Машинное обучение

Лекция 5

Логистическая регрессия и SVM

Евгений Егоров
https://github.com/WeaselCMC/ml_dpo_2021

План лекции

- Бинарная классификация
- Предсказание вероятностей
- Логистическая регрессия
- Метод опорных векторов

Линейная классификация

много классов → много бинарных задач

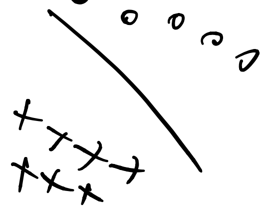
- Решаем задачу бинарной классификации:

$$\mathbb{Y} = \{-1, +1\}$$

+1 >0
-1 ≤0 „линейной"

- Линейная модель:

$$a(x) = \text{sign}(\langle w, x\rangle - t) = \text{sign}(\langle w, x\rangle)$$

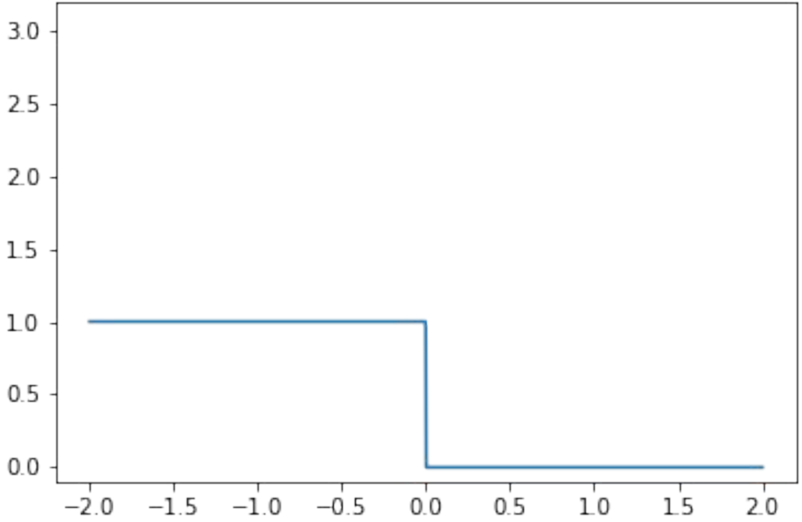


- Функция потерь:

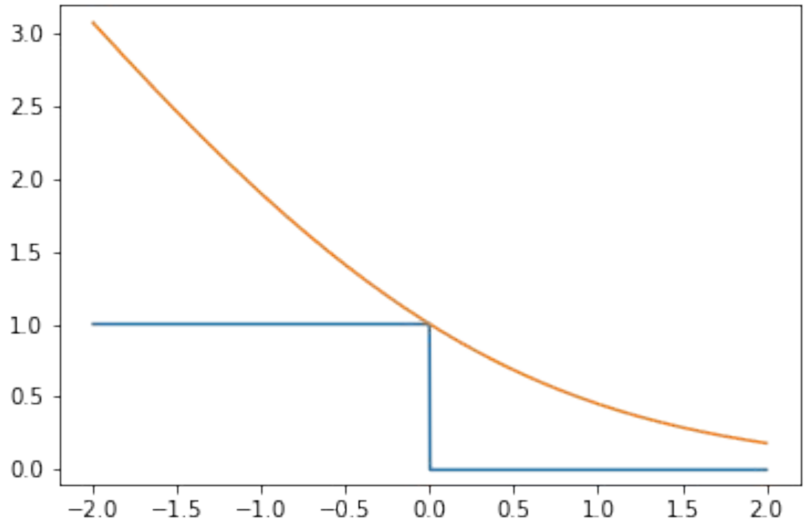
$$Q(w, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[\underbrace{y_i\langle w, x_i\rangle}_{M_i} < 0]$$

M_i = „отступ"

Линейная классификация

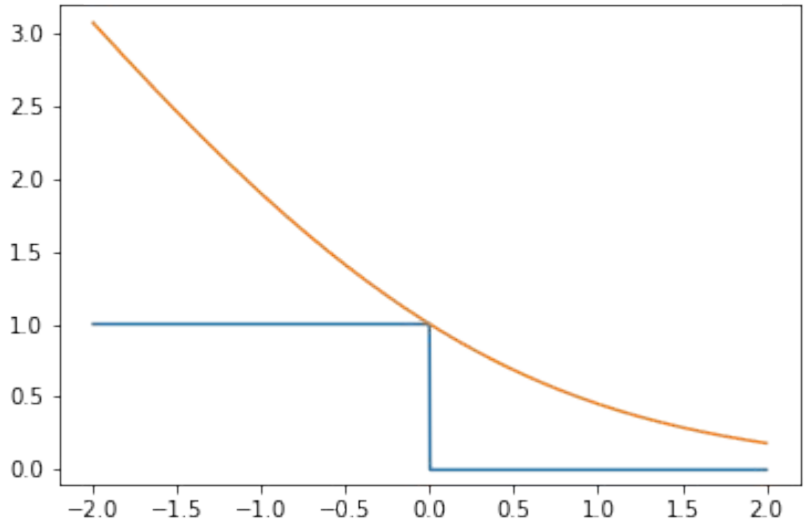


Линейная классификация



Линейная классификация

$$0 \le \frac{1}{\ell}\sum_{i=1}^{\ell}[y_i\langle w, x_i\rangle < 0] \le \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(y_i\langle w, x_i\rangle) \to \min_w$$



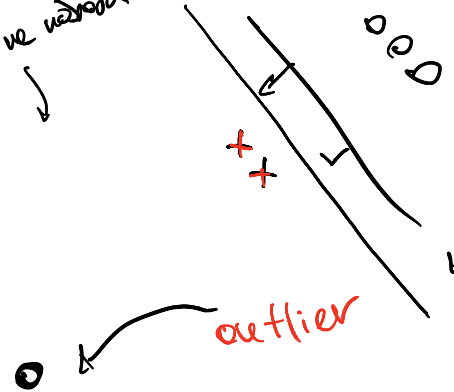
Линейная классификация

- Разные верхние оценки => разные модели

- Свойства модели => верхняя оценка

не подходят

хорош

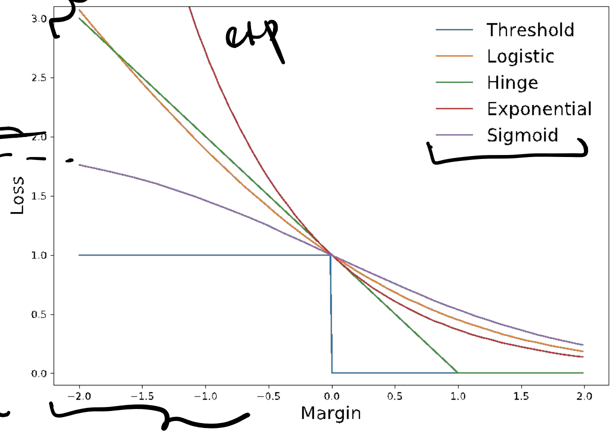


хуёвно

Быстрее всех

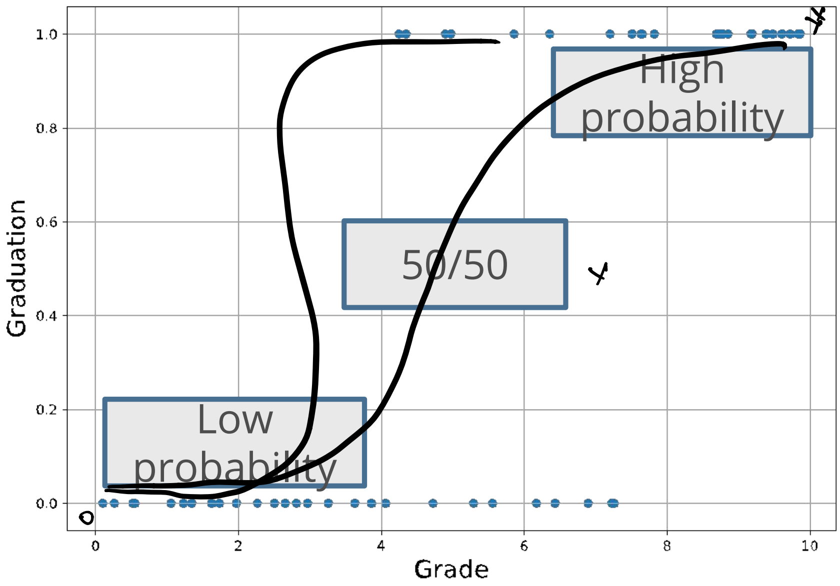
ограничен

где ошибки



$f(100\, \mathrm{sign}[<w, x>]) \in [0,1]$?

Предсказание вероятностей



Предсказание вероятностей

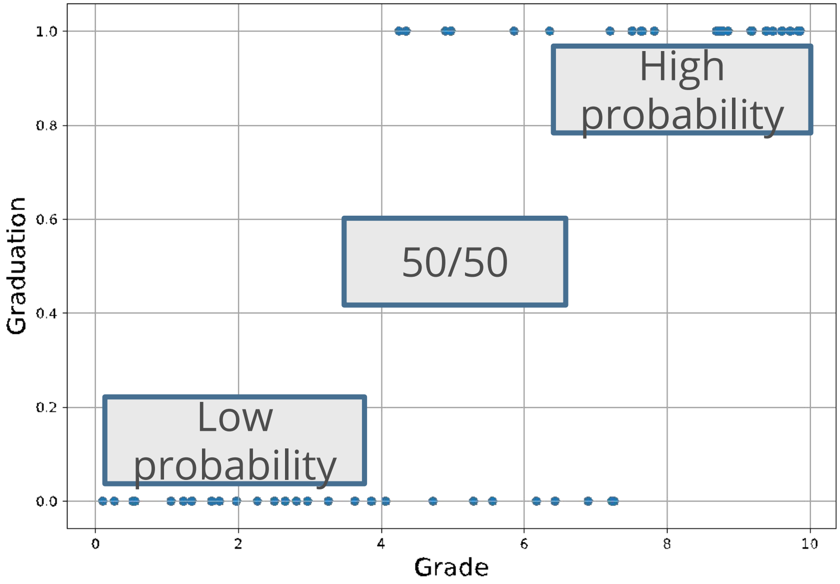
- Кредитный скоринг
- Стратегия: выдавать кредит только клиентам с вероятностью возврата > $\underbrace{0.9}$ ← не вероятность, просто „число"
- 10% невозвращенных кредитов - нормально

$\underbrace{10^6 \cdot 0.1}$

Предсказание вероятностей

- Реклама в интернете:
  - $b(x)$ — вероятность клика
  - $c(x)$ — прибыль в случае клика
  - $\underbrace{c(x)b(x)}$ — хотим оптимизировать ?

Предсказание вероятностей



$$\frac{1}{1+\exp(\theta)}$$

$\sigma(\cdot)$

$\Rightarrow$ 1) $[0,1]$ ✓

2) выходы (?) можно считать вероятностями

Логистическая регрессия

P.S

3) Всегда ли σ?

$a(x) \to \mathbb{R}$ $\Rightarrow$ калибровка модели

Pr.

0 1

Логистическая регрессия

- Решаем задачу бинарной классификации:

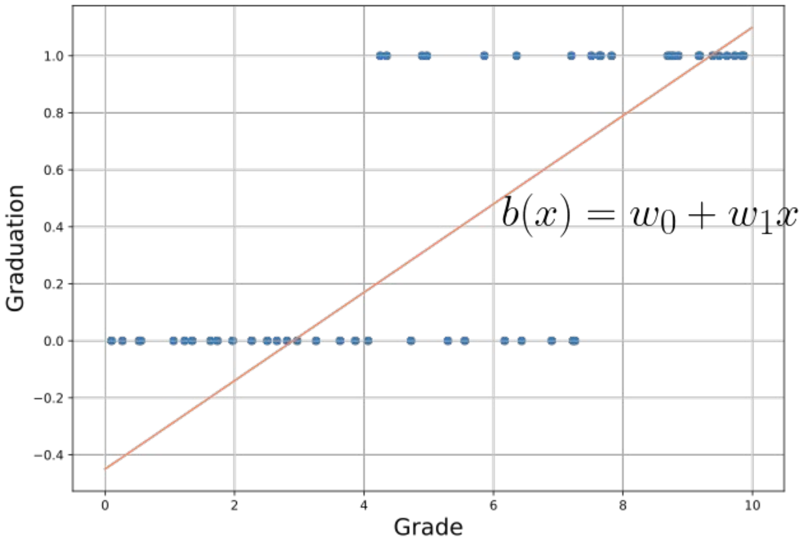
$$\mathbb{Y} = \{-1, +1\}$$

- Линейная модель:

$$a(x) = \operatorname{sign}(b(x)) = \operatorname{sign}(\langle w, x \rangle)$$

- Может использовать $b(x)=\langle w, x\rangle$ как оценку вероятности?

Логистическая регрессия



Логистическая регрессия

- Давайте переведем выход модели на отрезок [0, 1] ✓

$\in \mathbb{R}^k$

$w \to$ веса

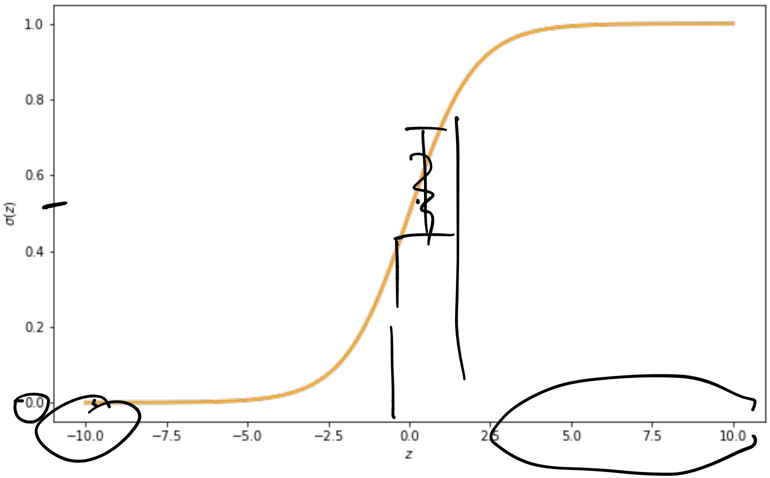
$x \to$ объект

$\in \mathbb{R}^k$

- Например, с помощью сигмоиды

$$\sigma(\langle w, x\rangle) = 1/(1+\exp(-\underbrace{\langle w, x\rangle}_{\text{линейность ЛР}}))$$

Сигмоида



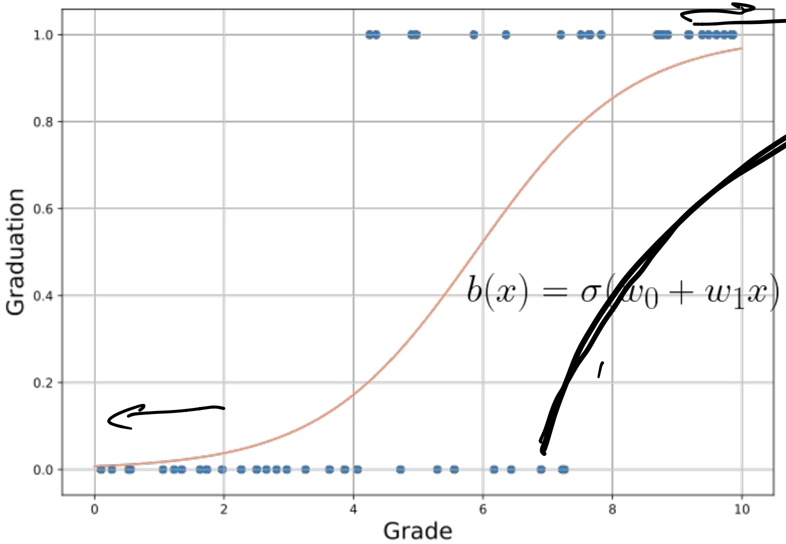
$\frac{1}{2}$

$Pr(y = 1)$

$sign[\langle w, x \rangle]$

Предсказание вероятностей

1. $\log \sigma$ вогнут
2. $+\|w\|^2$
3.



$$b(x) = \sigma(w_0 + w_1 x)$$

Логистическая регрессия

- Задача бинарной классификации: $\mathbb{Y} = \{-1, +1\}$

- Предсказание вероятностей

$$\mathbb{P}(y_i = 1) = \sigma(\langle w, x_i \rangle)$$

Теперь мы можем использовать метод максимального правдоподобия

Логистическая регрессия

$y_n \in \{-1, +1\} \vee \quad \{(x_n, y_n)\}_{n=1}^{N}$

$\{0, 1\} \vee$

$$\prod_{n=1}^{N_1} P_w(y_n = 1 \mid x_n) \prod_{n=1}^{N_2} P_w(y_n = -1 \mid x_n) \to \max_{w}$$

$\Rightarrow$ сумма

$$\max \sum_{n=1}^{N} \underbrace{\log \sigma(y_n \langle w, x_n \rangle)}_{\text{(выпуклая по } w\text{)}} \qquad \sigma = \frac{1}{1 + \exp(y_n \langle w, x_n \rangle)}$$

$$\nabla_w : \sum_{n=1}^{N} \frac{1 \cdot \cancel{\exp}(y_n \langle w, x_n \rangle)}{1 + \cancel{\exp}(y_n \langle w, x_n \rangle)} \, y_n x_n$$

$$\sum_{n=1}^{N} \left(1 - \sigma(y_n \langle w, x_n \rangle)\right) y_n x_n$$

$$(y_n - \mathbb{E}\, y \mid x_n)\, x_n$$

$(y_n - \mathbb{E}\, y(x))\, x_n$

$(y_n - 2\langle w, x_n \rangle)\, x_n$

Логистическая регрессия

$$-\sum_{i=1}^{\ell}\{[y_i = 1]\log\sigma(\langle w, x_i\rangle) + [y_i = -1]\log(1-\sigma(\langle w, x_i\rangle))\} =$$

$$-\sum_{i=1}^{\ell}\left\{[y_i = 1]\log\frac{1}{1+\exp(-\langle w, x\rangle)} + [y_i = -1]\log\left(1-\frac{1}{1+\exp(-\langle w, x\rangle)}\right)\right\} =$$

$$-\sum_{i=1}^{\ell}\left\{[y_i = 1]\log\frac{1}{1+\exp(-\langle w, x\rangle)} + [y_i = -1]\log\left(\frac{1}{1+\exp(\langle w, x\rangle)}\right)\right\} =$$

$$\sum_{i=1}^{\ell}\{[y_i = 1]\log(1+\exp(-\langle w, x\rangle)) + [y_i = -1]\log(1+\exp(\langle w, x\rangle))\} =$$

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

Логистическая регрессия

- Немало задач, где нужно предсказывать вероятности

- Можно применить сигмоиду к линейной модели, чтобы получать числа от 0 до 1

- Логистическая функция максимизирует отступы — то есть приближает вероятности к 1 или 0 в зависимости от класса

Логистическая регрессия и друзья

1. $\in \{0,1\}$ $\in \mathbb{R}$ $? \ \mathbb{N}_+$: GLM Def.

2. MLE D: $(y_n - \mathbb{E}\, y|x_n)\, x_n$?

$\sigma(\langle w, x_n \rangle)$

это модель

$$p(y_n) = \frac{\mu^{y_n}}{y_n!} \exp(-\mu) \qquad \log p(y_n) = y_n \cdot \log \mu - \mu$$

$$\theta = \log \mu, \quad \mu = \exp(\theta)$$

$$y_n \cdot \theta - \exp(\theta)$$

MLE $\log \prod p(y_n)$

$$\sum \log p(y_n|x_n) = \sum y_n \cdot \langle w, x_n \rangle - \exp(\langle w, x_n \rangle)$$

$$\sum_{n=1}^{N} \left(y_n \cdot \cancel{x_n} - \underbrace{\exp(\langle w, x_n \rangle)}_{\text{м.о.}} \right) \cdot \cancel{x_n}\ x_n$$

1. Ответ на вопрос (exp?)

2.

20??

2020 — здесь — нейронные сети

2014

2000

1995 SVM. → {необходимость создавать признаки руками.}

RAW Img

— obj

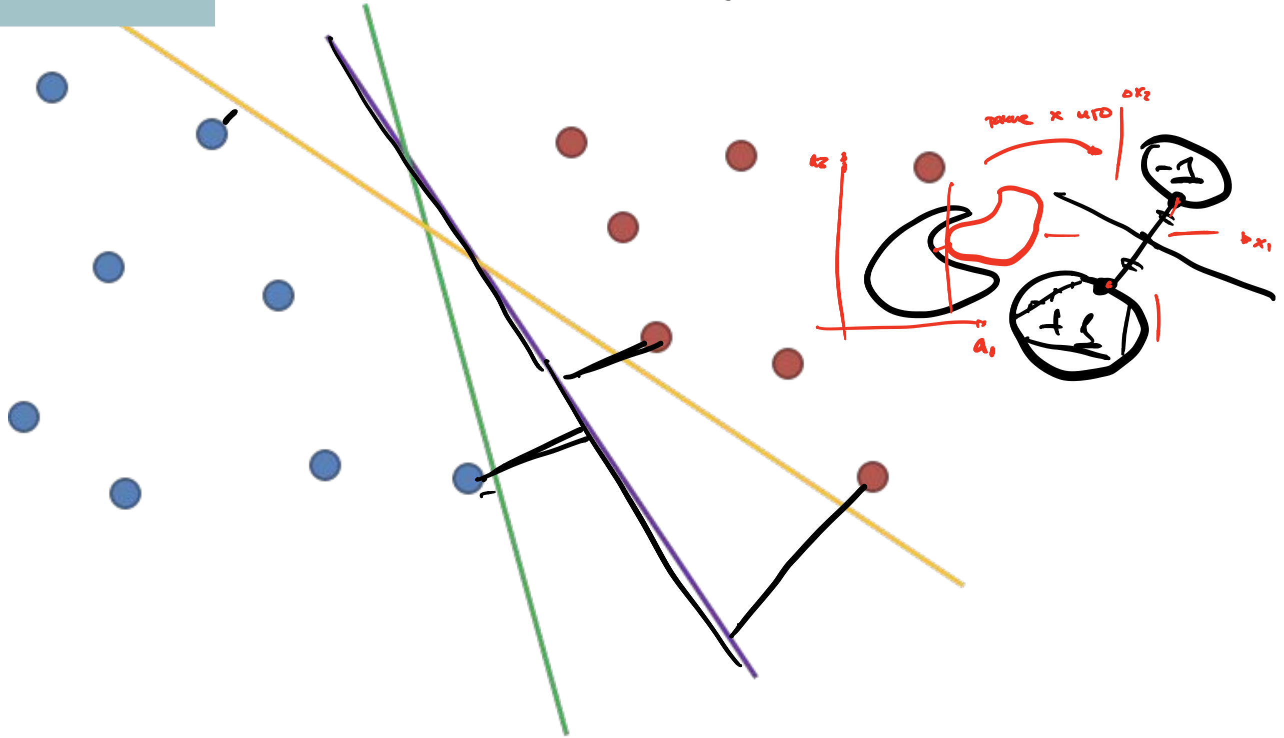
кто это сделал?

$\langle W, X_ч \rangle$

Support Vector Machine (SVM)

{Метод опорных векторов}

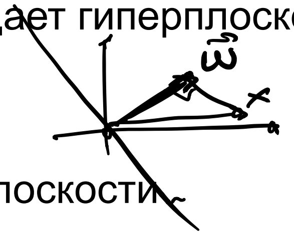
Какой классификатор лучше?



Отступ классификатора

Вспомним, что линейный классификатор задает гиперплоскость

$$\langle w, x\rangle = 0$$



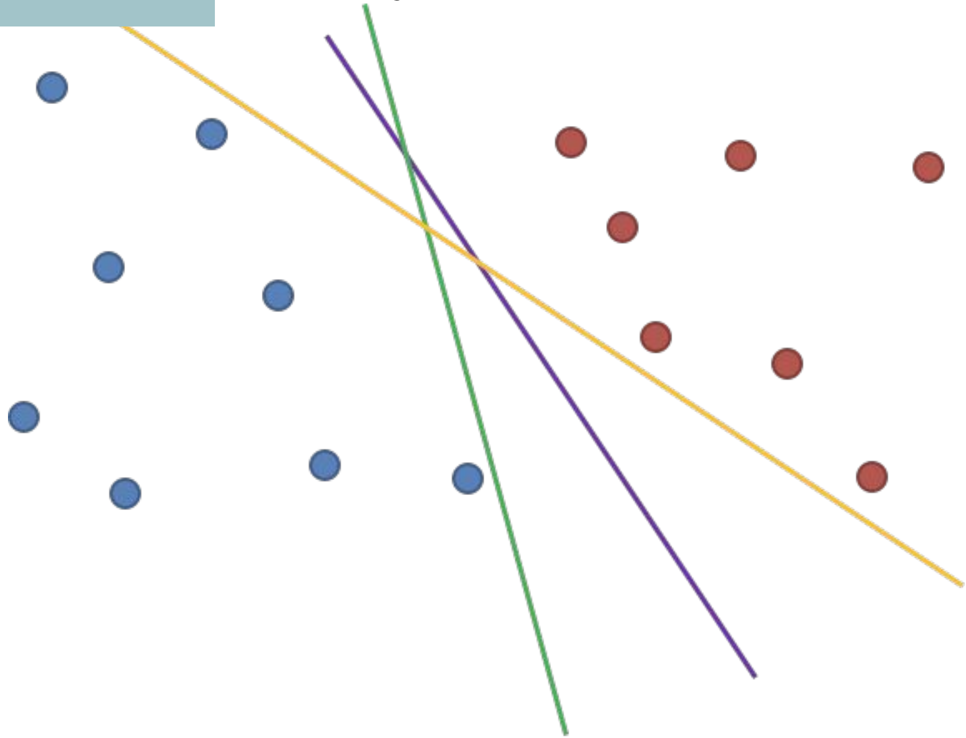
$|w| = 1.$

Расстояние от любой точки x до этой гиперплоскости

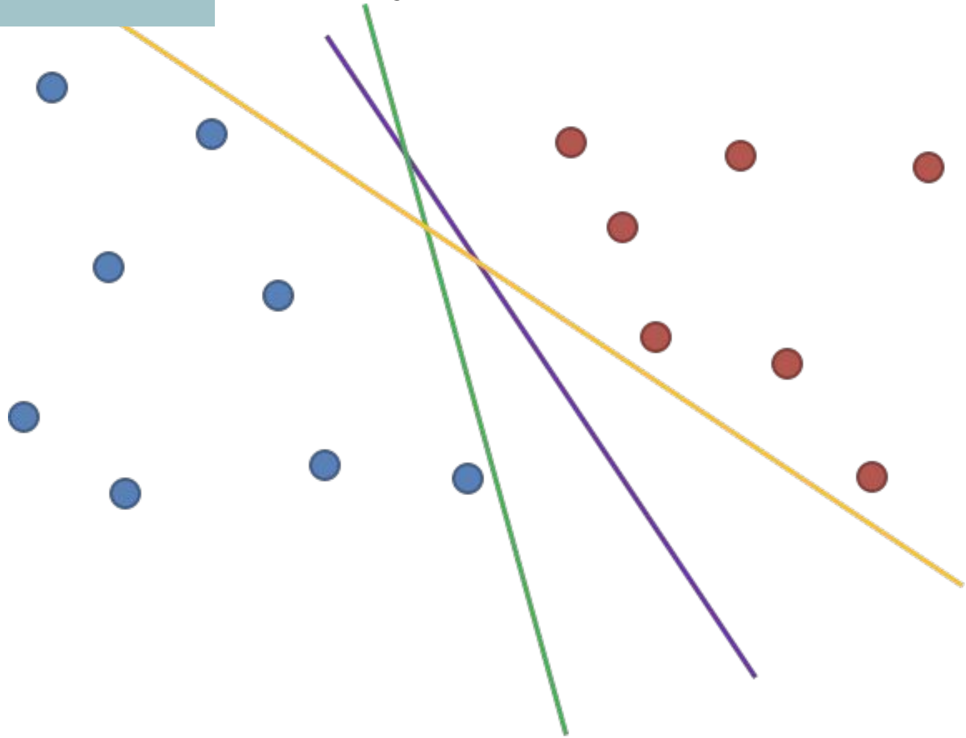
$$\frac{|\langle w, x\rangle|}{\|w\|}$$

Отступ классификатора — расстояние от гиперплоскости до ближайшего объекта

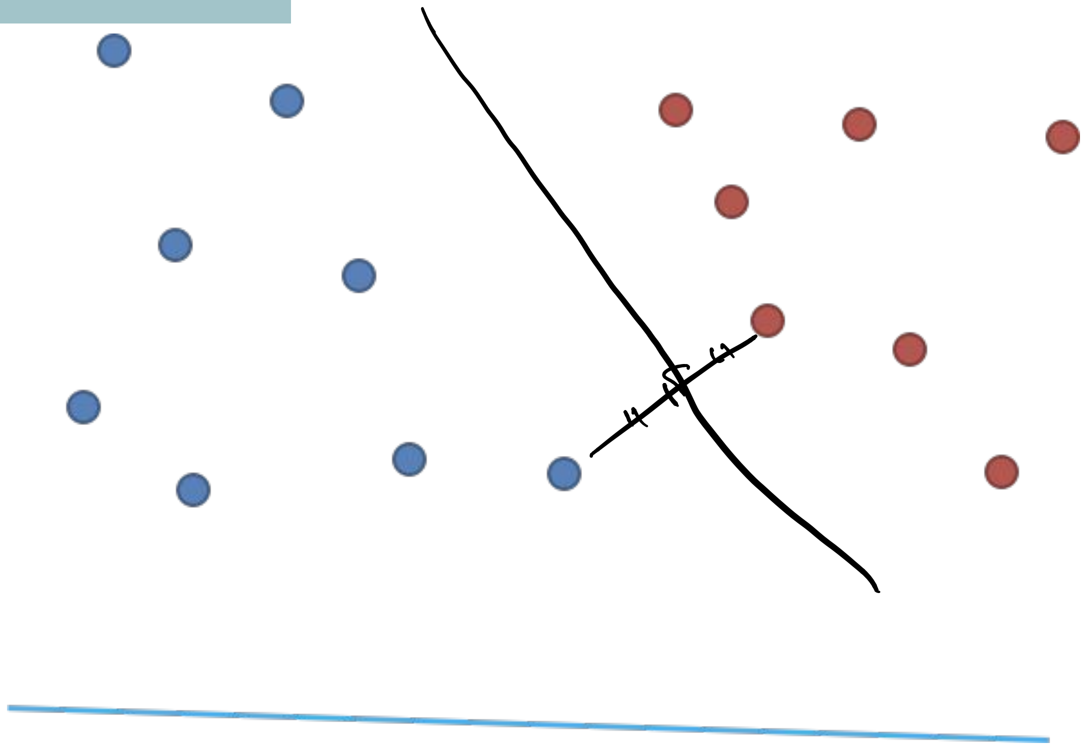
Какой классификатор лучше?



Какой классификатор ~~лучше?~~ имеет больший отступ?



Какой классификатор ~~лучше?~~ имеет больший отступ?



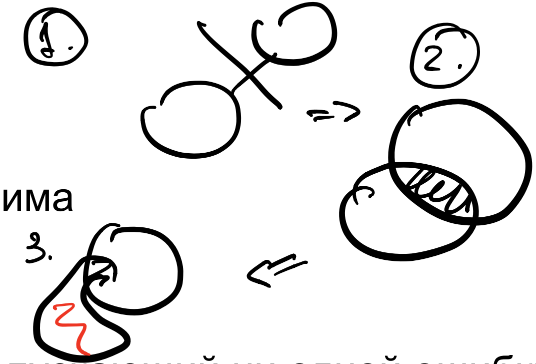
Отступ классификатора

- Будем максимизировать отступ классификатора — расстояние от гиперплоскости до ближайшего объекта

- При этом будет стараться сделать поменьше ошибок

- По сути, делаем как можно меньше предположений о модели, и верим, что это снизит вероятность переобучения

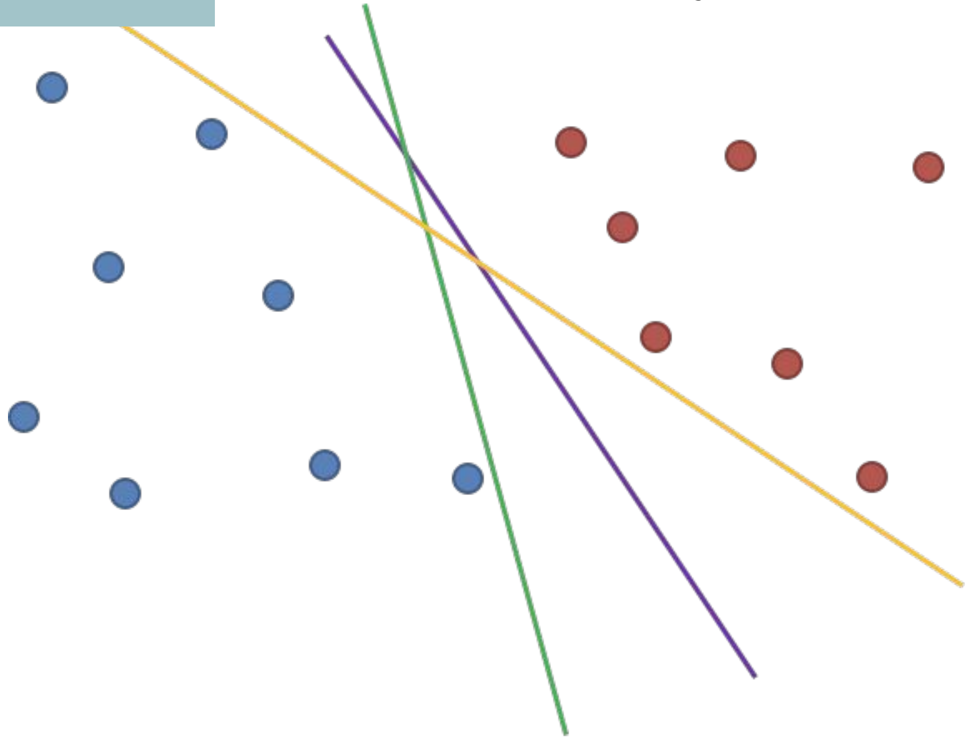
Простой случай



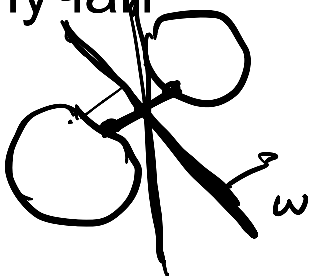
- Будем считать, что выборка линейно разделима

- Существует линейный классификатор, не допускающий ни одной ошибки

Линейно разделимый случай



Линейно разделимый случай



Требование 1: максимальный отступ

$$\min_i \frac{|\langle w, x_i\rangle|}{\|w\|}$$

Требование 2: Нет ошибок

$$y_i\langle w, x_i\rangle > 0, i = 1 \dots N$$

$\{-1, 1\}$

Небольшое предположение

Линейный классификатор:

$$a(x_i) = \mathrm{sign}(\langle w, x_i \rangle)$$

Если мы поделим w на число $a > 0$, то выходы классификатора никак не поменяются:

$$a(x_i) = \mathrm{sign}\left(\frac{\langle w, x_i \rangle}{a}\right) = \mathrm{sign}\left(\langle w, x_i \rangle\right)$$

Небольшое предположение

Если мы поделим w на число a > 0, то выходы классификатора никак не поменяются:

$$a(x_i) = \operatorname{sign}\left(\frac{\langle w, x_i\rangle}{a}\right) = \operatorname{sign}\left(\langle w, x_i\rangle\right)$$

Давайте поделим на $\min\limits_i |\langle w, x_i\rangle|$

$$\tilde{w} = \frac{w}{\min_i |\langle w, x_i\rangle|}$$

$$\min_i |\langle w, x_i\rangle| = 1$$

Небольшое предположение

Давайте поделим на $\min_i |\langle w, x_i \rangle|$

$$\min_i |\langle \tilde{w}, x_i \rangle| = 1$$

Отступ классификатора:

$$\min_i \frac{|\langle \tilde{w}, x_i \rangle|}{||\tilde{w}||} = \frac{\min_i |\langle \tilde{w}, x_i \rangle|}{||\tilde{w}||} = \frac{1}{||\tilde{w}||}$$

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_w$$ ✓

Требование 2: Нет ошибок

$$y_i \langle w, x_i \rangle > 0, i = 1 \ldots l$$ ✓

Требование 3:

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_w$$

Требование 2: Нет ошибок

$$y_i \langle w, x_i \rangle > 0, i = 1 \ldots l$$

Требование 3:

$$\min_i |\langle w, x_i \rangle| = 1$$

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: Нет ошибок

$$y_i \langle w, x_i \rangle > 0, i = 1 \ldots N$$

Требование 3:

$$|\langle w, x_i \rangle| \geq 1, \quad i = 1 \ldots N$$

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: $y_i \langle w, x_i \rangle > 0, i = 1 \dots l$

Требование 3: $|\langle w, x_i \rangle| \geq 1, \quad i = 1 \dots l$

Требование 2 + 3:

Линейно разделимый случай

Требование 1: максимальный отступ

$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2: $y_i \langle w, x_i \rangle > 0, \; i = 1 \ldots N$

+

Требование 3: $|\langle w, x_i \rangle| \geq 1, \quad i = 1 \ldots N$

Требование 2 + 3:

$$y_i \langle w, x_i \rangle \geq 1, \quad i = 1 \ldots N$$

Линейно разделимый случай

Требование 1: максимальный отступ

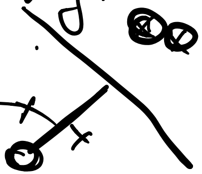
$$\frac{1}{\|w\|} \to \max_{w}$$

Требование 2 + 3: $y_i\langle w, x_i\rangle \geq 1, \quad i = 1 \dots l$

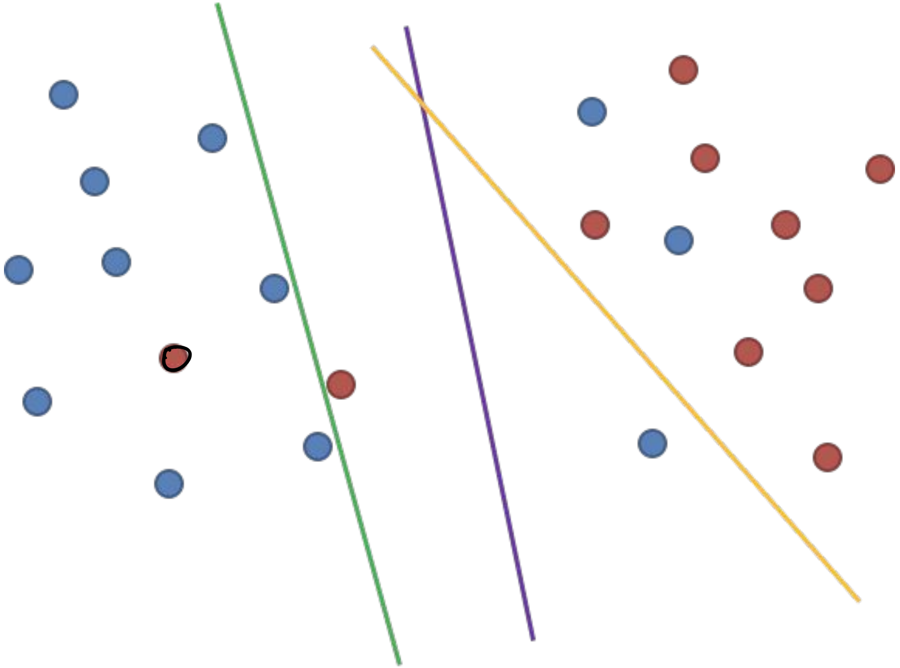
"l_2" reg.

Итого:

$$\begin{cases} \min_w \|w\| \\ \underbrace{y_i\langle w, x_i\rangle \geq 1}\, - \xi_i \end{cases}$$



Линейно неразделимый случай



Линейно неразделимый случай

- Любой линейный классификатор допускает хотя бы одну ошибку

$$\begin{cases} \min_w \|w\| \\ \boxed{y_i\langle w, x_i\rangle \geq 1} \end{cases}$$

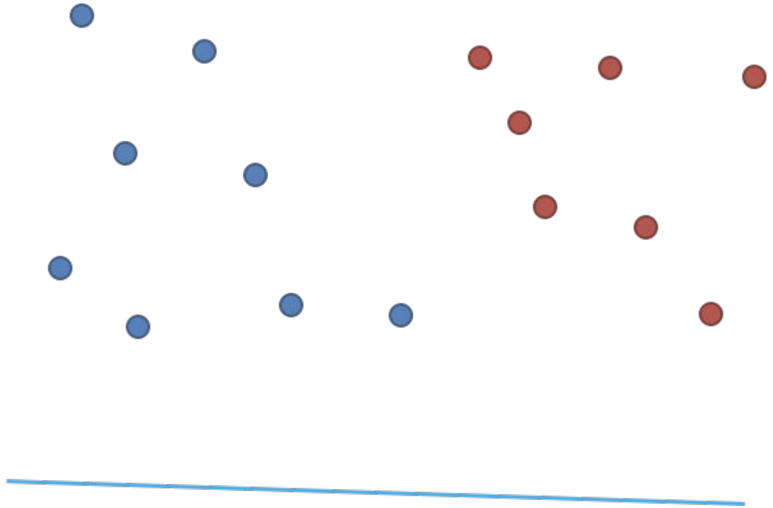
- Давайте смягчим условия

$$\begin{cases} \min_w \|w\| \\ y_i\langle w, x_i\rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

$\Rightarrow$ релаксация на долю ошибок.

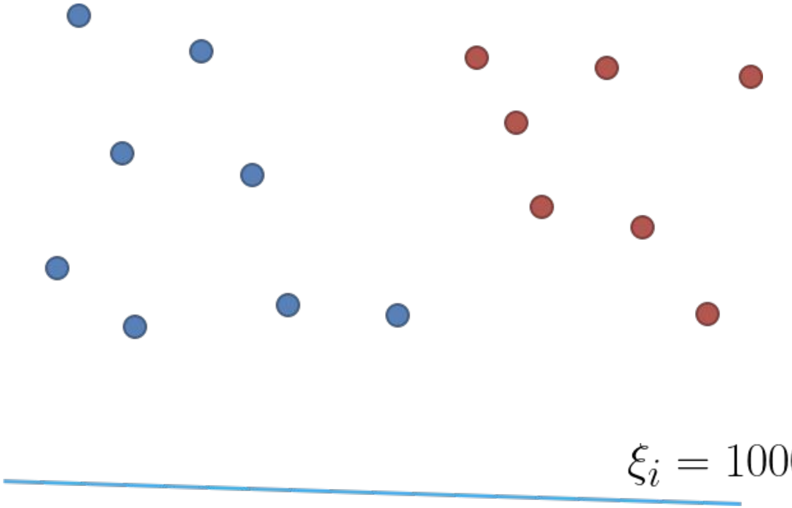
Линейно неразделимый случай

$$\begin{cases} \min_w \left[\|w\| + \sum_i \xi_i \right] \\ y_i \langle w, x_i \rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$



Линейно неразделимый случай

$$\begin{cases} \min_w \|w\| \\ y_i\langle w, x_i\rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$



$$\xi_i = 100000000$$

Линейно неразделимый случай

число объектов в выборке

$i = 1, \dots, N$

cvx opt

V.

ошибки модели

рег.

$$\min_w \|Y - X\beta\|_2^2 + \lambda \|w\|_2^2$$

$$\min_w \sum_i \xi_i + \frac{1}{C} \|w\|_2^2$$

$C =$

$$\begin{cases} \min_w \|w\| \\ y_i \langle w, x_i \rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

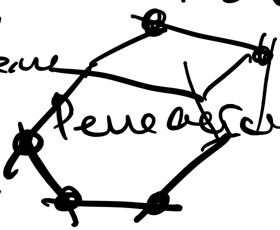
коэфф. регр.

polytype

норма весов

ошибки модели

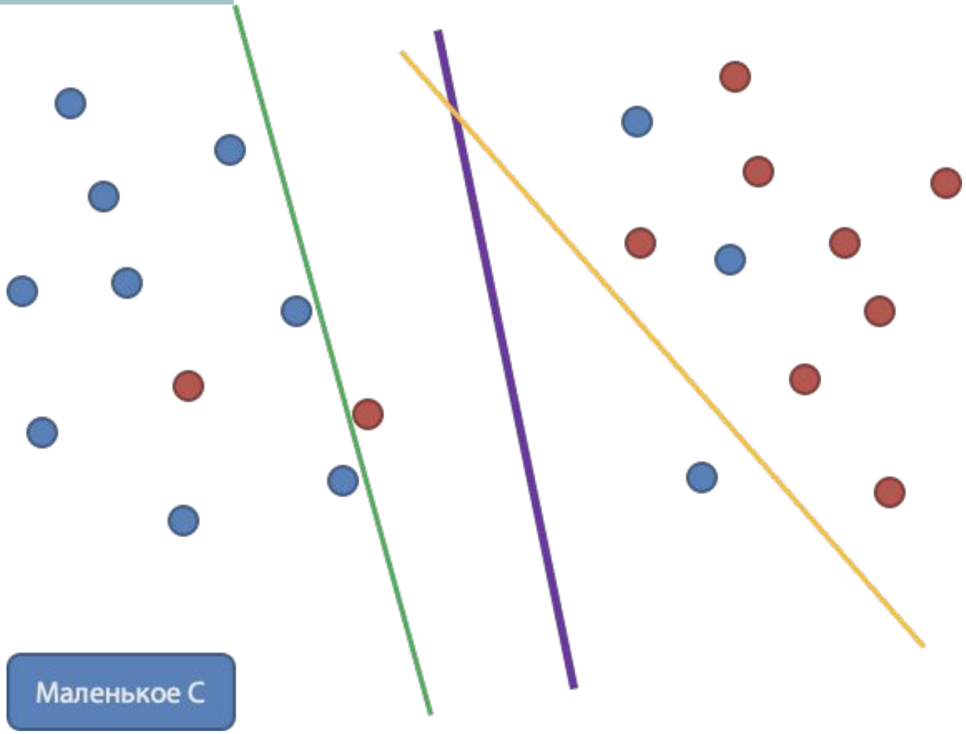
$$\begin{cases} \min_{w,\xi} \|w\|_2^2 + \boxed{C \sum_i \xi_i} \\ y_i \langle w, x_i \rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$



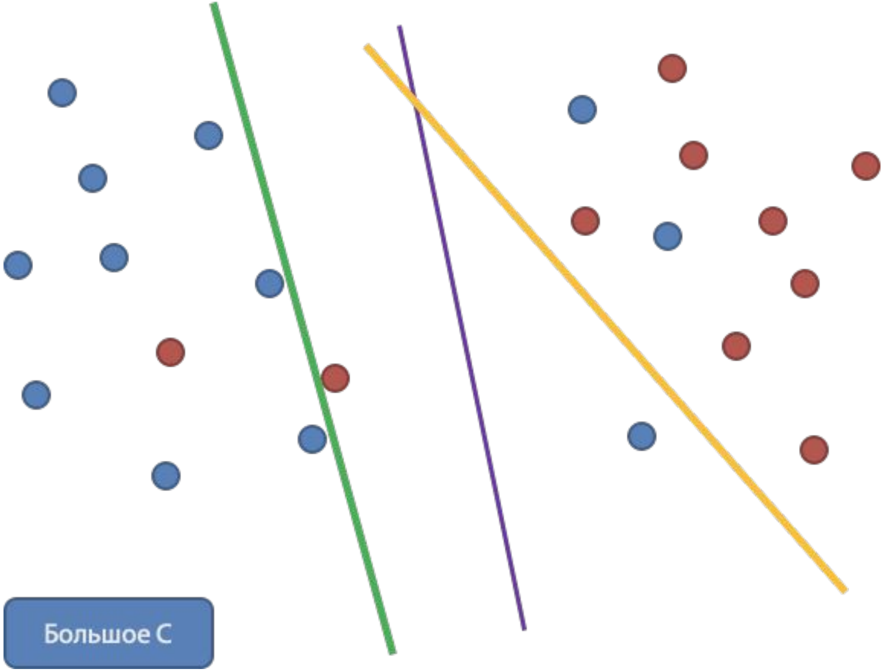
1. Решается

2. плохо масштабируется

Линейно неразделимый случай



Линейно неразделимый случай



Метод опорных векторов

$$\begin{cases} \min_{w,\xi} \|w\| + C \sum_i \xi_i \\ y_i \langle w, x_i \rangle \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Объединим ограничения

$$\xi_i \geq \max(0, 1 - y_i \langle w, x_i \rangle))$$

Итоговая задача оптимизации

$$\min_w \left(\|w\| + C \sum_i \max(0, 1 - y_i \langle w, x_i \rangle)) \right)$$

Метод опорных векторов

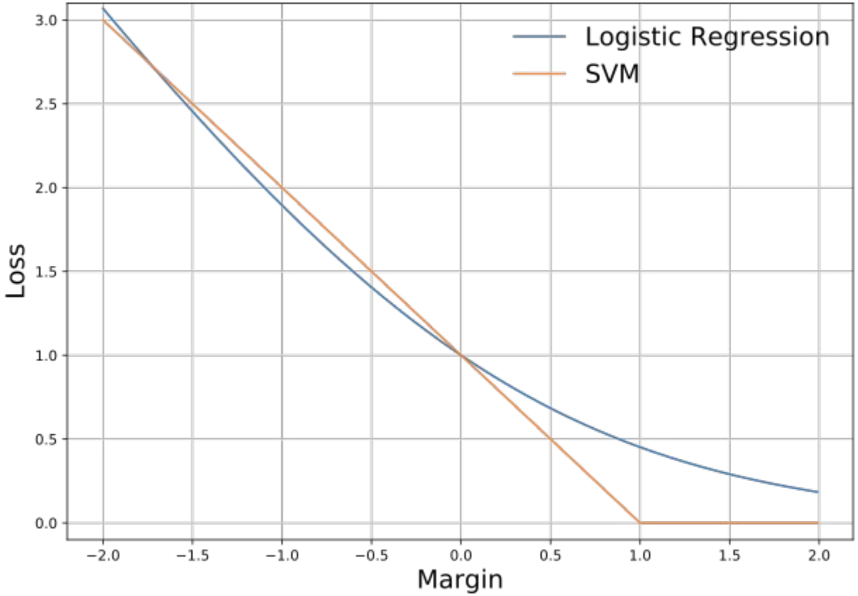
Итоговая задача оптимизации

$$\min_w \left(\|w\| + C \sum_i \max(0, 1 - y_i \langle w, x_i \rangle) \right)$$

Hingle Loss (верхняя оценка) + L2 регуляризация

$$\min_w \left(\sum_i \max(0, 1 - y_i \langle w, x_i \rangle) + \lambda \|w\| \right)$$

Сравнение логистической регрессии и SVM



Итого

- Метод опорных векторов основан на идее максимизации отступа классификатора
- В линейно неразделимом случае мы должны выбрать, что важнее, — ширина зазора или число ошибок
- По сути, SVM — это определённая функция потерь и регуляризация

1. SVM = $(\langle x_i, w\rangle) + \lambda\|w\|_2^2$

линейн.рег.

прямая задача

Xw

2. Двойственная

идея:

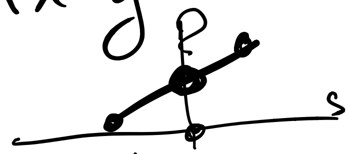
kernel

$y = Xw \qquad w^* = (X^T X + \lambda I)^{-1} X^T y$

$y = X(X^T X + \lambda I)^{-1} X^T y$

$= (\underbrace{XX^T}_{\sim} + \lambda I^{-1}) XX^T y$

3. $\langle x_i, x_j\rangle = \sum_n x_i^n x_j^n$



RBF.

$\langle x_i, x_j\rangle = \exp(-\gamma\|x_i - x_j\|^2)$ $\langle x_i, x_j\rangle$

$y_1 = \sum_{n=1}^{N} (\langle x_1, x_n\rangle) \cdot y_n$

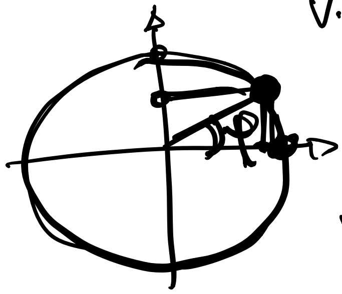
4. SVM kernel='rbf'

'poly'

$x_1^1 x_1^2 \quad x_1^{1^2} x_1^{2^2} \quad x_1 x$

признаки с переодичность

1)



v. $\begin{cases} \cos(\varphi+2\pi) \\ \sin(\varphi+2\pi) \end{cases}$

v. месяц покупки

время внутри дня

20T

TS

v $\cos\left(\frac{\varphi}{T}\right)$

режим: '2012-05-1 : 8:AM' → date.time

$\frac{H}{12}$ $\frac{H}{6}$ $\frac{H}{4}$

FT

$A\cos\left(\frac{H}{12}\right)\sin\left(\frac{H}{12}\right)$

2. DFT

3. Wavlets

Привет