Введение в анализ данных

Лекция 8

Линейная классификация

Евгений Соколов

sokolov.evg@gmail.com

НИУ ВШЭ, 2016

Организационное

- Скоро зачёт!
- Следите за обновлениями на wiki

План на сегодня

- Линейные модели в классификации
- Геометрия линейных моделей
- Обучение линейных классификаторов
- Логистическая регрессия
- Точность и полнота

Модель линейной классификации

Классификация

- $\mathbb{Y} = \{-1, +1\}$
- -1 — отрицательный класс
- $+1$ — положительный класс
- $a(x)$ должен возвращать одно из двух чисел

Линейная регрессия

$$a(x) = \boxed{w_0 + \sum_{j=1}^{d} w_j x^j}$$

Вещественное число!

Линейный классификатор

$$a(x) = \operatorname{sign}\left(w_0 + \sum_{j=1}^{d} w_j x^j\right)$$

Линейный классификатор

$$a(x) = \operatorname{sign}\left(w_0 + \sum_{j=1}^{d} w_j x^j\right)$$

Свободный коэффициент

Веса

Признаки

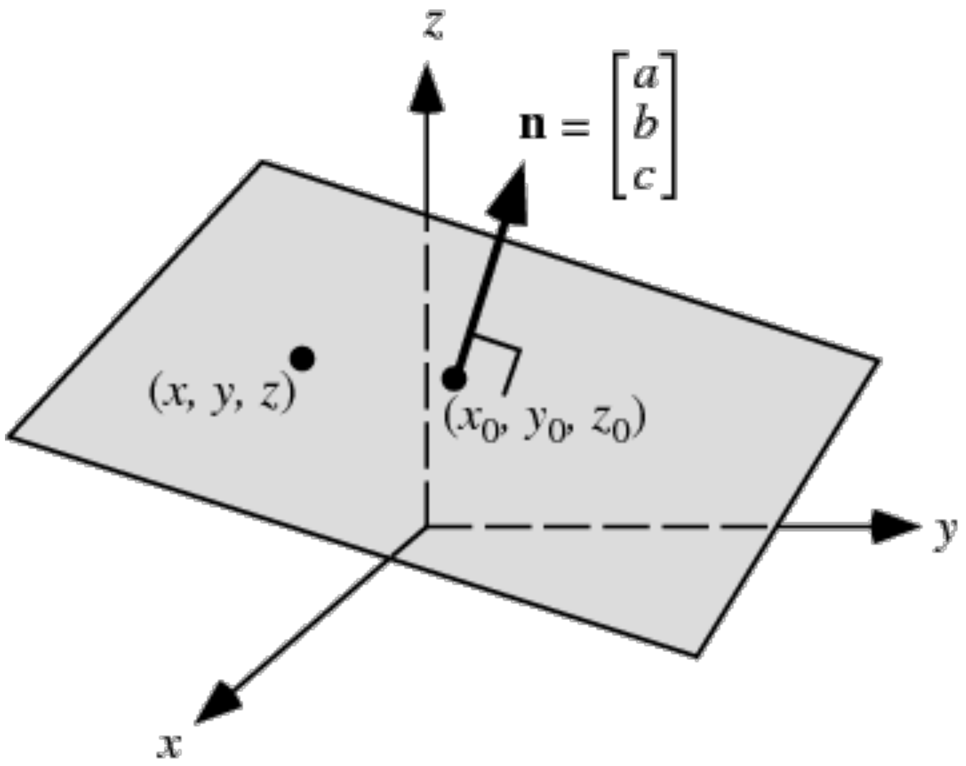
Линейный классификатор

- Добавим единичный признак

$$a(x) = \operatorname{sign} \sum_{j=1}^{d+1} w_j x^j = \operatorname{sign} \langle w, x \rangle$$

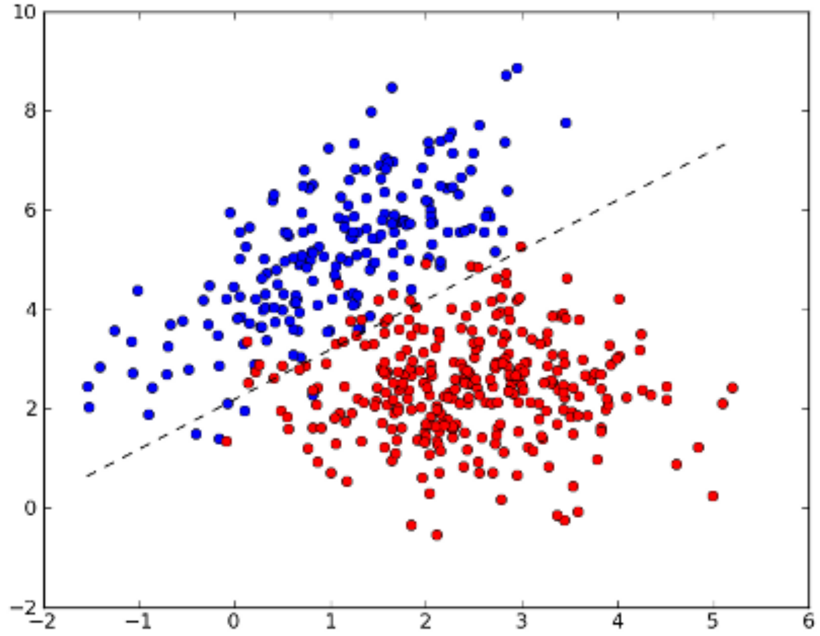
Геометрия линейного классификатора

Уравнение гиперплоскости: $\langle w, x\rangle = 0$



Геометрия линейного классификатора

- Линейный классификатор проводит гиперплоскость
- $\langle w, x \rangle < 0$ — объект «слева» от неё
- $\langle w, x \rangle > 0$ — объект «справа» от неё



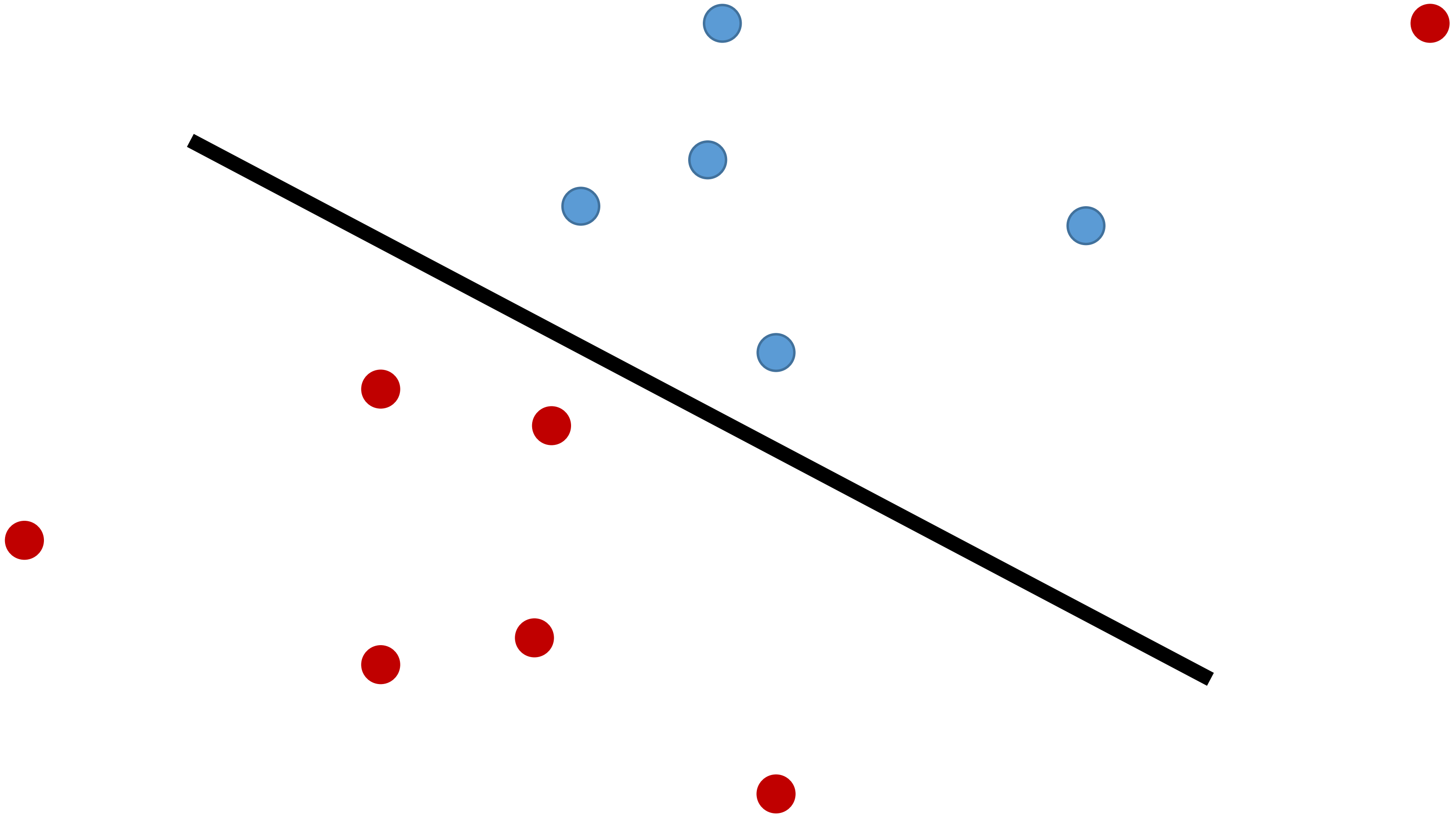
Геометрия линейного классификатора

- Расстояние от точки до гиперплоскости $\langle w, x \rangle = 0$:

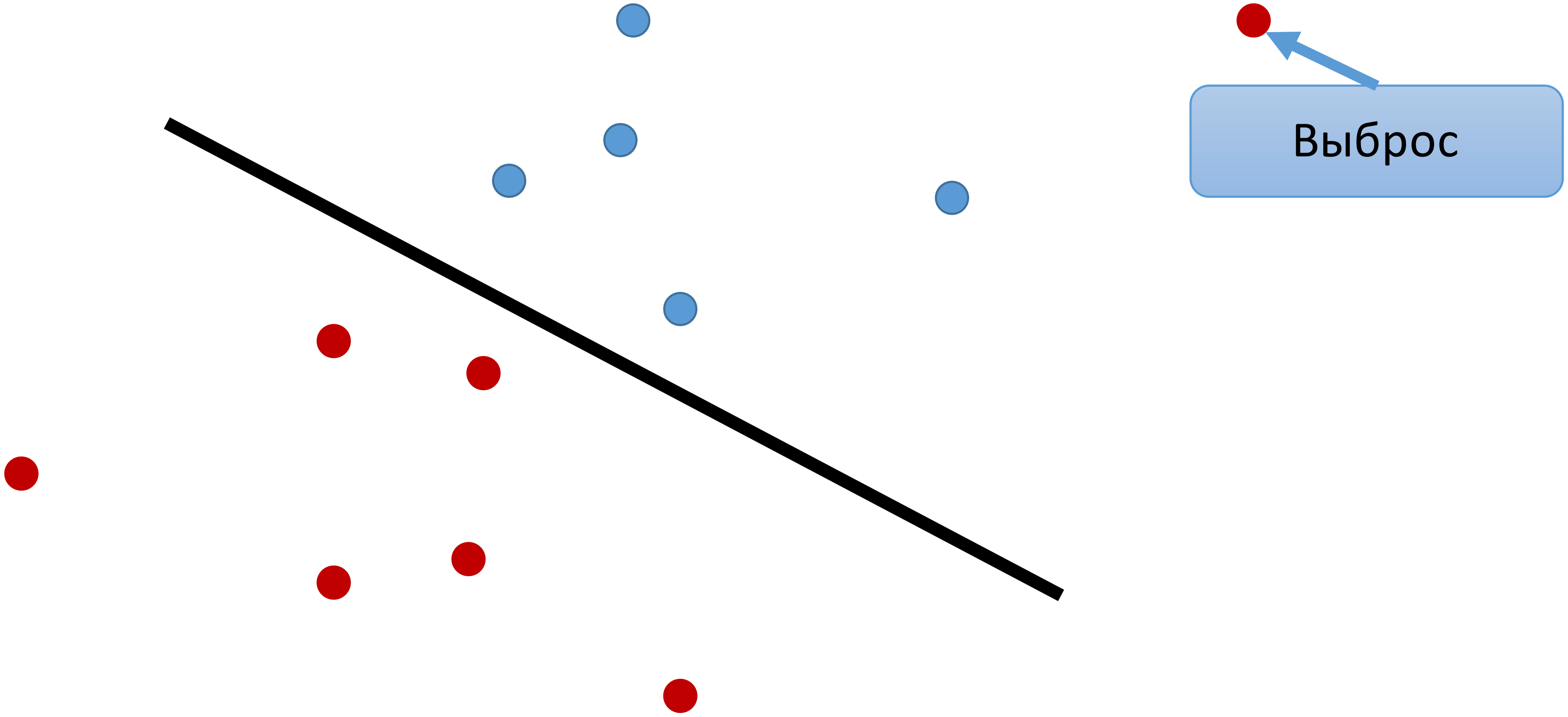
$$\frac{|\langle w, x \rangle|}{\|w\|}$$

- Чем больше $\langle w, x \rangle$, тем дальше объект от разделяющей гиперплоскости

Геометрия линейного классификатора

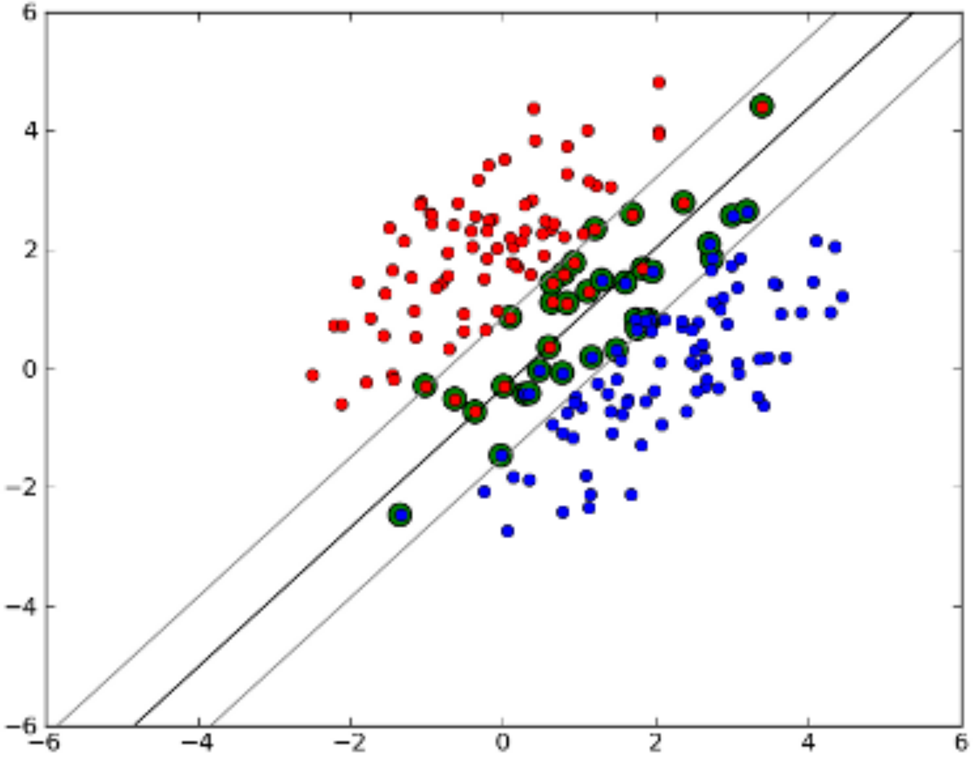


Геометрия линейного классификатора



Отступ

- $M_i = y_i \langle w, x_i \rangle$
- $M_i > 0$ — классификатор дает верный ответ
- $M_i < 0$ — классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности



Линейный классификатор

- Линейный классификатор разделяет два класса гиперплоскостью
- Чем больше отступ по модулю, тем дальше объект от гиперплоскости
- Знак отступа говорит о корректности предсказания

Функционал ошибки для классификации

Линейная регрессия

- Квадратичное отклонение:
$$L(a, y) = (a - y)^2$$
- Абсолютное отклонение:
$$L(a, y) = |a - y|$$

Линейная классификация

- Доля неправильных ответов:

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

Линейная классификация

$a(x)$	y
-1	-1
$+1$	$+1$
-1	-1
$\mathbf{+1}$	$\mathbf{-1}$
$+1$	$+1$

- Доля неправильных ответов:

$$\frac{1}{5} = 0.2$$

Линейная классификация

- Доля правильных ответов:

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

- На английском: accuracy

Линейная классификация

- Доля правильных ответов:

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

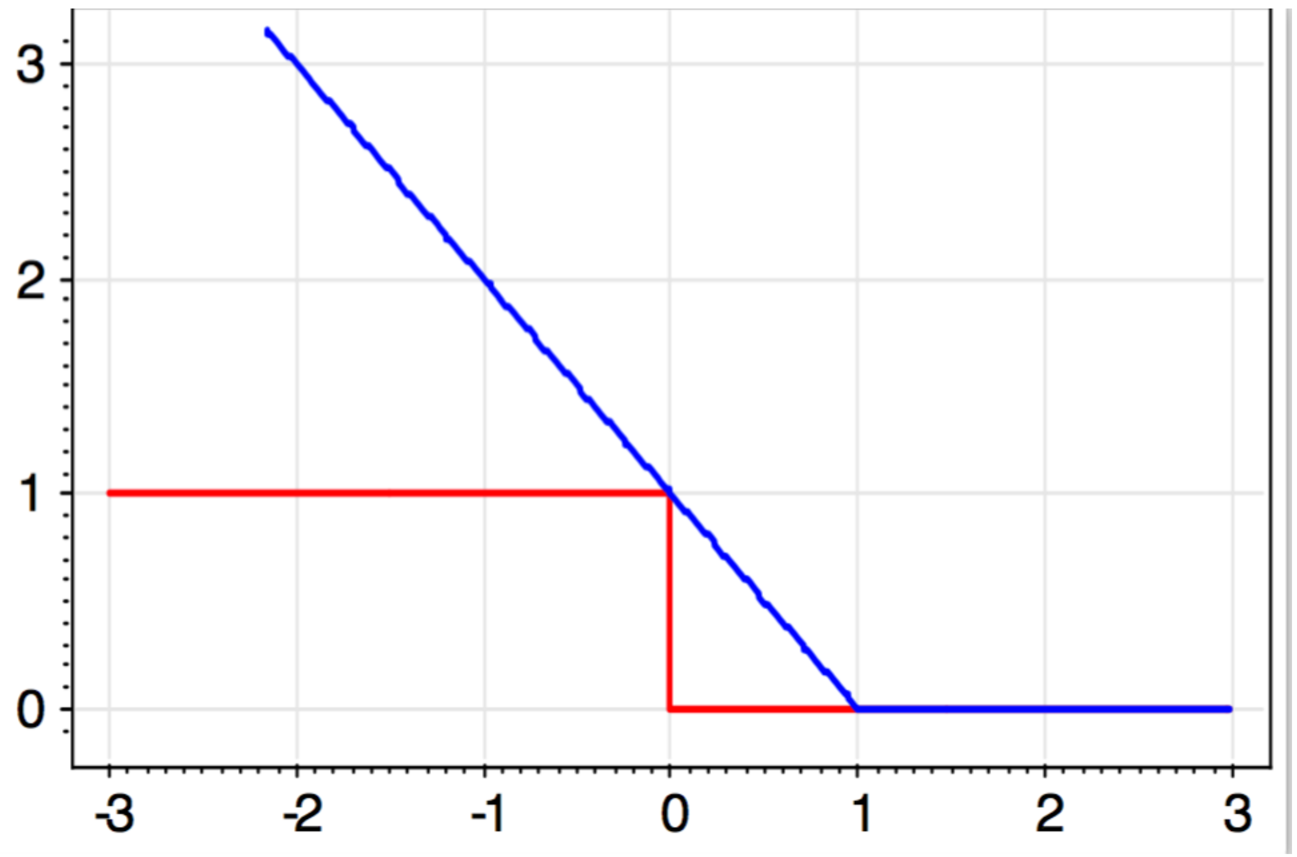
- На английском: accuracy
- ВАЖНО: не переводите это как «точность»!

Линейная классификация

- Доля неправильных ответов (через отступ):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0]$$

Пороговая функция потерь



Линейная классификация

- Доля неправильных ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0]$$

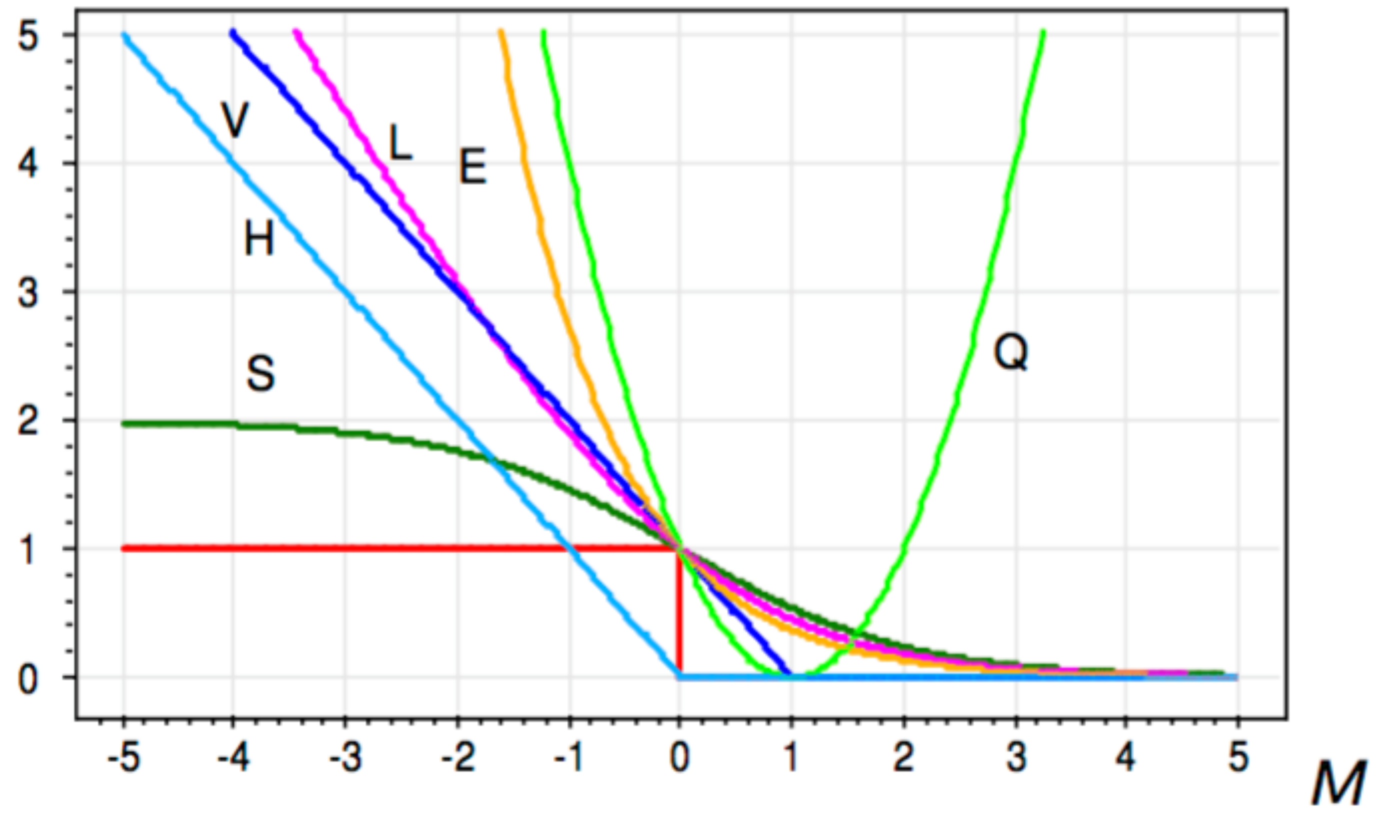
- Разрывная функция
- Непонятно, как оптимизировать

Оценка функции потерь

- Возьмем любую гладкую оценку пороговой функции:

$$[M < 0] \leq \tilde{L}(M) = \tilde{L}(y\langle w, x\rangle)$$

Примеры оценок



Примеры оценок

- $\tilde{L}(M) = \log_2(1 + \exp(-M))$ — логистическая
- $\tilde{L}(M) = \exp(-M)$ — экспоненциальная
- $\tilde{L}(M) = \max(0, 1 - M)$ — кусочно-линейная

Оценка функции потерь

- Возьмем любую гладкую оценку пороговой функции:

$$[M < 0] \leq \tilde{L}(M)$$

- Оценим через нее функционал ошибки:

$$Q(a, X) \leq \tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(M_i)$$

Оценка функции потерь

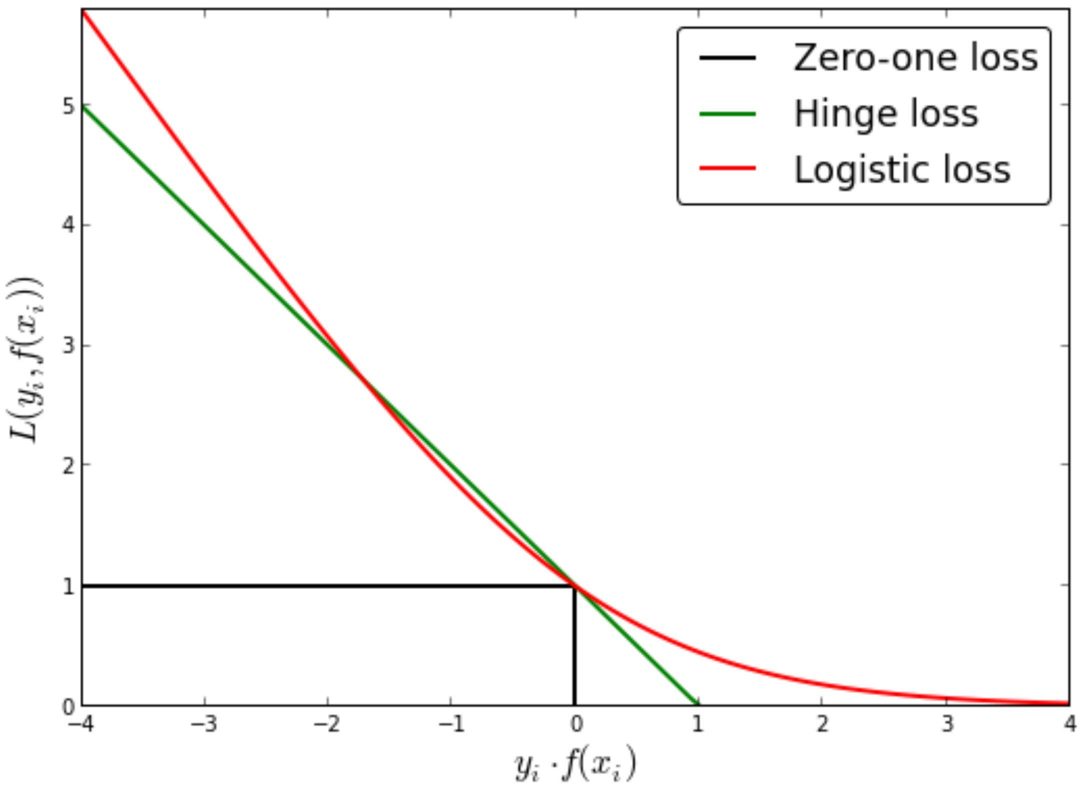
$$\frac{1}{\ell}\sum_{i=1}^{\ell}[M_i < 0] \leq \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(M_i) \to \min_a$$

Минимизируем верхнюю оценку

Надеемся, что доля ошибок тоже уменьшится

Примеры оценок

- $\tilde{L}(a, y) = \ln(1 + \exp(-ya))$ — логистическая



Логистическая функция потерь

$$\tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \log_2(1 + \exp(-y_i \langle w, x_i \rangle))$$

1. Выписали индикатор ошибки через отступ
2. Заменили пороговую функцию потерь на гладкую функцию

Обучение

- Обучение — с помощью любых методов оптимизации
- Например, градиентный спуск

- Борьба с переобучением: регуляризация (так же, как в линейной регрессии)

Логистическая регрессия

Логистическая регрессия

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log_2(1+\exp(-y_i\,\langle w, x_i\rangle)) \to \min_w$$

Оценивание вероятностей

- $P(y = 1 \mid x) = \pi(x)$

Оценивание вероятностей

- Кредитный скоринг
- Стратегия: выдавать кредит только клиентам с $\pi(x) > 0.9$
- 10% невозвращённых кредитов — нормально

Оценивание вероятностей

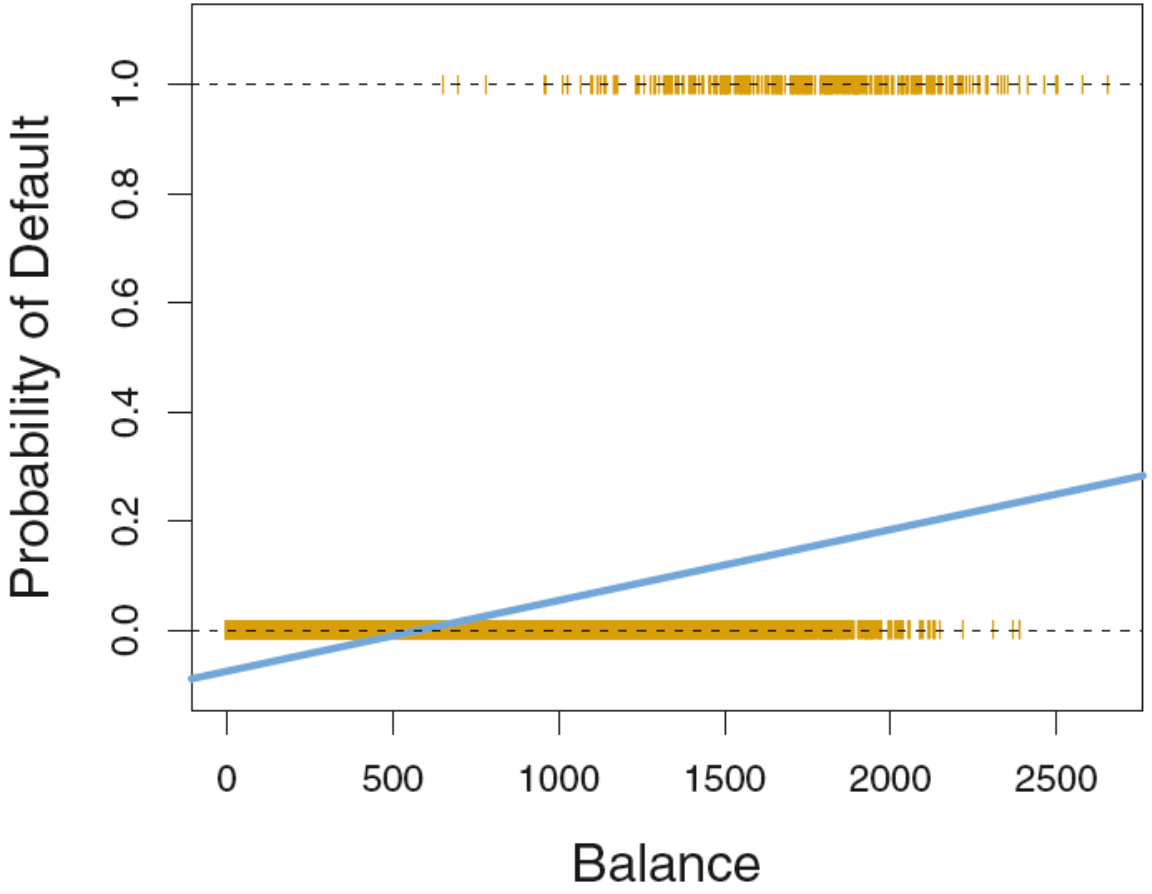
- Баннерная реклама
- $\pi(x)$ — вероятность, что пользователь кликнет по рекламе
- $c(x)$ — прибыль в случае клика
- $\pi(x)c(x)$ — хотим оптимизировать

Оценивание вероятностей

- $P(y = 1 \mid x) = \pi(x)$
- $\pi(x)$ — вещественное число
- Классификатор не подходит

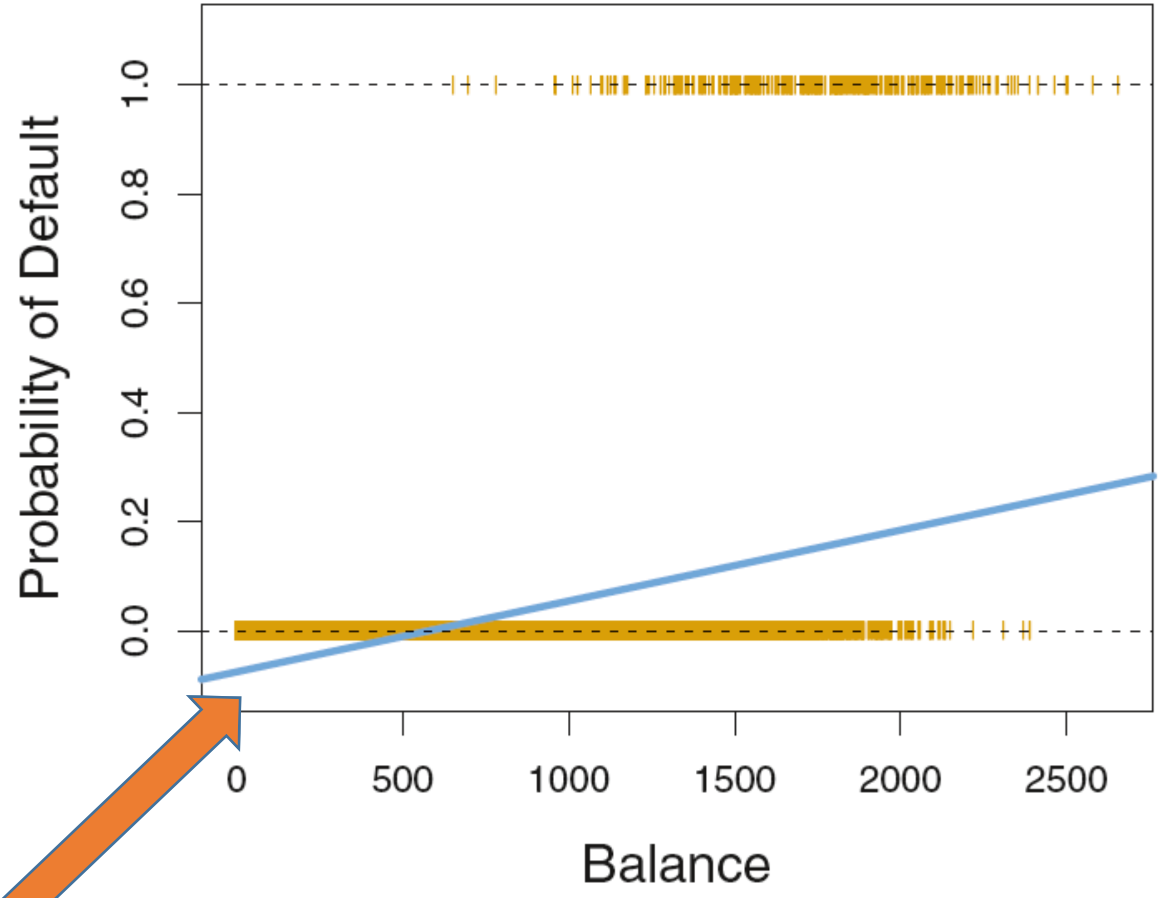
Регрессия?

- $\pi(x) \approx \langle w, x \rangle = w_1 x + w_0$



Регрессия?

- $\pi(x) \approx \langle w, x \rangle = w_1 x + w_0$



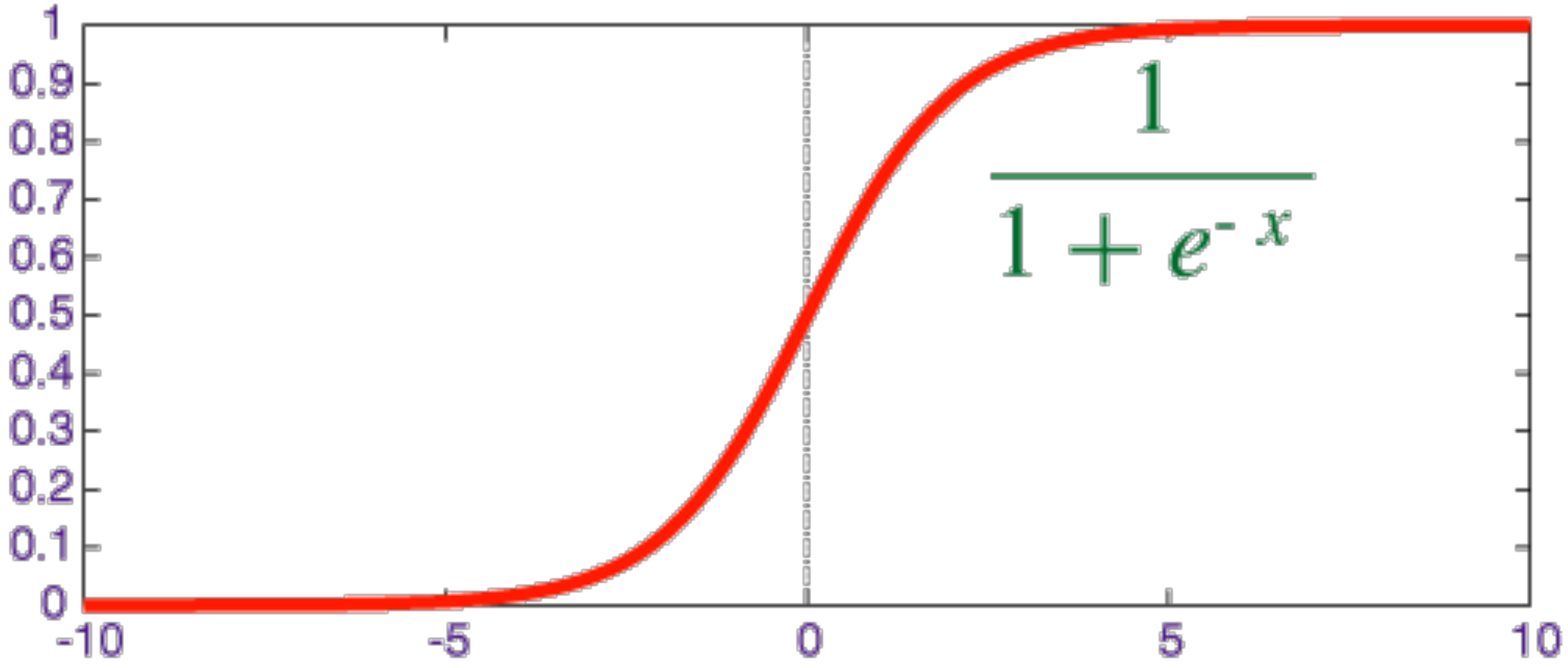
Отрицательная вероятность о_О

Регрессия?

$$\pi(x) \approx \sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

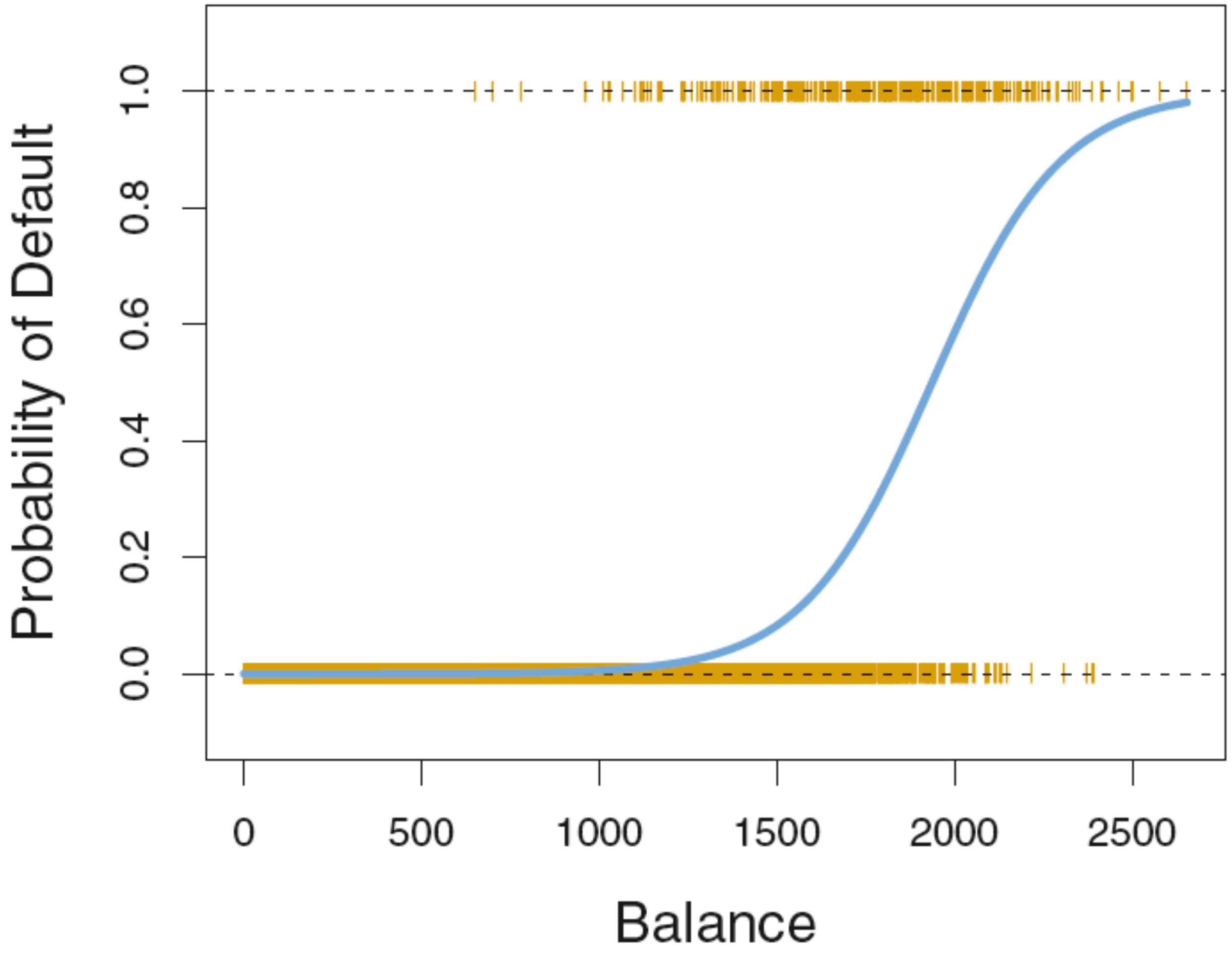
Сигмоида

Сигмоида



Логистическая регрессия

- $\pi(x) \approx \sigma(\langle w, x \rangle)$



Логистическая регрессия

- Как оптимизировать?
- Если $y_i = +1$, то $\langle w, x_i \rangle \to +\infty$
- Если $y_i = -1$, то $\langle w, x_i \rangle \to -\infty$

Логистическая регрессия

- Как оптимизировать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

Логистическая регрессия

- Как оптимизировать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

$$\sum_{i=1}^{\ell} \left\{ [y_i = 1] \sigma(\langle w, x_i \rangle) + [y_i = -1] \big(1 - \sigma(\langle w, x_i \rangle)\big) \right\} \to \max_w$$

Логистическая регрессия

- Как оптимизировать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

$$\sum_{i=1}^{\ell} \left\{ [y_i = 1]\sigma(\langle w, x_i \rangle) + [y_i = -1]\big(1 - \sigma(\langle w, x_i \rangle)\big) \right\} \to \max_w$$

- Слишком слабый штраф
- Если $y_i = +1$ и $\sigma(\langle w, x_i \rangle) = 0$, то штраф $= 1$

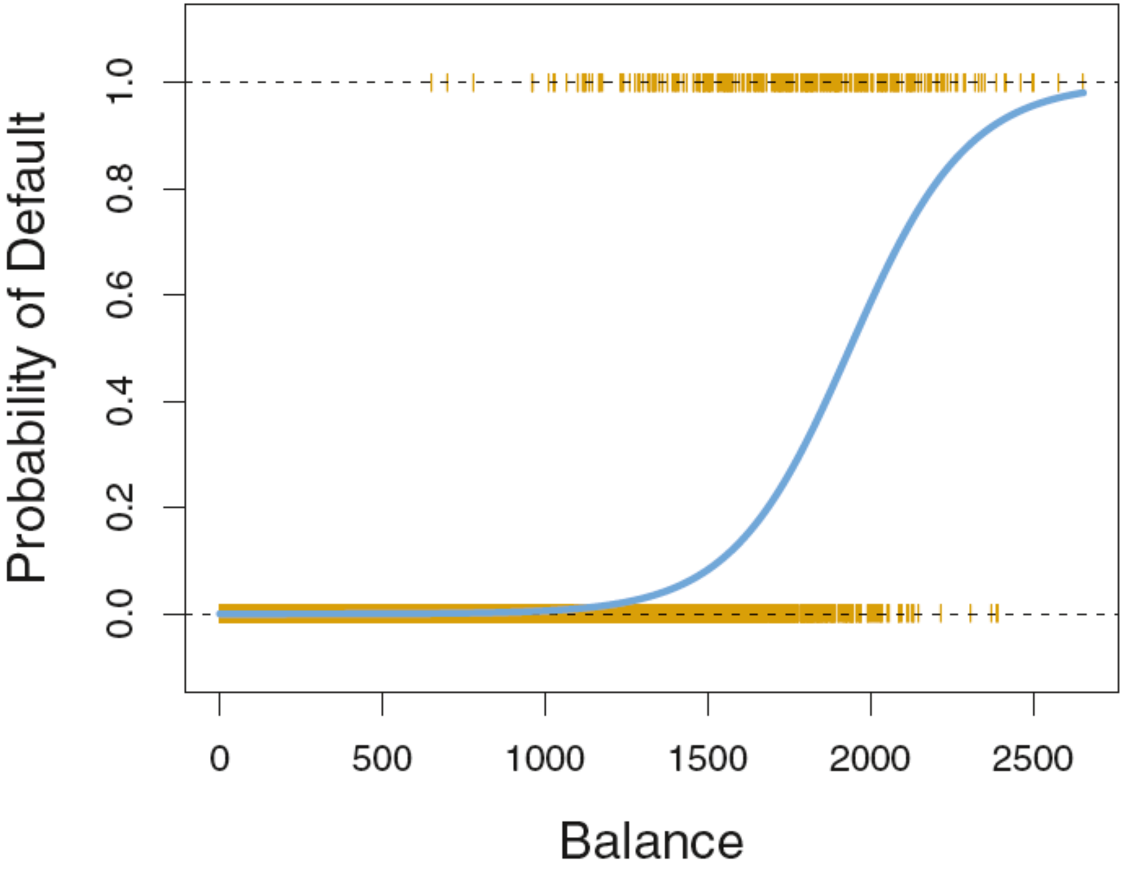
Логистическая регрессия

- Как оптимизировать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

$$\sum_{i=1}^{\ell} \left\{ [y_i = 1] \log_2 \sigma(\langle w, x_i \rangle) + [y_i = -1] \log_2 \big(1 - \sigma(\langle w, x_i \rangle)\big) \right\} \to \max_w$$

- Если $y_i = +1$ и $\sigma(\langle w, x_i \rangle) = 0$, то штраф $= -\infty$

Логистическая регрессия



Логистическая регрессия

- Если вспомнить арифметику, то получим эквивалентную задачу:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log_2(1+\exp(-y_i\,\langle w, x_i\rangle)) \to \min_w$$

Логистическая регрессия

- Линейная модель классификации: $a(x) = \operatorname{sign} \langle w, x \rangle$
- Позволяет оценивать вероятности: $\pi(x) = \sigma(\langle w, x \rangle)$
- Обучение: градиентный спуск

Метрики качества классификации

Качество классификации

- Доля неправильных ответов:

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

Качество классификации

- Доля правильных ответов (accuracy):

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

Несбалансированные выборки

- Пример:
  - Класс -1: 950 объектов
  - Класс +1: 50 объектов
- $a(x) = -1$
- Доля правильных ответов: 0.95

Несбалансированные выборки

- q_0 — доля объектов самого крупного класса
- Для разумных алгоритмов:

$$\text{accuracy} \in [q_0, 1]$$

- Если получили большой accuracy — посмотрите на баланс классов

Цены ошибок

- Пример: кредитный скоринг
- Модель 1:
  - 80 кредитов вернули
  - 20 кредитов не вернули
- Модель 2:
  - 48 кредитов вернули
  - 2 кредита не вернули
- Кто лучше?

Цены ошибок

- Что хуже?
  - Выдать кредит «плохому» клиенту
  - Не выдать кредит «хорошему» клиенту
- Доля верных ответов не учитывает цены ошибок

Матрица ошибок

	$\boldsymbol{y = 1}$	$\boldsymbol{y = -1}$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

Матрица ошибок

- Модель $a_1(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	80	20
$a(x) = -1$	20	80

- Модель $a_2(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	48	2
$a(x) = -1$	52	98

Точность (precision)

- Можно ли доверять классификатору при $a(x) = 1$?

$$\text{precision}(a, X) = \frac{TP}{TP + FP}$$

Точность (precision)

- Модель $a_1(x)$:

	$\boldsymbol{y = 1}$	$\boldsymbol{y = -1}$
$a(x) = 1$	80	20
$a(x) = -1$	20	80

- $\text{precision}(a_1, X) = 0.8$

- Модель $a_2(x)$:

	$\boldsymbol{y = 1}$	$\boldsymbol{y = -1}$
$a(x) = 1$	48	2
$a(x) = -1$	52	98

- $\text{precision}(a_2, X) = 0.96$

Полнота (recall)

- Как много положительных объектов находит классификатор?

$$\mathrm{recall}(a, X) = \frac{TP}{TP + FN}$$

Полнота (recall)

- Модель $a_1(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	80	20
$a(x) = -1$	20	80

- $\text{recall}(a_1, X) = 0.8$

- Модель $a_2(x)$:

	$y = 1$	$y = -1$
$a(x) = 1$	48	2
$a(x) = -1$	52	98

- $\text{recall}(a_2, X) = 0.48$

Кредитный скоринг

- Неудачных кредитов должно быть не больше 5%
- Ограничение: $\mathrm{precision}(a, X) \geq 0.95$
- Максимизируем полноту

Медицинская диагностика

- Надо найти не менее 80% больных
- Ограничение: $\mathrm{recall}(a, X) \geq 0.8$
- Максимизируем точность

Несбалансированные выборки

- $\text{accuracy}(a, X) = 0.99$
- $\text{precision}(a, X) = 0.33$
- $\text{recall}(a, X) = 0.1$

	$y = 1$	$y = -1$
$a(x) = 1$	10	20
$a(x) = -1$	90	10000

Резюме

- Линейные классификаторы разделяют классы гиперплоскостью
- Логистическая регрессия — классификация и оценка вероятности
- Качество классификации: доля правильных ответов, точность и полнота

На следующей лекции

- Оценивание качества и подбор гиперпараметров
- Кросс-валидация
- Подробнее про точность и полноту
- Качество оценок вероятности: площади под кривыми