Лекция 4

Метод опорных векторов и логистическая регрессия

4.1. Метод опорных векторов

4.1. Метод опорных векторов

Дано: обучающая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$,

x_i - объекты, векторы выборки $X = \mathbb{R}^n$,

y_i - метки классов, элементы множества $Y=\{-1, +1\}$

Найти: параметры $w \in \mathbb{R}^n$, $w_0 \in \mathbb{R}$ линейной модели классификации

$$a(x; w, w_0) = \operatorname{sign}\big(\langle x, w\rangle - w_0\big)$$

Критерий: минимизация эмпирического риска

$$\sum_{i=1}^{\ell} \big[a(x_i; w, w_0) \neq y_i\big] = \sum_{i=1}^{\ell} \big[M_i(w, w_0) < 0\big] \to \min_{w, w_0}$$

где $M_i(w, w_0) = \big(\langle x_i, w\rangle - w_0\big) y_i$ - отступ (margin) объекта x_i

$b(x) = \langle x, w\rangle - w_0$ - дискриминантная функция

4.1. Метод опорных векторов

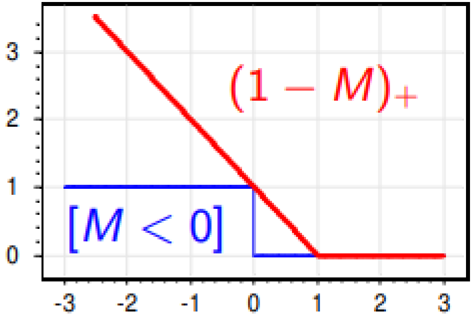
Эмпирический риск - кусочно-постоянная функция.

Заменим его оценкой сверху, непрерывной по параметрам:

$$Q(w, w_0) = \sum_{i=1}^{\ell} [M_i(w, w_0) < 0] \leqslant$$

$$\leqslant \sum_{i=1}^{\ell} \left(1 - M_i(w, w_0)\right)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w, w_0}$$

- Аппроксимация штрафует объекты классов за приближение к границе классов, увеличивая зазор между классами
- Регуляризация штрафует неустойчивые решения в случае мультиколлинеарности



4.1. Метод опорных векторов

Линейный классификатор:

$$a(x, w) = \mathrm{sign}\big(\langle w, x\rangle - w_0\big), \quad w, x \in \mathbb{R}^n,\ w_0 \in \mathbb{R}$$

Пусть выборка $X^\ell = (x_i, y_i)_{i=1}^{\ell}$ линейно разделима

$$\exists w, w_0 : \quad M_i(w, w_0) = y_i\big(\langle w, x_i\rangle - w_0\big) > 0, \quad i = 1, \ldots, \ell$$

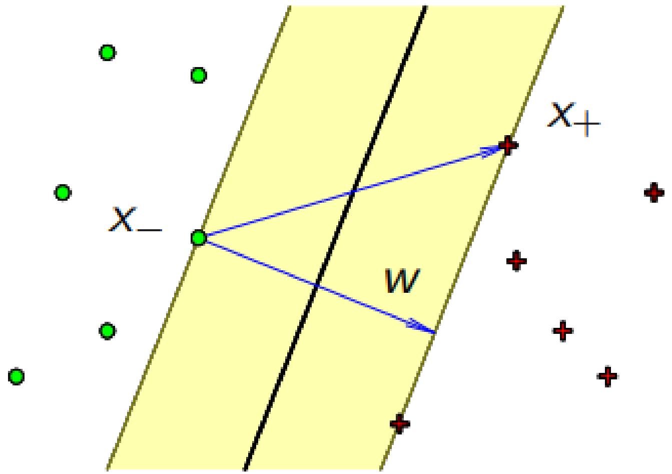
Нормировка: $\min\limits_{i=1,\ldots,\ell} M_i(w, w_0) = 1$

Разделяющая полоса:

$$\{x:\ -1 \leqslant \langle w, x\rangle - w_0 \leqslant 1\}$$

Ширина полосы:

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{2}{\|w\|} \to \max$$



4.1. Метод опорных векторов

Постановка задачи в случае линейно разделимой выборки

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,w_0}; \\ M_i(w,w_0) \geqslant 1, \quad i=1,\dots,\ell. \end{cases}$$

Общий случай - линейно неразделимая выборка:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi}; \\ M_i(w,w_0) \geqslant 1-\xi_i, \quad i=1,\dots,\ell; \\ \xi_i \geqslant 0, \quad i=1,\dots,\ell. \end{cases}$$

Исключая ξ_i, получаем задачу безусловной минимизации

$$C\sum_{i=1}^{\ell}\left(1-M_i(w,w_0)\right)_+ + \frac{1}{2}\|w\|^2 \to \min_{w,w_0}$$

4.1. Метод опорных векторов

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_{x}; \\ g_i(x) \leqslant 0, \quad i = 1, \dots, m; \\ h_j(x) = 0, \quad j = 1, \dots, k. \end{cases}$$

Необходимые условия, если x - точка локального минимума, то существуют множители

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \; h_j(x) = 0; \; \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \; \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \; \text{(условие дополняющей нежёсткости)} \end{cases}$$

4.1. Метод опорных векторов

Функция лагранжа:

$$\mathscr{L}(w, w_0, \xi; \lambda, \eta) =$$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i \bigl(M_i(w, w_0) - 1\bigr) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

λ_i - переменные, двойственные к ограничениям $M_i \geqslant 1 - \xi_i$;

η_i - переменные, двойственные к ограничениям $\xi_i \geqslant 0$.

$$\begin{cases} \frac{\partial \mathscr{L}}{\partial w} = 0, \quad \frac{\partial \mathscr{L}}{\partial w_0} = 0, \quad \frac{\partial \mathscr{L}}{\partial \xi} = 0; \\ \xi_i \geqslant 0, \quad \lambda_i \geqslant 0, \quad \eta_i \geqslant 0, \quad i = 1, \dots, \ell; \\ \lambda_i = 0 \text{ либо } M_i(w, w_0) = 1 - \xi_i, \quad i = 1, \dots, \ell; \\ \eta_i = 0 \text{ либо } \xi_i = 0, \quad i = 1, \dots, \ell; \end{cases}$$

4.1. Метод опорных векторов

$$\begin{cases} -\sum_{i=1}^{\ell} \lambda_i + \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \to \min_{\lambda}; \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \dots, \ell; \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases}$$

Решив эту задачу численно относительно λ_i получаем линейный классификатор

$$a(x) = \operatorname{sign}\Bigl(\sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle - w_0 \Bigr),$$

где $w_0 = \sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x_j \rangle - y_j$ для такого j, что $\lambda_j > 0$, $M_j = 1$

Объект x называется опорным, если $\lambda_i \neq 0$

4.2. Обобщение для нелинейного случая

$$\begin{cases} -\sum\limits_{i=1}^{\ell} \lambda_i + \frac{1}{2} \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \to \min\limits_{\lambda}; \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \dots, \ell; \\ \sum\limits_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases}$$

Решив эту задачу численно относительно λ_i получаем линейный классификатор

$$a(x) = \mathrm{sign}\Big(\sum_{i=1}^{\ell} \lambda_i y_i K(x_i, x) - w_0 \Big),$$

где $w_0 = \sum\limits_{i=1}^{\ell} \lambda_i y_i K(x_i, x_j) - y_j$ для такого j, что $\lambda_j > 0$, $M_j = 1$

Объект x называется опорным, если $\lambda_i \neq 0$

4.2. Обобщение для нелинейного случая

Ф-ия от пары объектов $K(x, x')$ называется ядром, если она представима в виде скалярного произведения

$$K(x, x') = \langle \psi(x), \psi(x') \rangle$$

при некотором преобразовании $\psi : X \to H$ из пространства признаков X в спрямляющее пространство H

Возможная интерпретация: признак $f_i(x) = K(x_i, x)$ - это оценка близости объекта x к объекту x_i. Выбирая опорные объекты, SVM осуществляет отбор признаков в линейном классификаторе

$$a(x) = \mathrm{sign}\left(\sum_{i=1}^{\ell} \lambda_i y_i K(x_i, x) - w_0 \right)$$

4.2. Обобщение для нелинейного случая

Ядра в SVM расширяют линейную модель классификации:

- $K(x, x') = (\langle x, x' \rangle + 1)^d$ - полиномиальная разделяющая поверхность степени ≤ *d*;
- $K(x, x') = \sigma(\langle x, x' \rangle)$ - нейронная сеть с заданной ф-ией активации *σ(z)* (*K* не при всех *σ* является ядром);
- $K(x, x') = \mathrm{th}(k_1 \langle x, x' \rangle - k_0)$, $k_0, k_1 \geqslant 0$ - нейросеть с сигмоидными ф-ми активации;
- $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ - сеть радиальных базисных ф-ий (RBF ядро)

4.2. Обобщение для нелинейного случая

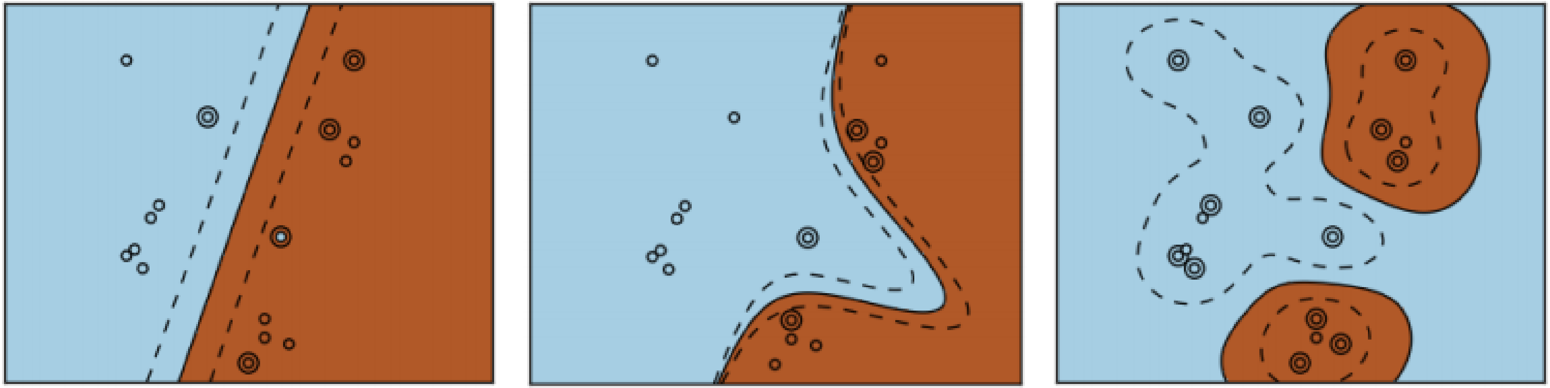
Гиперплоскость в спрямляющем пространстве соответствует нелинейной разделяющей поверхности в исходном.

Примеры с различными ядрами *K(x, x’)*

линейное $\langle x, x' \rangle$

полиномиальное $(\langle x, x' \rangle + 1)^d, \quad d{=}3$

гауссовское (RBF) $\exp(-\gamma \|x - x'\|^2)$



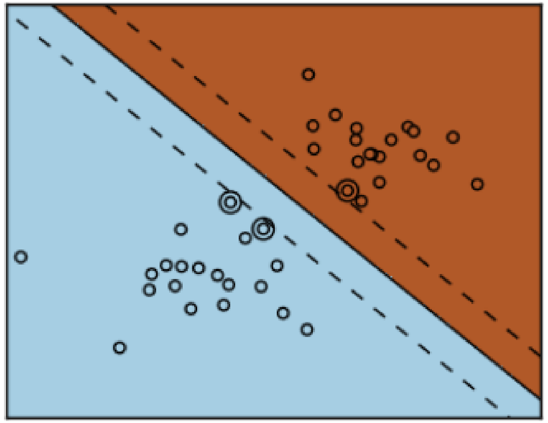
4.2. Обобщение для нелинейного случая

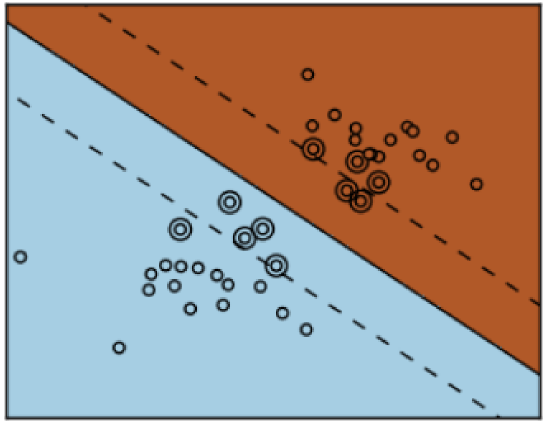
SVM - аппроксимация и регуляризация эмпирического риска:

$$\sum_{i=1}^{\ell}\left(1 - M_i(w, w_0)\right)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w, w_0}.$$

большой C
слабая регуляризация

малый C
сильная регуляризация





4.2. Обобщение для нелинейного случая

Преимущества:

- Задача выпуклого квадратичного программирования имеет единственное решение.
- Выделяется множество опорных векторов.
- Имеются эффективные численные методы для SVM
- Изящное обобщение н анелинейные классификаторы

Недостатки:

- Опорными векторами могут становиться выбросы
- Нет отбора признаков в исходном пространстве X
- Приходится подбирать константу C

4.3. Многомерная линейная регрессия

4.3. Многомерная линейная регрессия

$f_1(x)$ …. $f_n(x)$ - числовые признаки

Модель многомерной линейной регрессии:

$$f(x,\alpha) = \sum_{j=1}^{n} \alpha_j f_j(x), \qquad \alpha \in \mathbb{R}^n.$$

Матричные обозначения:

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell\times 1}{y} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \underset{n\times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}$$

Функционал квадрата ошибки:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} \big(f(x_i,\alpha) - y_i\big)^2 = \|F\alpha - y\|^2 \to \min_{\alpha}$$

4.3. Многомерная линейная регрессия

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^{\mathsf{T}}(F\alpha - y) = 0,$$

откуда следует нормальная система задачи МНК:

$$F^{\mathsf{T}}F\alpha = F^{\mathsf{T}}y,$$

где $F^{\mathsf{T}}F$ — $n \times n$- матрица.

Решение системы: $\alpha^* = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}y = F^{+}y$

Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2,$

где $P_F = FF^{+} = F(F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}$ - проекционная матрица.

4.4. Логистическая регрессия

4.4. Логистическая регрессия

Обучающая выборка: $X^\ell = (x_i, y_i)_{i=1}^\ell, \;\; x_i \in \mathbb{R}^n, \;\; y_i \in \{-1, +1\}$

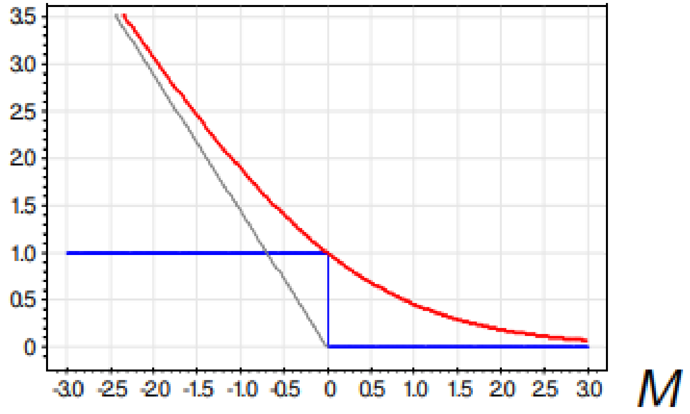
- Линейная модель классификации: $a(x, w) = \operatorname{sign}\langle x, w\rangle$
- Непрерывная аппроксимация линейной ф-ии потерь

$$Q(w) = \sum_{i=1}^{\ell} [a(x_i, w) y_i < 0] \leqslant \sum_{i=1}^{\ell} \mathscr{L}(\langle x_i, w\rangle y_i) \to \min_w$$

Отступ (margin) объекта x_i: $M_i(w) = \langle x_i, w\rangle y_i$

Логарифмическая ф-ия потерь,
как ф-ия отступа M:

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



4.4. Логистическая регрессия

$(x_i, y_i)_{i=1}^{\ell} \sim p(x, y; w)$ - выборка независимых наблюдений

Принцип максимума правдоподобия:

$$L(w) = \log \prod_{i=1}^{\ell} p(x_i, y_i; w) = \sum_{i=1}^{\ell} \log P(y_i|x_i; w) p(x_i) \to \max_w .$$

Вероятностная модель порождения данных с параметром *w*:

- *p(x)* - не зависит от параметра *w*;
- *p(y|x;w)* описывается линейной моделью классификации.

$$Q(w) = \sum_{i=1}^{\ell} \log\left(1 + \exp(-\langle w, x_i \rangle y_i)\right) \to \min_w .$$

$$P(y_i|x_i; w) = \frac{1}{1 + \exp\left(-\langle x_i, w \rangle y_i\right)} = \sigma\left(\langle x_i, w \rangle y_i\right),$$ - сигмоидная $\sigma(M) = \frac{1}{1+e^{-M}}$

Тогда задачи $Q(w) \to \min$ и $L(w) \to \max$ эквивалентны

4.4. Логистическая регрессия

$$p = \frac{1}{1+e^{-z}} \; ,$$

где $z = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n + a$,

$$p(X) = \frac{e^{(\beta_o+\beta_1 x)}}{e^{(\beta_o+\beta_1 x)} + 1}$$

$$p(e^{(\beta_o+\beta_1 x)} + 1) = e^{(\beta_o+\beta_1 x)}$$

$$p.e^{(\beta_o+\beta_1 x)} + p = e^{(\beta_o+\beta_1 x)}$$

$$p = e^{(\beta_o+\beta_1 x)} - p.e^{(\beta_o+\beta_1 x)}$$

$$p = e^{(\beta_o+\beta_1 x)}(1-p)$$

$$\frac{p}{1-p} = e^{(\beta_o+\beta_1 x)}$$

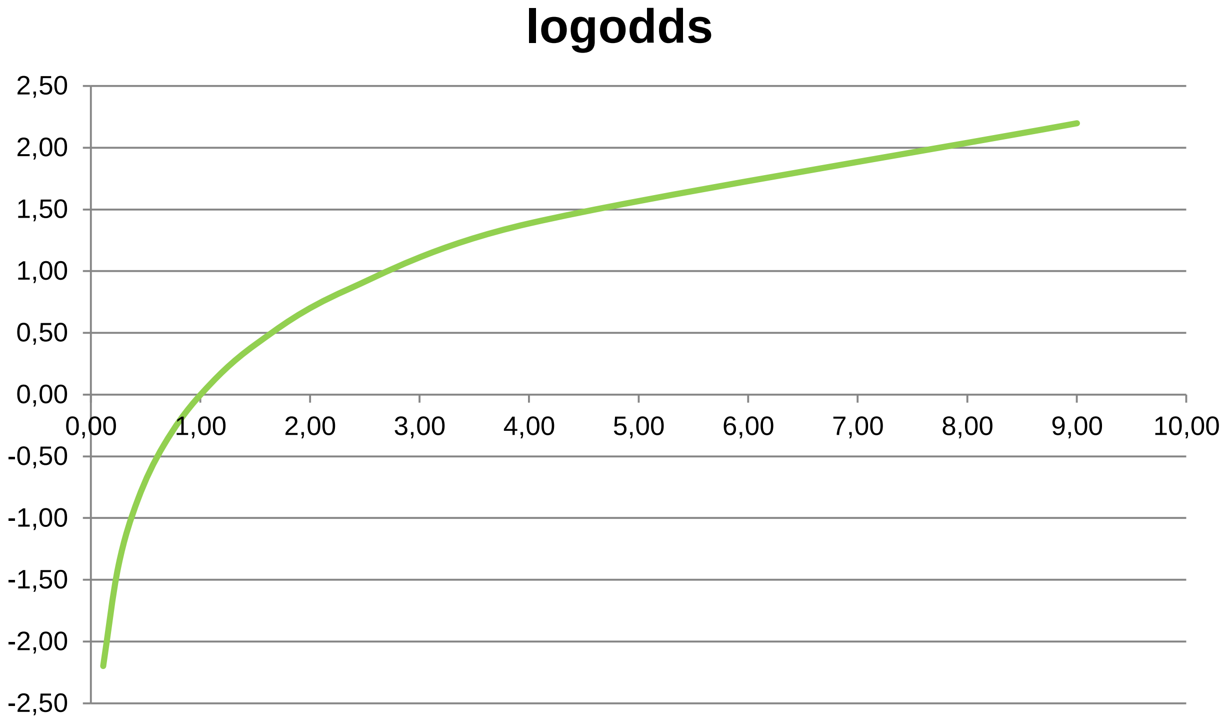
$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

4.4. Логистическая регрессия

p	odds
0,1	0,11
0,2	0,25
0,3	0,43
0,4	0,67
0,5	1,00
0,6	1,50
0,7	2,33
0,8	4,00
0,9	9,00

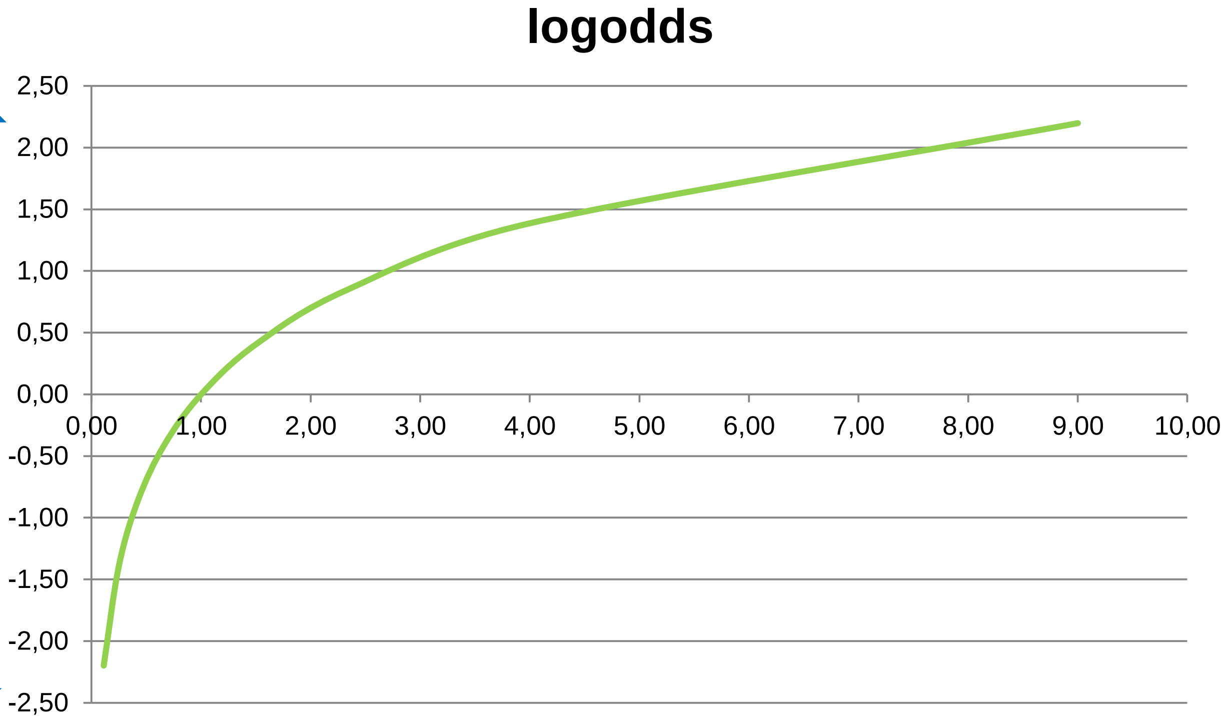
4.4. Логистическая регрессия

p	odds	logodds
0,1	0,11	-2,20
0,2	0,25	-1,39
0,3	0,43	-0,85
0,4	0,67	-0,41
0,5	1,00	0,00
0,6	1,50	0,41
0,7	2,33	0,85
0,8	4,00	1,39
0,9	9,00	2,20

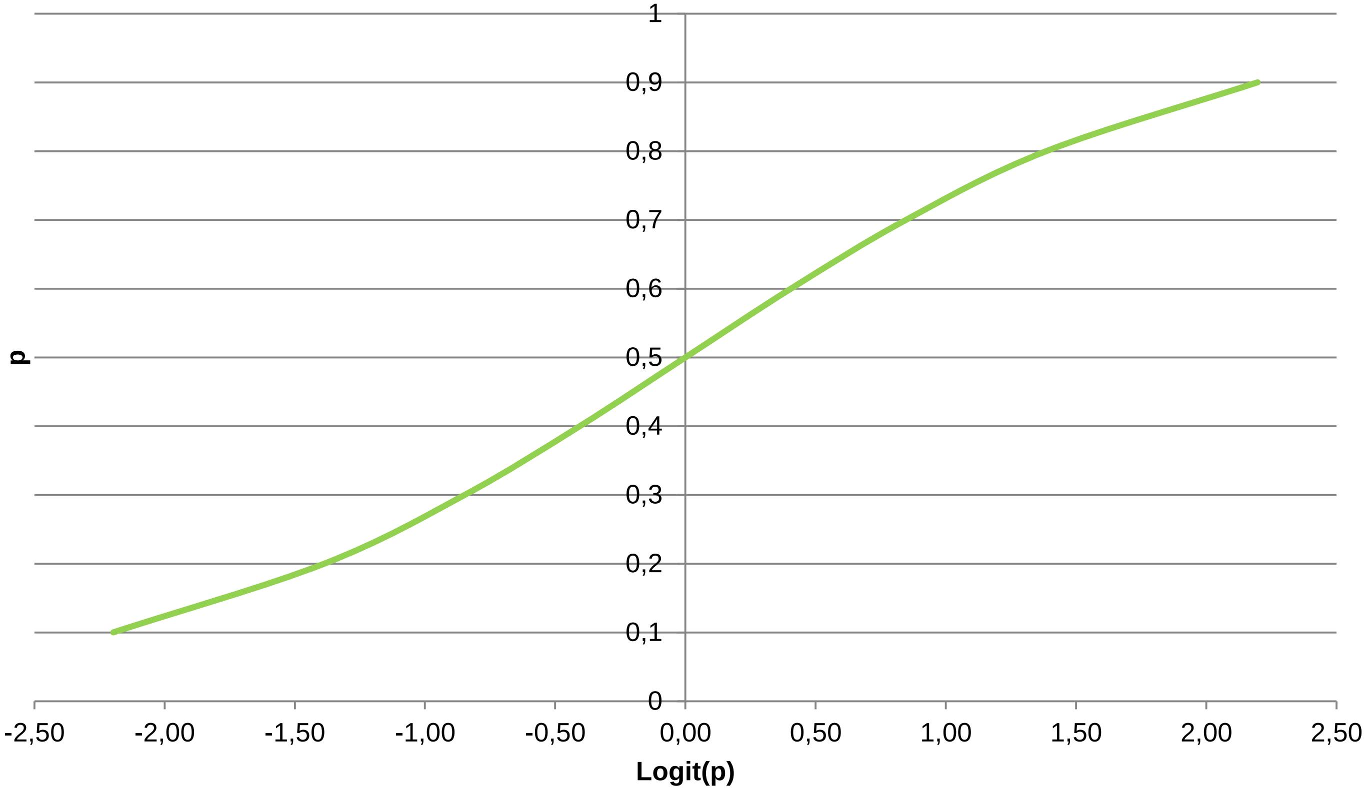


4.4. Логистическая регрессия

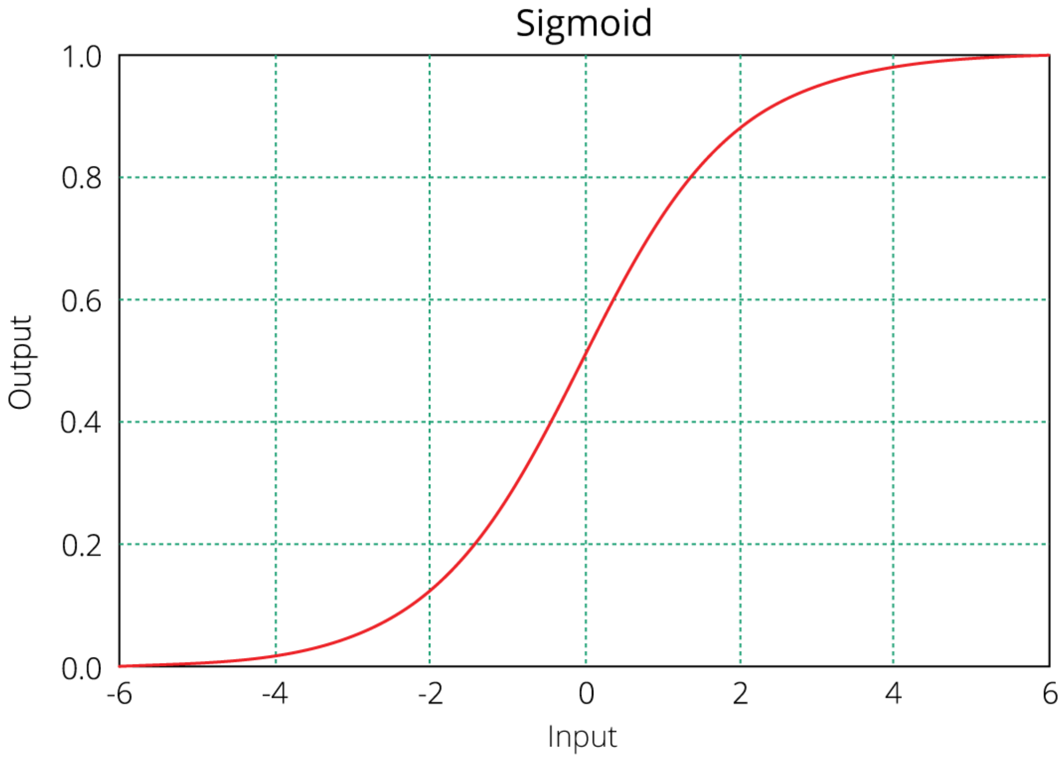
p	odds	logodds
0,1	0,11	-2,20
0,2	0,25	-1,39
0,3	0,43	-0,85
0,4	0,67	-0,41
0,5	1,00	0,00
0,6	1,50	0,41
0,7	2,33	0,85
0,8	4,00	1,39
0,9	9,00	2,20



4.4. Логистическая регрессия



4.4. Логистическая регрессия



4.4. Логистическая регрессия

- Логистическая регрессия - линейный классификатор, оценивающий апостериорные вероятности классов *P(y|x)*, необходимые в прикладных задачах оценивания рисков.
  $|y - PD|^2 \rightarrow \min_B.$
- Регуляризация улучшает обобщающую способность:
  - L2-регуляризация - при мультиколлинеарности признаков
  - L1-регуляризация - для отбора признакв
  - Elasticnet - для менее агрессивного отбора признаков

$$|y - PD|^2 + \lambda_1 |B|_1 + \lambda_2 |B|_2^2 \rightarrow \min_B$$