

Машинное обучение

Лекция 3

Линейные модели (Продолжение)

Власов Кирилл Вячеславович



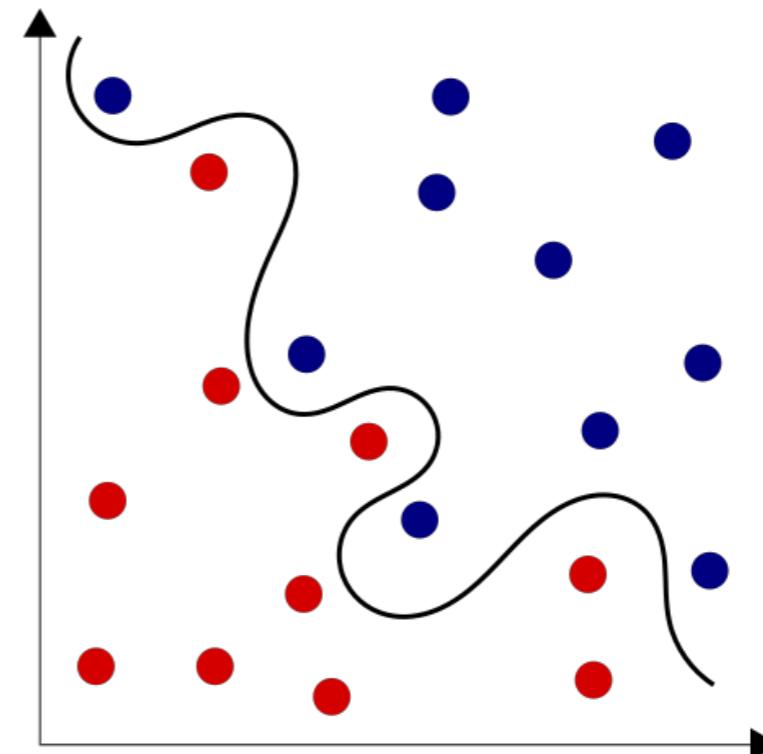
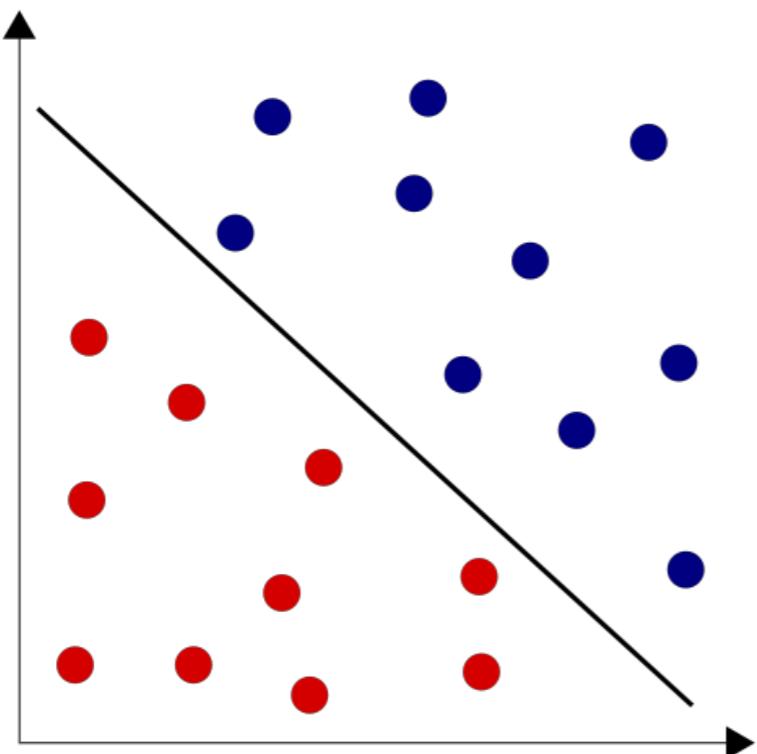
2018

Линейные модели в задаче классификации

Линейные модели в задаче классификации

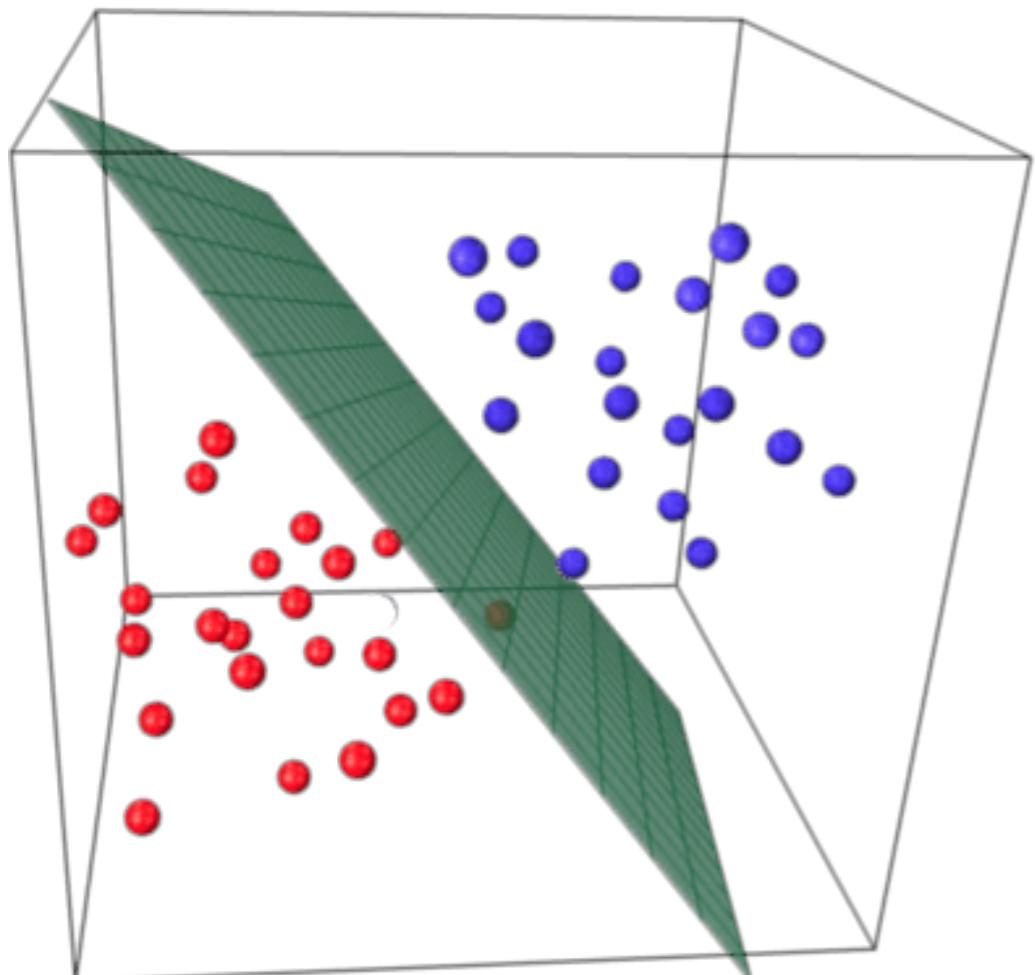
Основная идея:

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:

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

\vec{w} – нормаль гиперплоскости

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

Линейные модели в задаче классификации

$$a(x) = \text{sign}(\langle w, x \rangle + x_0) = \text{sign}(\vec{w}^T \cdot x)$$

доля правильных ответов (accuracy):

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

Линейные модели в задаче классификации

$$a(x) = \text{sign}(\langle w, x \rangle + x_0) = \text{sign}(\vec{w}^T \cdot x)$$

доля правильных ответов (accuracy):

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

Линейные модели в задаче классификации

$$a(x) = \text{sign}(\langle w, x \rangle + x_0) = \text{sign}(\vec{w}^T \cdot x)$$

доля правильных ответов (accuracy):

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

доля неправильных ответов:

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

Проблемы:

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

Линейные модели в задаче классификации

$$a(x) = \text{sign}(\langle w, x \rangle + x_0) = \text{sign}(\vec{w}^T \cdot x)$$

доля правильных ответов (accuracy):

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

доля неправильных ответов:

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

$$\mathcal{Q}(a, X) = \frac{1}{l} \sum_{i=1}^l \underbrace{[y_i \langle w, x_i \rangle < 0]}_{M_i} \rightarrow \min \quad M_i = y_i \langle w, x_i \rangle \text{ -- отступ (margin)}$$

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

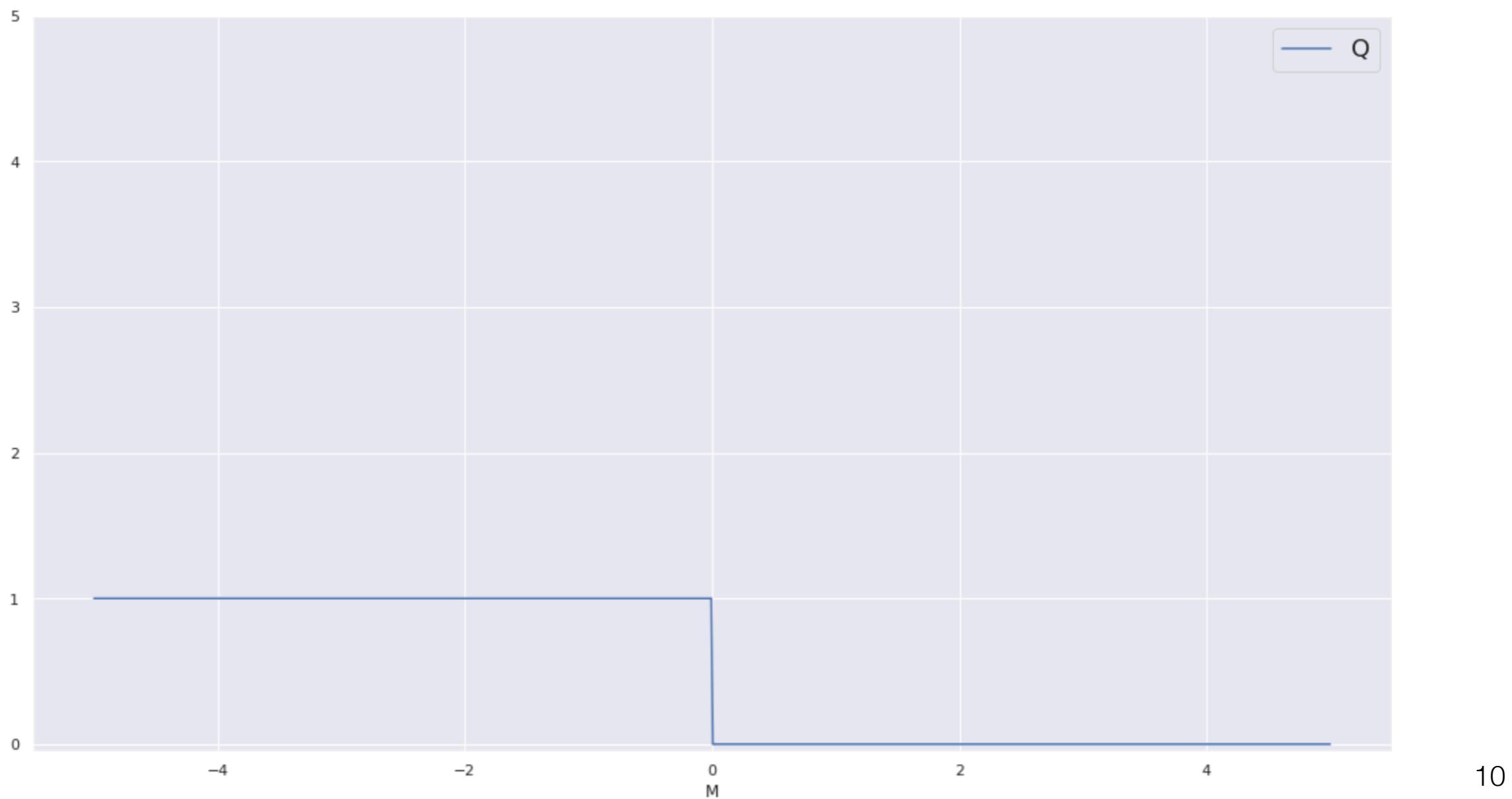
$$\mathcal{Q}(a, X) = \frac{1}{l} \sum_{i=1}^l \underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0 \rightarrow \min$$

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

Линейные модели в задаче классификации

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

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



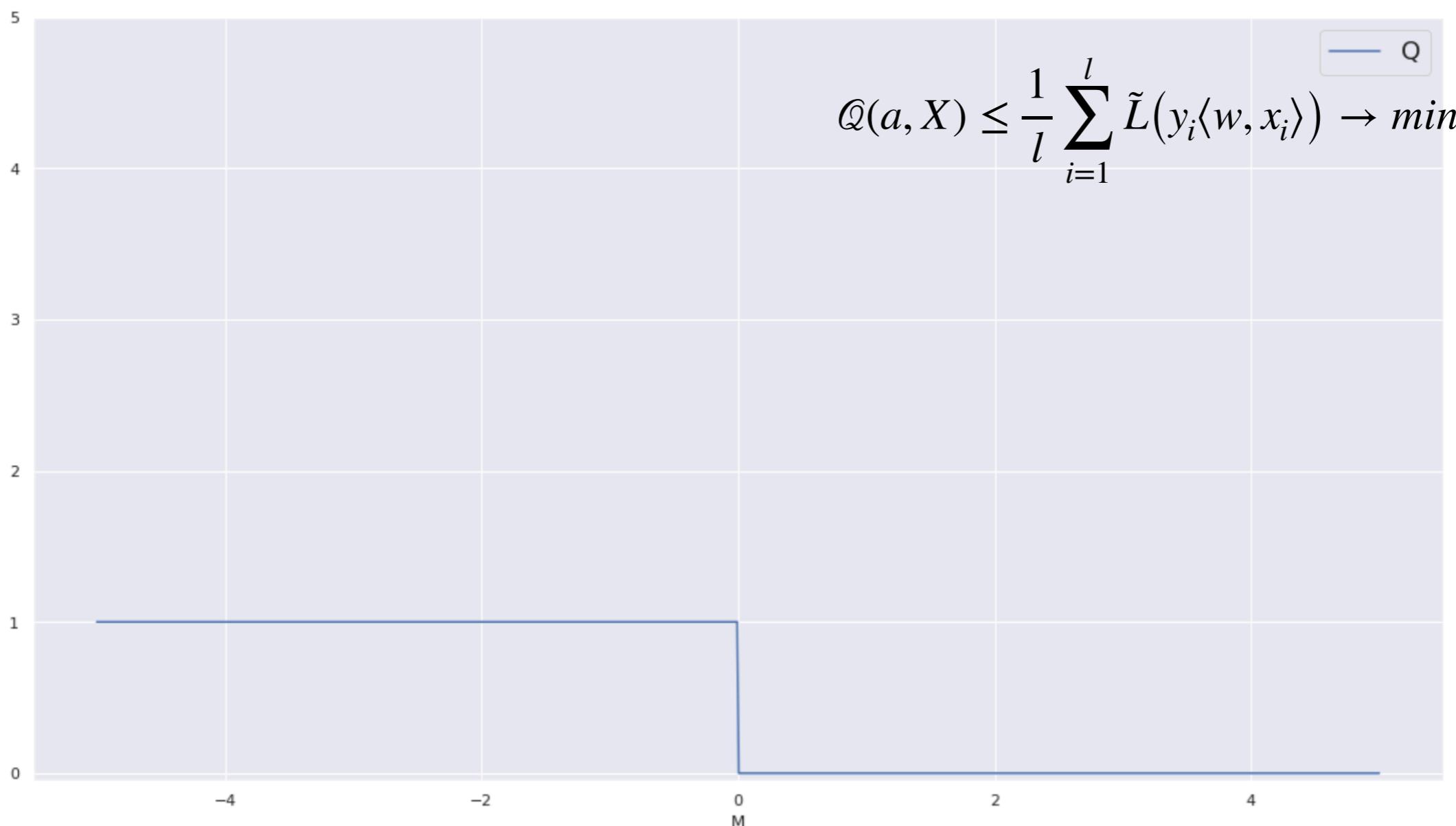
Линейные модели в задаче классификации

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

$L(M) \leq \tilde{L}(M)$ – 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



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

Линейные модели в задаче классификации

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

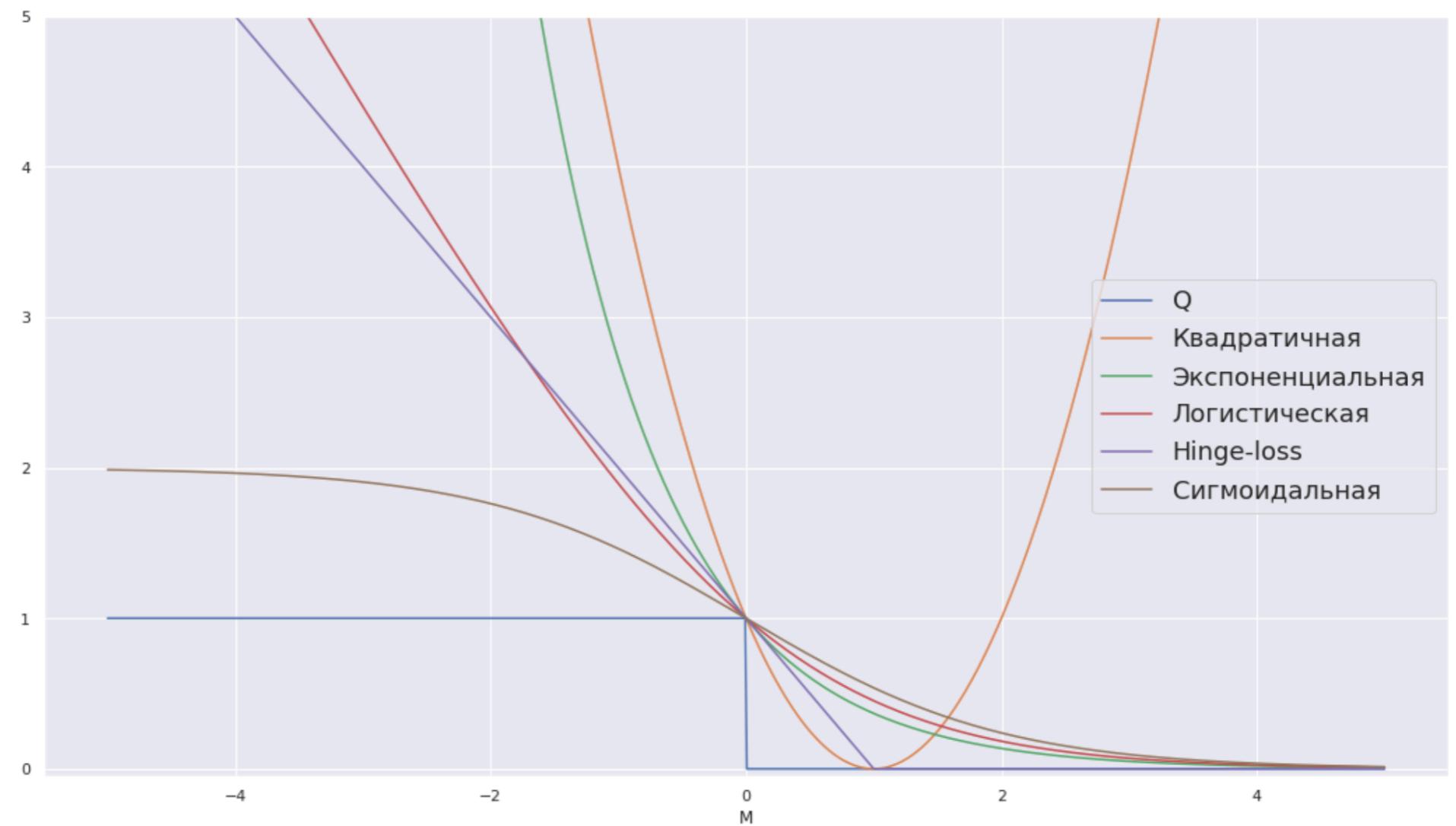
$L(M) = [M < 0]$ – пороговая функция потерь

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

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

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

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



Линейные модели в задаче классификации

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

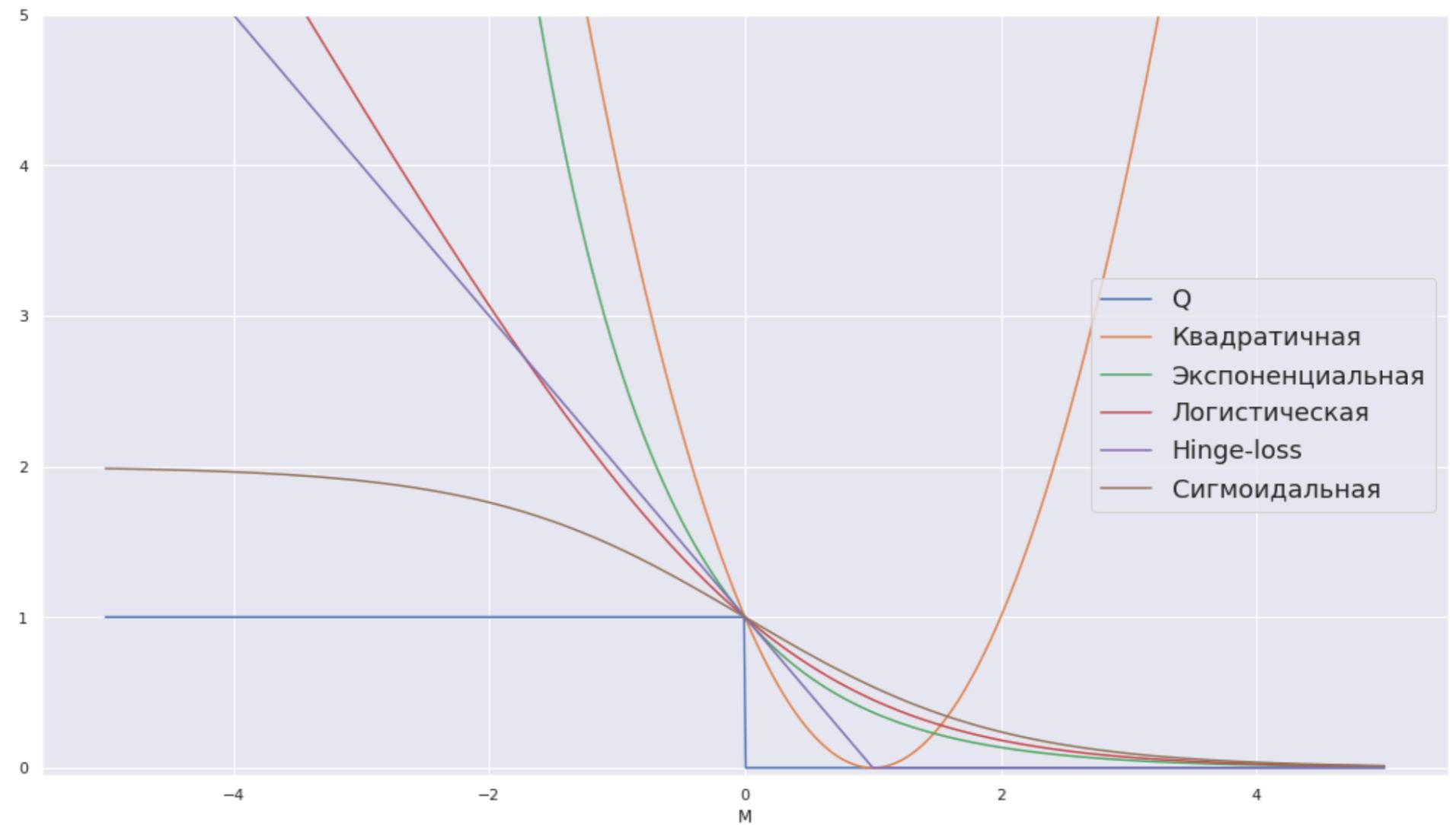
$L(M) = [M < 0]$ – пороговая функции потерь

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

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

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

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



Линейные модели в задаче классификации

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

Минимизация эмпирического риска:

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

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

Линейные модели в задаче классификации

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

Минимизация эмпирического риска:

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

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

Как оценить апостериорную вероятность принадлежности к классам, с помощью взвешенной суммы признаков?

$$\vec{w}^T \cdot x \in R$$

Линейные модели в задаче классификации

Шансы

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

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

Линейные модели в задаче классификации

Шансы

$$odds = \frac{p}{1 - p} \in [0; \infty] \quad \ln(odds) \in R$$
$$\vec{w}^T \cdot x \in R$$

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

Линейные модели в задаче классификации

Шансы

$$odds = \frac{p}{1 - p} \in [0; \infty] \quad \ln(odds) \in R$$
$$\vec{w}^T \cdot x \in R$$

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

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

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

$$\ln(odds_+) = -\ln(odds_-) = \vec{w}^T \cdot x$$

Линейные модели в задаче классификации

Шансы

$$odds = \frac{p}{1 - p} \in [0; \infty] \quad \ln(odds) \in R$$
$$\vec{w}^T \cdot x \in R$$

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

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

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

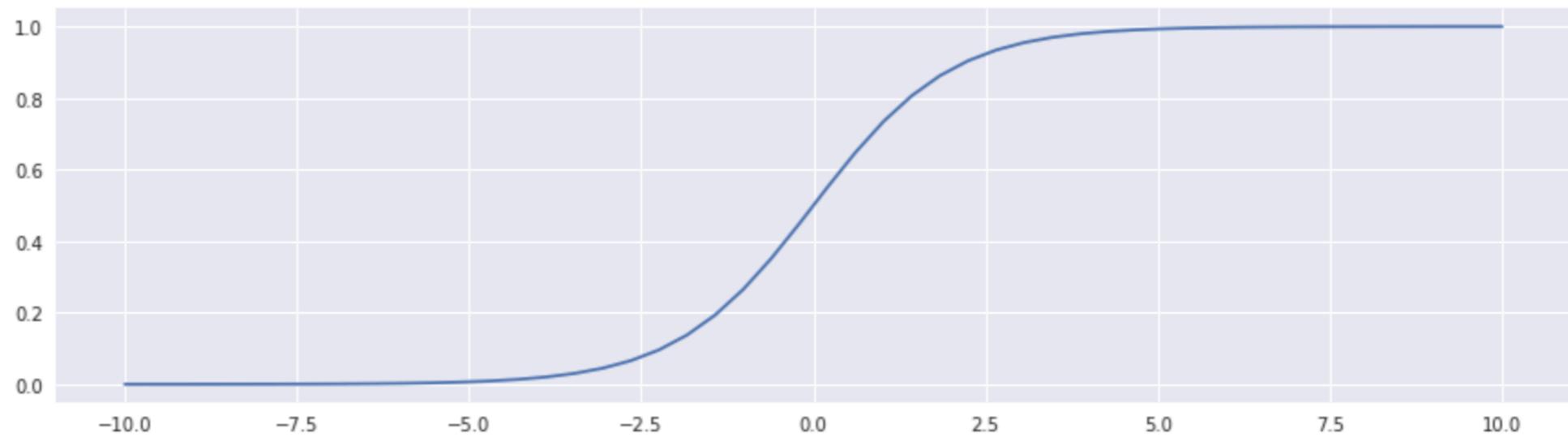
$$\ln(odds_+) = -\ln(odds_-) = w^T \cdot x$$

$$odds = e^{w^T x} \Rightarrow p = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

Линейные модели в задаче классификации

$$p = \frac{e^{w^T x}}{1 + e^{w^T x}} = \frac{1}{1 - e^{-w^T x}}$$

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



Линейные модели в задаче классификации

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

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



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

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

Делаем предположение: объекты приходят независимо, из одного распределения

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

Линейные модели в задаче классификации

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

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



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

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

Делаем предположение: объекты приходят независимо, из одного распределения

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

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

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

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

Линейные модели в задаче классификации

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

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



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

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

Делаем предположение: объекты приходят независимо, из одного распределения

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

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

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

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

Линейные модели в задаче классификации

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

$$P(\vec{y} | X, w) = \prod_{i=1}^l P(y = y_i | x_i, w) = \prod_{i=1}^l a_i^{y_i} (1 - a_i)^{(1-y_i)} \quad \mathbb{Y} = \{0; 1\}$$

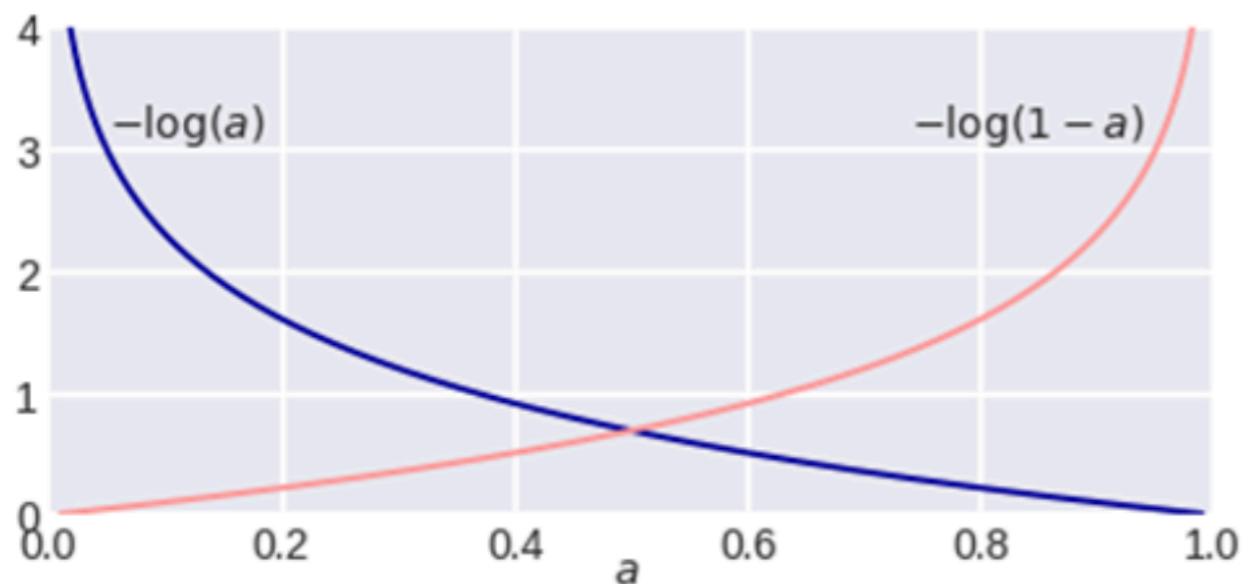
$$P(\vec{y} | X, w) = \sum_{i=1}^l \log a_i^{y_i} (1 - a_i)^{(1-y_i)} = \sum_{i=1}^l y_i \log a_i + (1 - y_i) \log(1 - a_i) \rightarrow \max$$

$$\mathcal{Q}(a, X) = \frac{1}{l} \sum_{i=1}^l -y_i \log a_i - (1 - y_i) \log(1 - a_i) \rightarrow \min$$

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

Линейные модели в задаче классификации

$$-\begin{cases} \log a_i, & y_i = 1, \\ \log(1 - a_i), & y_i = 0. \end{cases}$$



если для объекта 1го класса мы предсказываем нулевую вероятность принадлежности к этому классу или, наоборот, для объекта 0го – единичную вероятность принадлежности к классу 1, то ошибка равна бесконечности! Таким образом, **грубая ошибка на одном объекте сразу делает алгоритм бесполезным.**

Метрики качества в задачах классификации

Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	90	20
	No	10	50

Выборка: Всего 170

Положительного класса 100

Отрицательного класса 70

Прогноз:

Положительного класса 110

Отрицательного класса 60

Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Доля правильных ответов (*accuracy*):

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

Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Доля правильных ответов (*accuracy*):

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

Сколько в нашем примере?
О чём говорит эта цифра?

		Actual class	
		Yes	No
Predicted class	Yes	90	20
	No	10	50

Метрики качества в задачах классификации

Еще один пример:

		Actual class	
		Yes	No
Predicted class	Yes	90	5
	No	10	5

Доля правильных ответов (accuracy):

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Метрики качества в задачах классификации

Еще один пример:

		Actual class	
		Yes	No
Predicted class	Yes	90	5
	No	10	5

Доля правильных ответов (accuracy):

$$\text{accuracy} = \frac{90 + 5}{90 + 5 + 10 + 5} = 86,4$$

Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	90	5
	No	10	5

Доля правильных ответов (*accuracy*):

$$\text{accuracy} = \frac{90 + 5}{90 + 5 + 10 + 5} = 86,4$$

Давайте всегда будем предсказывать константным значением (110 объектов и все положительные). Посчитайте accuracy

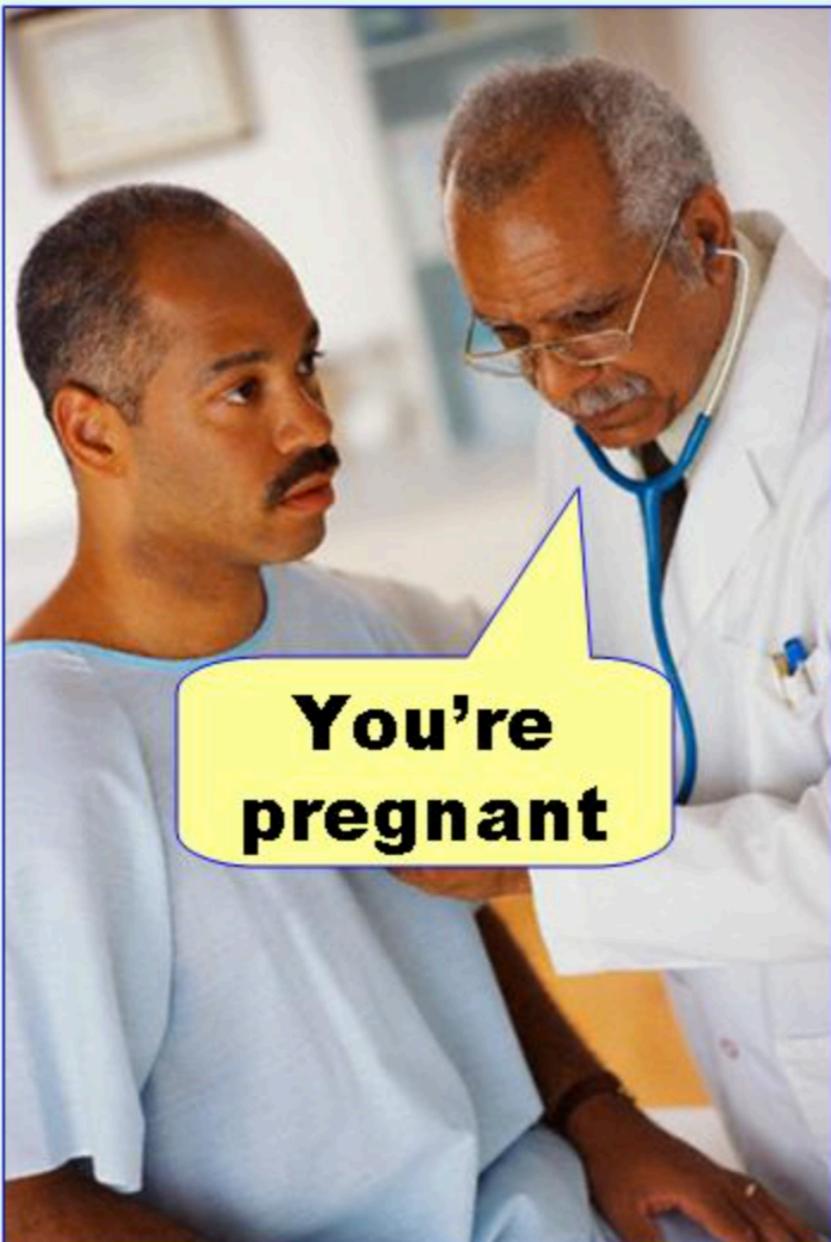
Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class		
		Yes	No	
Predicted class	Yes	True Positive (TP)	False Positive (FP)	Ошибка I-ого рода
	No	False Negative (FN)	True Negative (TN)	Ошибка II-ого рода

Метрики качества в задачах классификации

Type I error
(false positive)



Type II error
(false negative)



Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Точность (*precision*):

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

доля объектов, предсказанных как положительные, действительно является положительными.

Полнота (*recall*):

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

Доля положительных объектов, которую выделил классификатор

Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Точность (*precision*):

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

Полнота (*recall*):

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

Какая ошибку важнее оптимизировать?

Например:

1. Решаем задачу, уйдет ли от нас клиент. Какая цена ошибки?
2. Выявление фрода. Заблокировать хорошего клиента или пропустить злоумышленника?
3. Кредитный скоринг. Выдать кредит злостному неплательщику или не выдать положительному?

Метрики качества в задачах классификации

Матрица ошибок (*confusion matrix*):

		Actual class	
		Yes	No
Predicted class	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

F-мера:

$$F_\beta = (1 + \beta^2) \frac{precision \times recall}{\beta^2 precision + recall}$$

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

β в данном случае определяет вес точности в метрике, а при $\beta=1$ это среднее гармоническое

F-мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю.

Метрики качества в задачах классификации

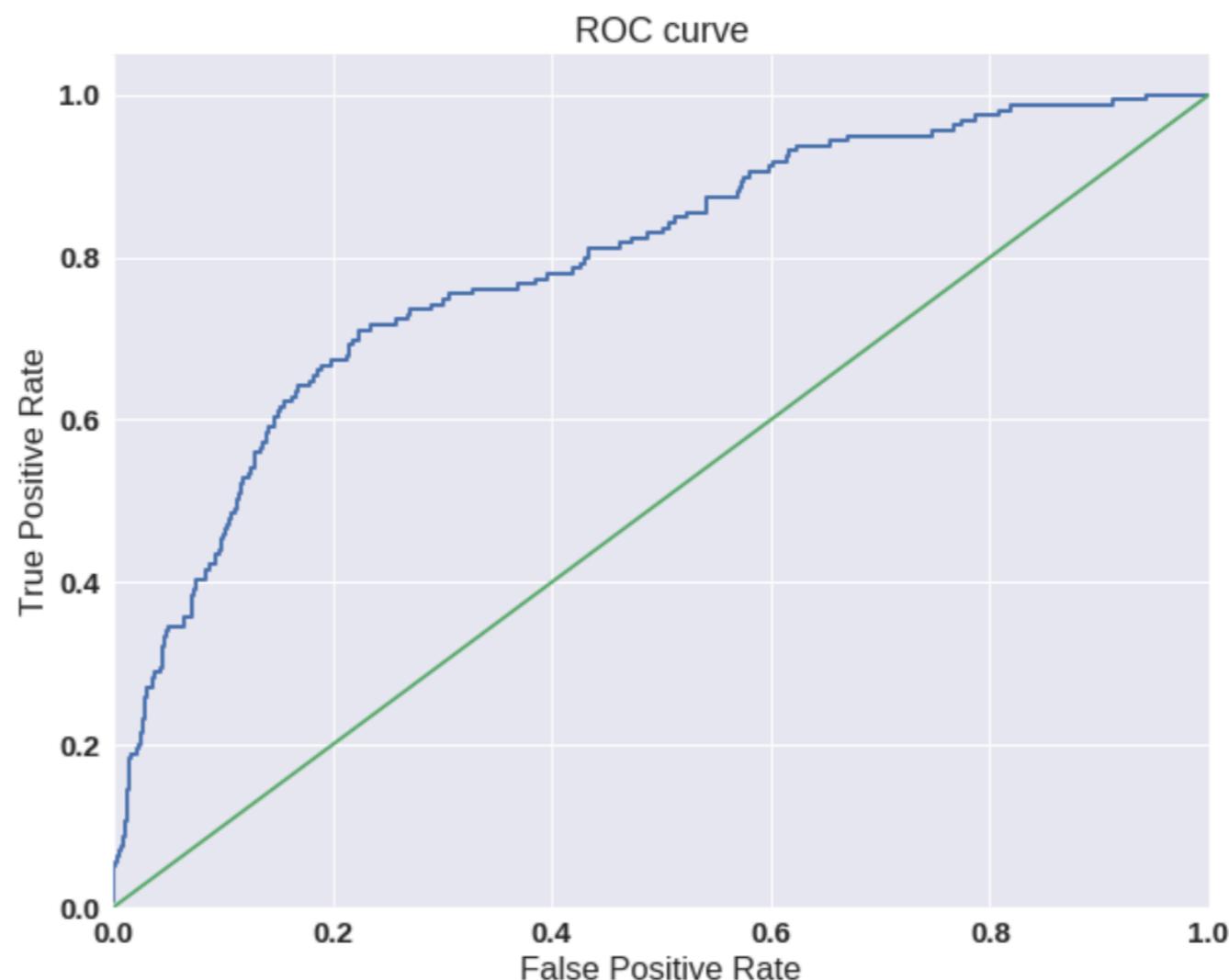
ROC AUC

или площадь (*Area Under Curve*) под кривой ошибок (*Receiver Operating Characteristic curve*).

Кривая ошибок (*Receiver Operating Characteristic curve*) представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR):

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

$$FPR = \frac{FP}{FP + TN}$$



Обобщение для многомерного случая

Sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}} \in [0; 1]$$

Softmax

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \in [0; 1] \quad \sum_{k=1}^K \sigma(z)_k = 1$$

$$\mathcal{L} = \frac{1}{l} \sum_{i=1}^l -y_i \log a_i - (1 - y_i) \log(1 - a_i)$$

$$\mathcal{L} = \frac{1}{l} \sum_{i=1}^l \sum_{k=1}^K y_{ik} \log a_{ik}$$