Машинное обучение

Лекция 5

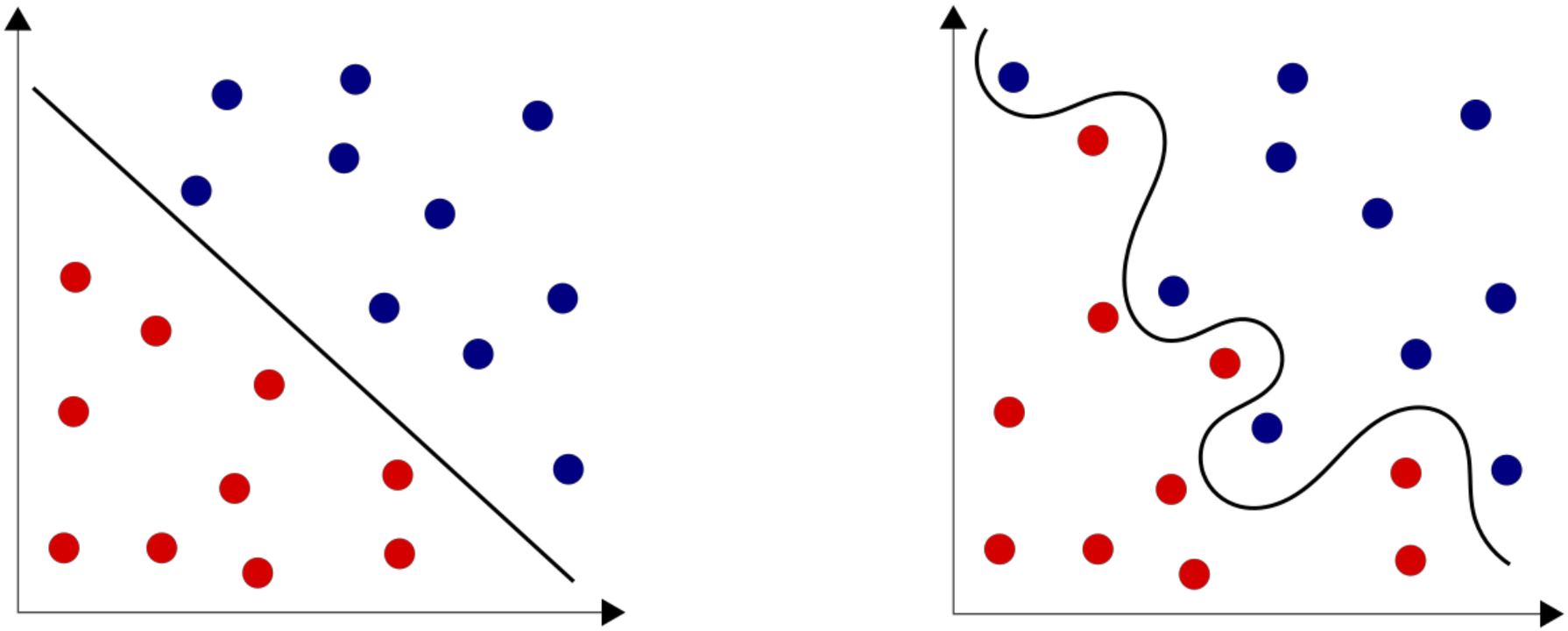
Линейные модели (Продолжение)

Власов Кирилл Вячеславович

2018

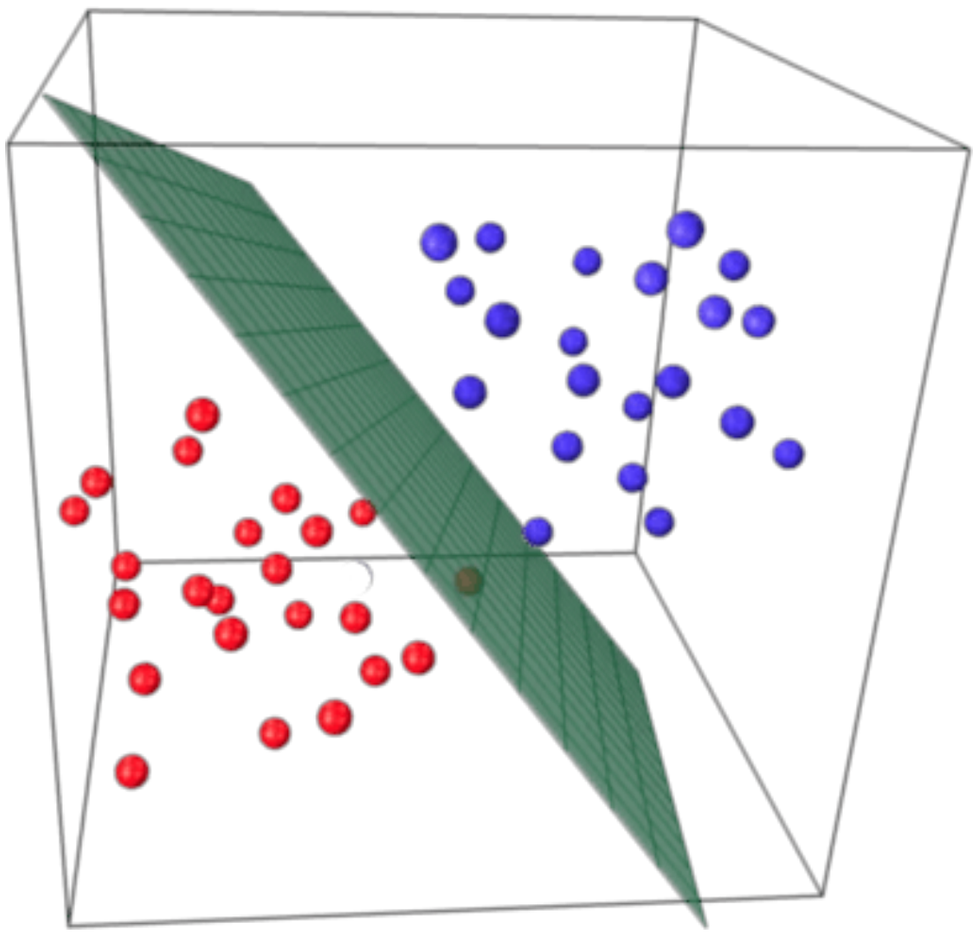
Линейные модели в задаче классификации

Основная идея:
Предполагаем, что существует такая гиперплоскость, которая делит пространство на два полупространства в каждом из которых одно из двух значений целевого класса.



Если существует гиперплоскость которой можно разделить пространство на два класса без ошибок, то обучающая выборка называется *линейно разделимой*

Линейные модели в задаче классификации



Дана обучающая выборка:

$$X_l = \{ (x_1 , y_1), \dots , (x_l , y_l) \}$$

Для задачи классификации - Целевая переменная задана конечным числом меток

$$(x_1 , y_1) \in \mathbb{R}^m \times \mathbb{Y}, \mathbb{Y} = \{-1; 1\}$$

Простейший классификатор:

$$a(x) = sign(\langle w, x \rangle + x_0) = sign\left(\overrightarrow{w}^T \cdot x\right)$$

$\overrightarrow{w}$ – нормаль гиперплоскости

$\overrightarrow{w}^T \cdot x_i$ – расстояние от гиперплоскости до x_i, знак показывает отношение к классу

Рассмотрим пример!

Линейные модели в задаче классификации

$$a(x) = sign(\langle w, x\rangle + x_0) = sign\left(\overrightarrow{w}^T \cdot x\right)$$

доля правильных ответов (accuracy):

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

Линейные модели в задаче классификации

$$a(x) = sign(\langle w, x \rangle + x_0) = sign\left(\overrightarrow{w}^T \cdot x\right)$$

доля правильных ответов (accuracy):

$$\mathcal{Q}(a, X) = \frac{1}{l} \sum_{i=1}^{l} \left[a(x_i) = y_i\right] \rightarrow max$$

Линейные модели в задаче классификации

$$a(x) = sign(\langle w, x\rangle + x_0) = sign\left(\overrightarrow{w}^T \cdot x\right)$$

доля правильных ответов (accuracy):

$$\mathcal{Q}(a, X) = \frac{1}{l}\sum_{i=1}^{l}\left[a(x_i) = y_i\right] \to max$$

доля неправильных ответов:

$$\mathcal{Q}(a, X) = \frac{1}{l}\sum_{i=1}^{l}\left[a(x_i) \neq y_i\right] = \frac{1}{l}\sum_{i=1}^{l}\left[sign\left(\langle w, x_i\rangle\right) \neq y_i\right] \to min$$

<u>Проблемы:</u>

1. *Функционал дискретный относительно весов ⇒ мы не сможем искать минимум с помощью градиентных методов.*
2. *Функционал может иметь несколько глобальных минимумов ⇒ может быть много способов добиться оптимального количества ошибок.*

Линейные модели в задаче классификации

$$a(x) = sign(\langle w, x\rangle + x_0) = sign\left(\overrightarrow{w}^T \cdot x\right)$$

доля правильных ответов (accuracy):

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

доля неправильных ответов:

$$\mathcal{Q}(a, X) = \frac{1}{l}\sum_{i=1}^{l}\left[a(x_i) \neq y_i\right] = \frac{1}{l}\sum_{i=1}^{l}\left[sign\left(\langle w, x_i\rangle\right) \neq y_i\right] \to min$$

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

$M_i = y_i\langle w, x_i\rangle$ – отступ (margin)

Знак отступа говорит о корректности ответа классификатора (положительный отступ соответствует правильному ответу, отрицательный неправильному)
абсолютная величина M – характеризует степень уверенности классификатора в своём ответе.

Линейные модели в задаче классификации

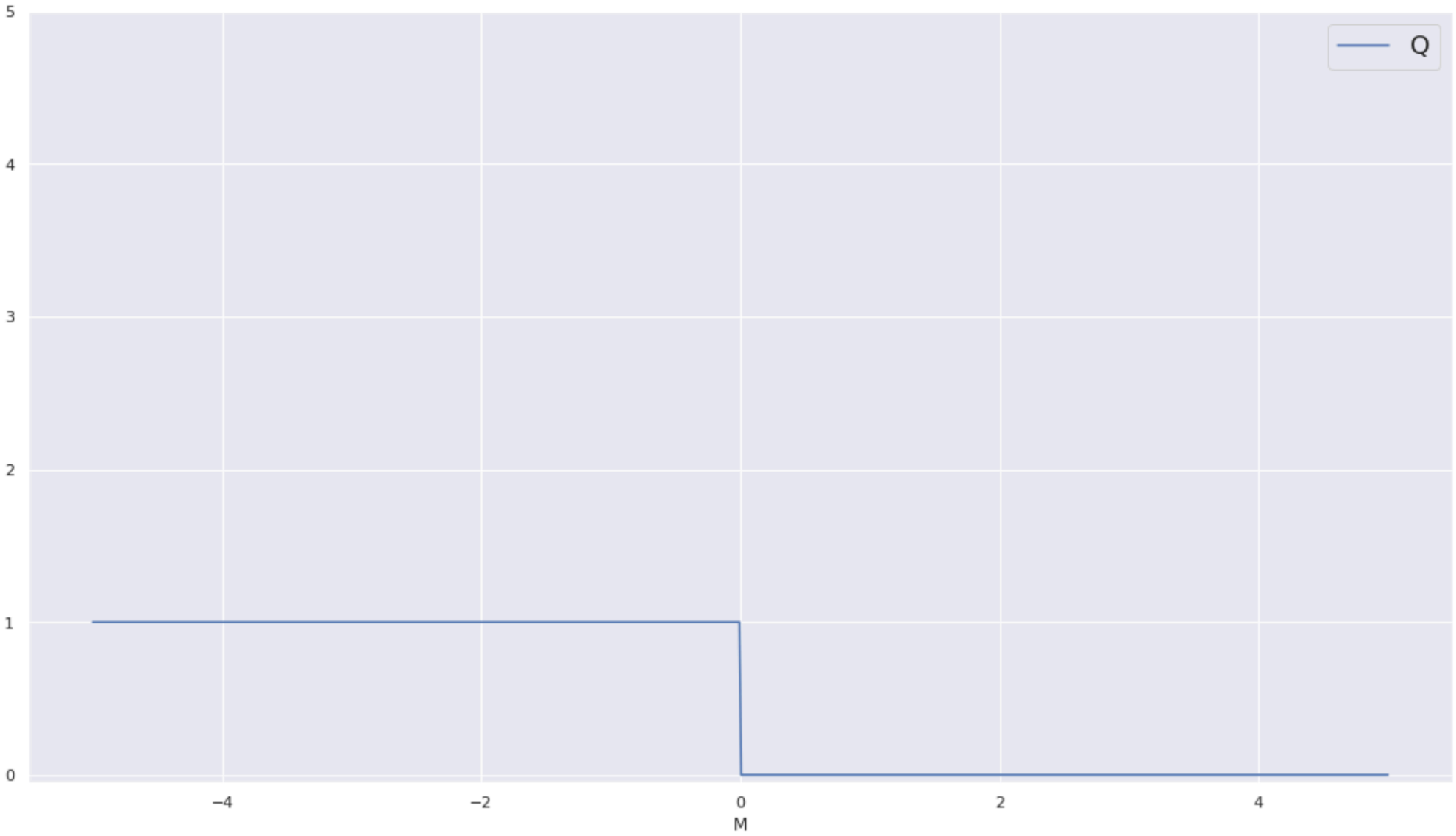
$$\mathcal{Q}(a, X) = \frac{1}{l} \sum_{i=1}^{l} \left[\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0 \right] \to min$$

$$L(M) = [M < 0]$$

Линейные модели в задаче классификации

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

$L(M) = [M < 0]$ – *пороговая функции потерь*



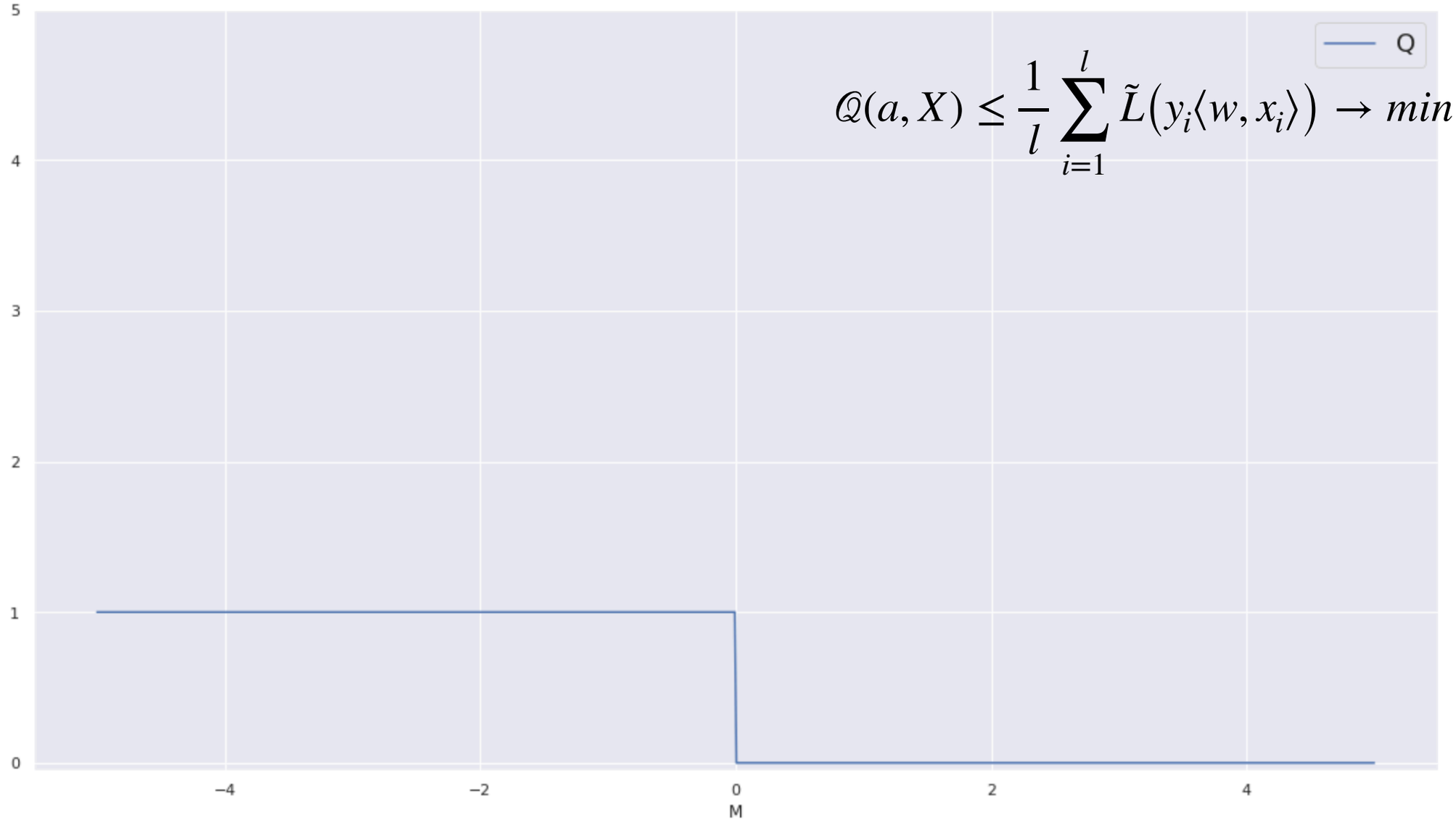
Линейные модели в задаче классификации

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

$L(M) \le \tilde{L}(M)$ – *верхняя оценка функции потерь*

$L(M) = [M < 0]$ – *пороговая функции потерь*

Если верхнюю оценку удастся приблизить к нулю, то и доля неправильных ответов тоже будет близка к нулю



$$\mathcal{Q}(a, X) \le \frac{1}{l}\sum_{i=1}^{l}\tilde{L}\left(y_i\langle w, x_i\rangle\right) \to min$$

Линейные модели в задаче классификации

$$\tilde{L}(M) = (1 - M)^2$$

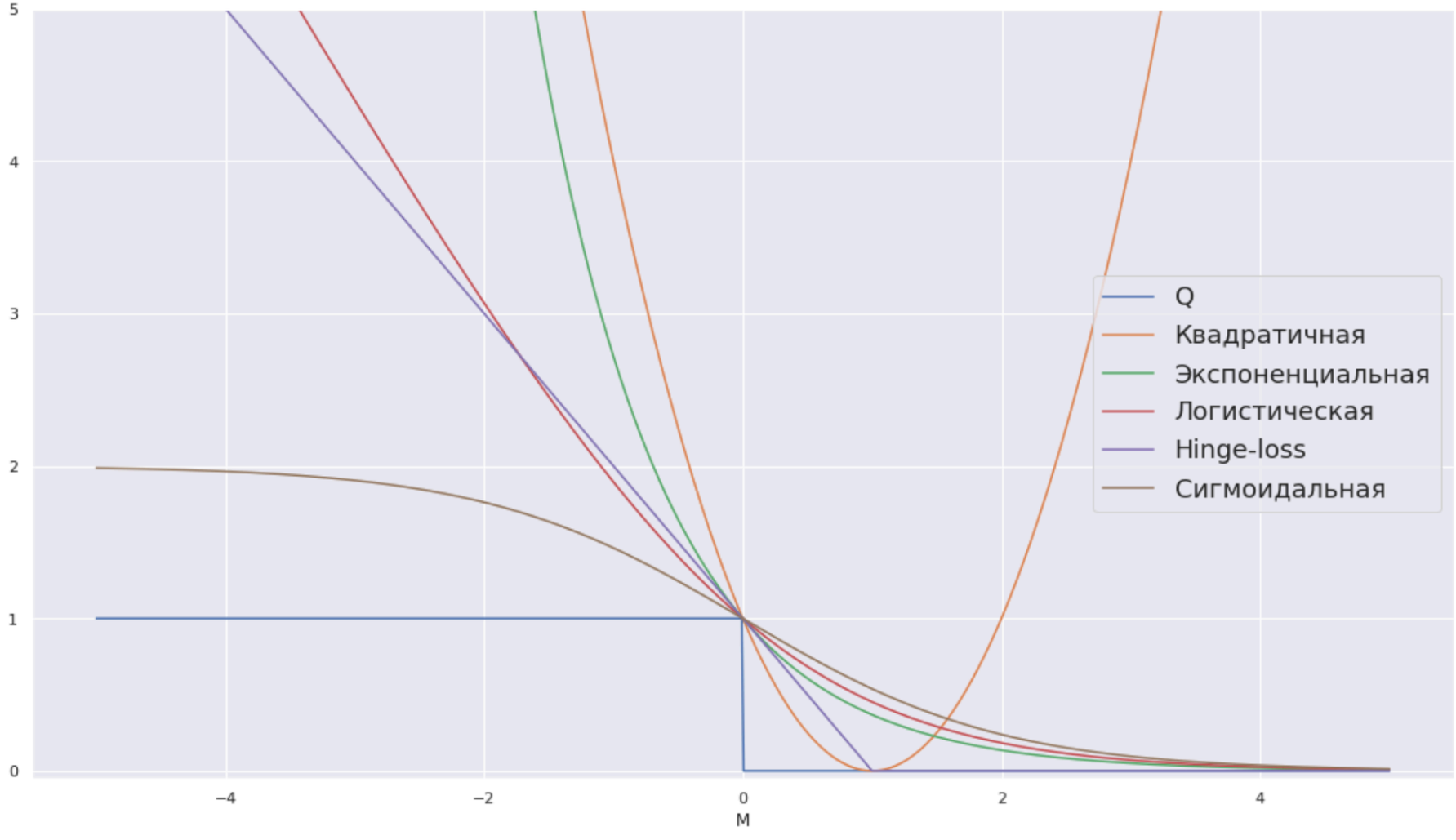
$$\tilde{L}(M) = e^{-M}$$

$$\tilde{L}(M) = \log(1 + e^{-M})$$

$$\tilde{L}(M) = (1 - M)_+ = \max(0, 1 - M)$$

$$\tilde{L}(M) = \frac{2}{1 + e^{-M}}$$

$$L(M) = [M < 0]$$ – *пороговая функции потерь*



Рассмотрим пример!

Линейные модели в задаче классификации

$$\tilde{L}(M) = (1 - M)^2$$

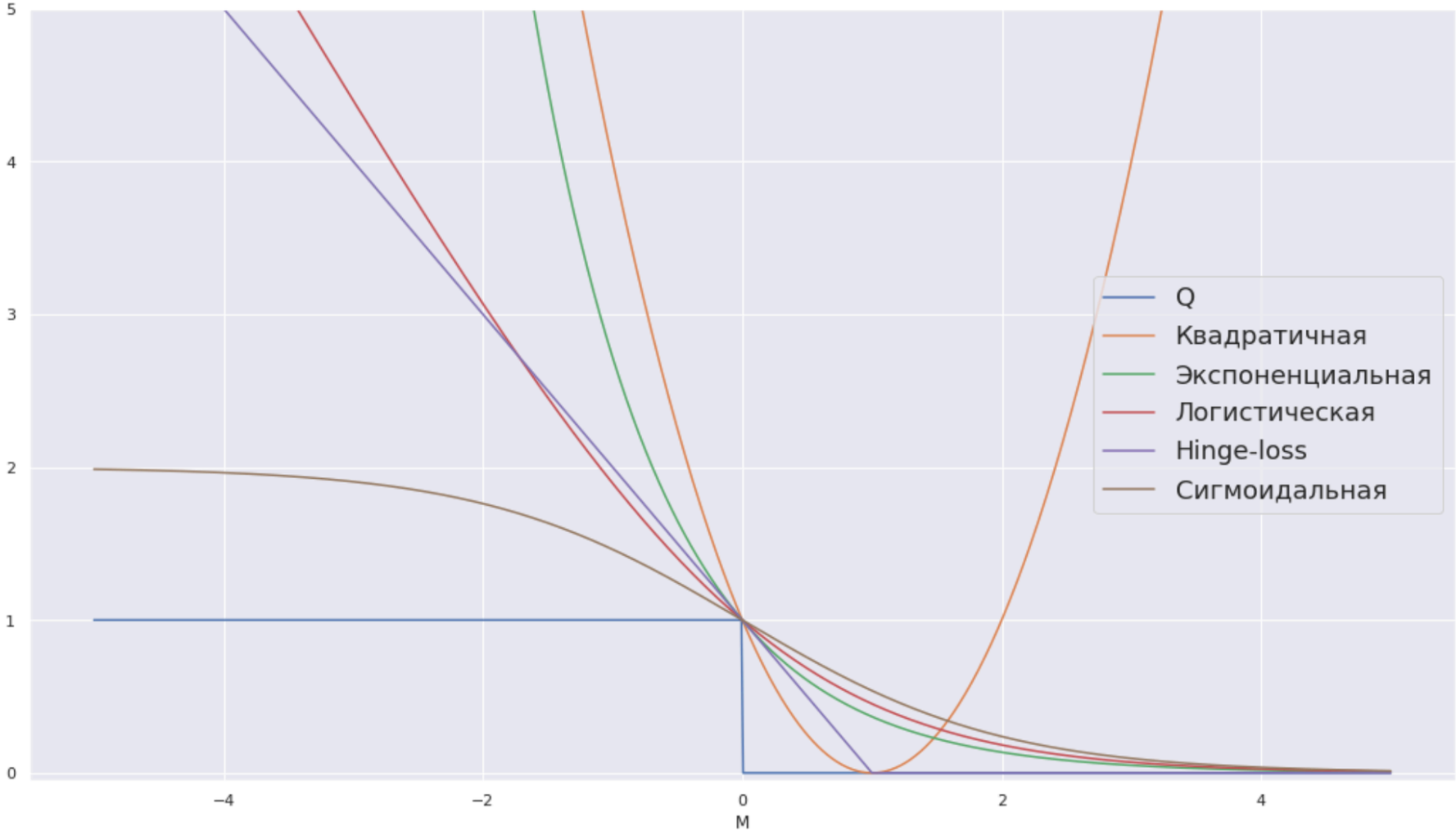
$$\tilde{L}(M) = e^{-M}$$

$$\tilde{L}(M) = \log(1 + e^{-M})$$

$$\tilde{L}(M) = (1 - M)_+ = \max(0, 1 - M)$$

$$\tilde{L}(M) = \frac{2}{1 + e^{-M}}$$

$L(M) = [M < 0]$ – *пороговая функции потерь*



Линейные модели в задаче классификации

$$\tilde{L}(M) = \log(1 + e^{-M})$$

Минимизация эмпирического риска:

$$\mathcal{Q}(a, X) = \frac{1}{l} \sum_{i=1}^{l} \log(1 + e^{-y_i \langle w, x_i \rangle}) \to min$$

$$a(x) = sign\left(\vec{w}^T \cdot x\right)$$

Линейные модели в задаче классификации

$$\tilde{L}(M) = \log(1 + e^{-M})$$

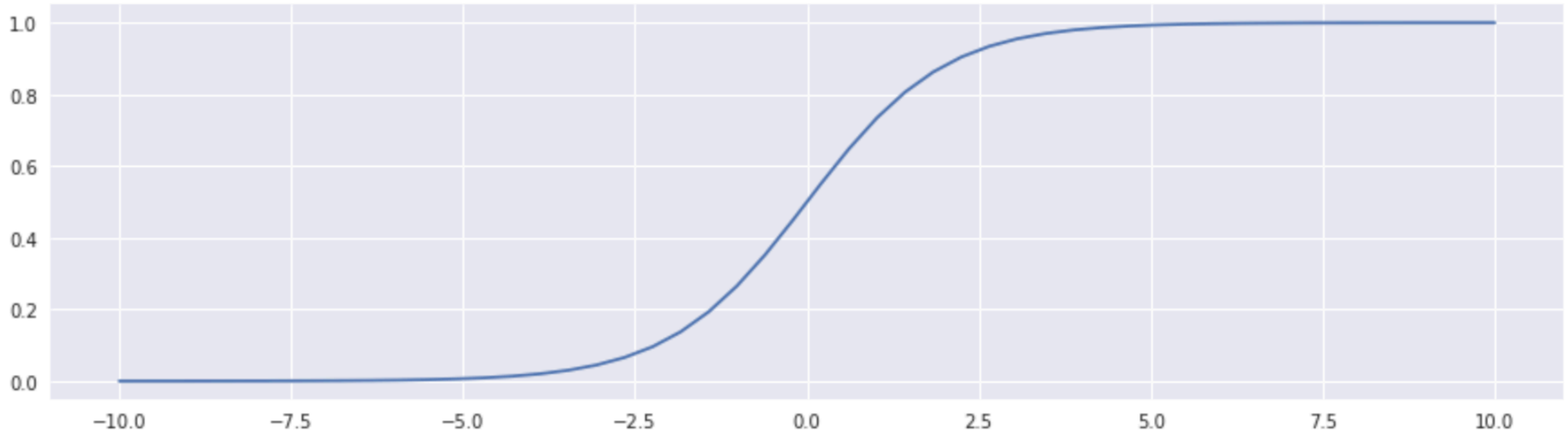
Минимизация эмпирического риска:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} \log(1 + e^{-y_i \langle w, x_i \rangle}) \to min$$

$$a(x) = sign\left(\vec{w}^T \cdot x\right)$$

Оценка апостериорных вероятностей принадлежности классам

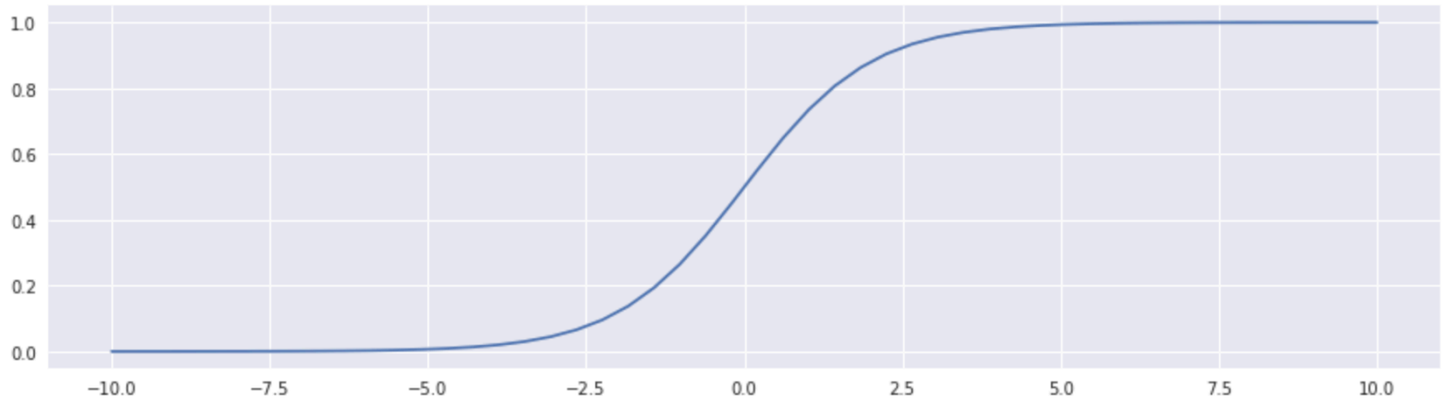
$$\vec{w}^T \cdot x \in R \qquad f(x_i, w) = \sigma(z) = \frac{1}{1 + e^{-z}} \in [0; 1]$$



Линейные модели в задаче классификации

$$y_i = f(x_i, w) + \varepsilon$$

$$f(x_i, w) = \sigma(z) = \frac{1}{1 + e^{-z}}$$



Шансы

$$odds = \frac{p}{1-p} \in [0; \infty] \qquad \ln(odds) \in R$$

где p вероятность, что событие состоится (в нашем случае, что класс примет значение 1):

$$\ln(odds_+) = ln(\frac{p}{1-p}) = ln(p) - ln(1-p)$$

$$\ln(odds\_) = ln(\frac{1-p}{p}) = ln(1-p) - ln(p)$$

$$\ln(odds_+) = -\ln(odds\_) = w^T \cdot x \qquad odds = e^{w^T x} \quad \Rightarrow \quad p = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

Линейные модели в задаче классификации

$$P(y_i = 1 \,|\, x_i, w) = \sigma(w^T x) \qquad P(y_i = -1 \,|\, x_i, w) = \sigma(-w^T x)$$

$$P(y = y_i \,|\, x_i, w) = \sigma(y_i w^T x)$$

Правдоподобие (вероятность наблюдать вектор *y* при заданных значениях *X* и *w*)

Делаем предположение: объекты приходят независимо, из одного распределения

$$P(\overrightarrow{y} \,|\, X, w) = \prod_{i=1}^{l} P(y = y_i \,|\, x_i, w) \to max$$

Так как логарифм монотонно возрастающая функция, то оценка *w* максимизирующая логарифм, будет максимизировать и само правдоподобие

$$\log P(\overrightarrow{y} \,|\, X, w) = \sum_{i=1}^{l} \log \sigma(y_i w^T x) = \sum_{i=1}^{l} \log \frac{1}{1 + e^{-y_i \langle w, x_i \rangle}} = -\sum_{i=1}^{l} \log(1 + e^{-y_i \langle w, x_i \rangle})$$

$$\mathcal{Q}(a, X) = \frac{1}{l} \sum_{i=1}^{l} \log(1 + e^{-y_i \langle w, x_i \rangle}) \to min$$

Рассмотрим пример!

Ссылки

Открытый курс машинного обучения

Репозитории Евгения Соколова