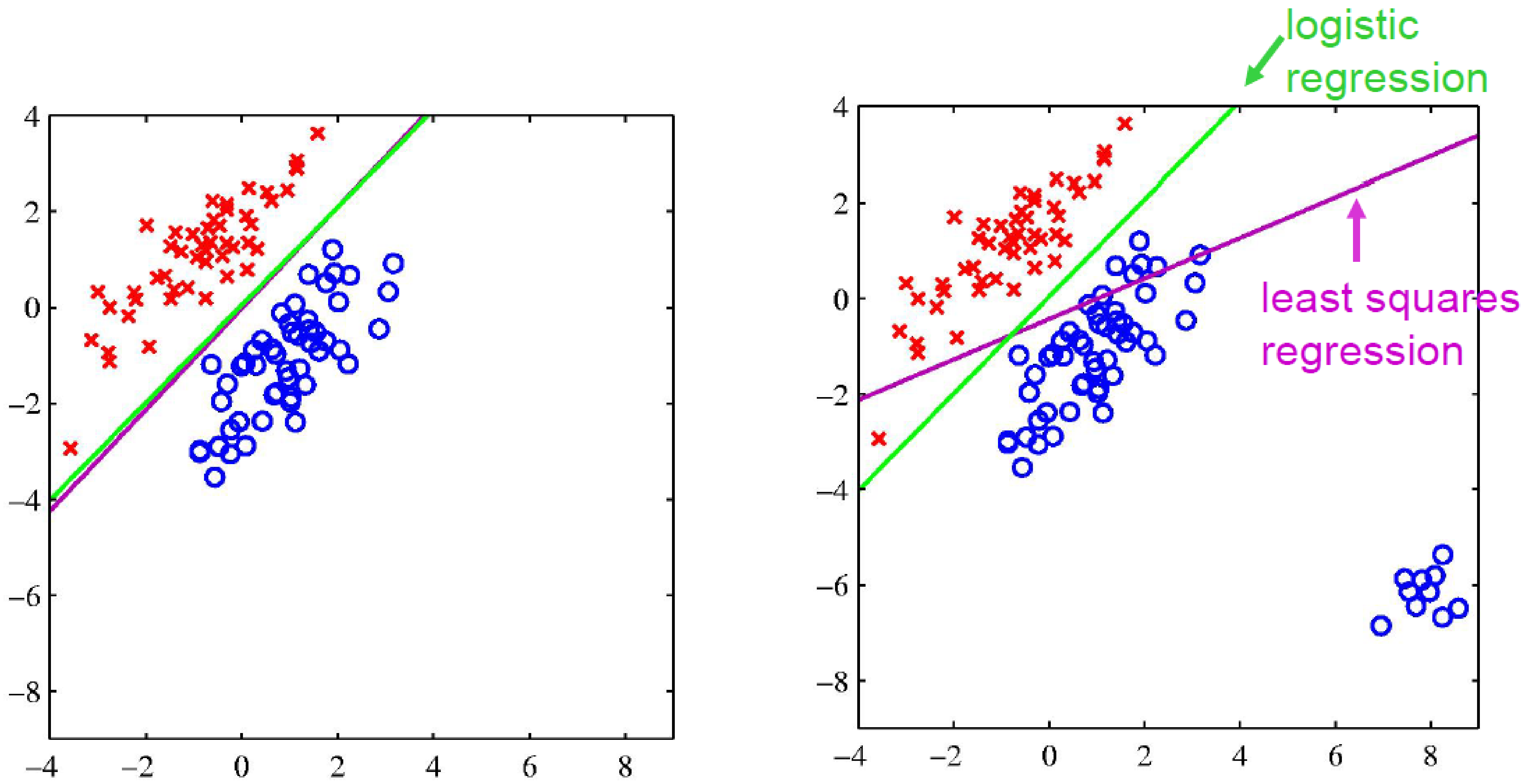
$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = \frac{\exp(-\mathbf{w}^T\mathbf{x})}{1 + \exp(-\mathbf{w}^T\mathbf{x})}.$$

Решающая граница

- Как выглядит решающая граница логистической регрессии?
- $P(y=0|\mathbf{x}) = P(y=1|\mathbf{x}) = 0.5$
- $P(y=1|\mathbf{x}) = g(\mathbf{w}^T\mathbf{x} + w_0) = 0.5$, где $g(z) = \frac{1}{1+\exp(-z)}$
- Решающая граница $\mathbf{w}^T\mathbf{x} + w_0 = 0$
- У логистической регрессии **линейная решающая граница**.



Сравнение линейной и логистической регрессии



- Логистическая регрессия менее подвержена выбросам.

Обучение логистической регрессии

- На входе имеются p-мерные вектора признаков $\mathbf{x} \in \mathbb{R}^p$
- Требуется обучить вектор коэффициентов $\mathbf{w} = [w_0, w_1, \ldots, w_p]$
- Логистическая регрессия является вероятностной моделью
- Можно воспользоваться **методом максимального правдоподобия**

Условное правдоподобие

- Целевые переменные $y^{(i)} \in \{0,1\}$. Вероятность появления набора данных $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})\}$ можно описать ***функцией правдоподобия***

$$CL(\mathbf{w}) = \prod_{i=1}^{n} p\left(y^{(i)} | \mathbf{x}^{(i)}; \mathbf{w}\right).$$

- Мы можем переписать каждую отдельную вероятность в следующем виде

$$p\left(y^{(i)} | \mathbf{x}^{(i)}; \mathbf{w}\right) = p\left(y = 1 | \mathbf{x}^{(i)}; \mathbf{w}\right)^{y^{(i)}} \times p\left(y = 0 | \mathbf{x}^{(i)}; \mathbf{w}\right)^{1-y^{(i)}}$$
$$= p\left(y = 1 | \mathbf{x}^{(i)}; \mathbf{w}\right)^{y^{(i)}} \times \left(1 - p\left(y = 1 | \mathbf{x}^{(i)}; \mathbf{w}\right)\right)^{1-y^{(i)}}$$

- Мы можем обучить модель максимизировав правдоподобие:

$$\max_{\mathbf{w}} CL(\mathbf{w}) = \max_{\mathbf{w}} \prod_{i=1}^{n} p\left(y^{(i)} | \mathbf{x}^{(i)}; \mathbf{w}\right)$$

- Проще работать с логарифмом функции правдоподобия $\log CL(\mathbf{w})$.

Функция потерь (loss function)

$$CL(\mathbf{w}) = \prod_{i=1}^{n} p\left(y^{(i)}|\mathbf{x}^{(i)};\mathbf{w}\right)$$
$$= \prod_{i=1}^{n} p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right)^{y^{(i)}} \times \left(1 - p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right)\right)^{1-y^{(i)}}.$$

- Обычно задачу максимизации заменяют задачей минимизации, можно записать следующую **функцию потерь**:

$$NLL(\mathbf{w}) = -\log CL(\mathbf{w})$$
$$= -\log \prod_{i=1}^{n} p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right)^{y^{(i)}} \times \left(1 - p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right)\right)^{1-y^{(i)}}$$
$$= -\sum_{i=1}^{n} y^{(i)} \log p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right) - \sum_{i=1}^{n}\left(1 - y^{(i)}\right) \log\left(1 - p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right)\right)$$
$$= -\sum_{i=1}^{n} y^{(i)} \log p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right) + \left(1 - y^{(i)}\right) \log\left(1 - p\left(y = 1|\mathbf{x}^{(i)};\mathbf{w}\right)\right).$$

Градиентный спуск

$$\min_{\mathbf{w}} NLL(\mathbf{w})$$

- Градиентный спуск: на каждой итерации вычисляется направление наиболь-шего убывание и выполняется движение в этом направлении с шагом η:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \frac{\partial NLL}{\partial \mathbf{w}},$$

- Какова зависимость от $\mathbf{w}$?

$$p(y = 1|\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x} + w_0)}$$

- Можно показать, что

$$\frac{\partial NLL}{\partial \mathbf{w}} = \sum_{i=1}^{n} \left(p\left(y = 1|\mathbf{x}^{(i)}; \mathbf{w}\right) - y_i\right)\mathbf{x}^{(i)}$$

Выводы

Преимущества логистической регрессии

- Легко расширяется на многоклассовую классификацию;
- Вероятностная интерпретация на предсказание класса;
- Высокая скорость обучения;
- Быстрая классификация (быстрый вывод, inference);
- Хорошая точность на множестве наборов данных;
- Устойчивость к переобучению;
- Значение коэффициентов может быть интерпретировано, как индикатор важности признака;

Менее хорошее свойство

- Линейная решающая граница (слишком просто для сложных задач).