

Задачи классификации. SVM



Дмитрий Меркушов
TeamLead @ Antispam ML

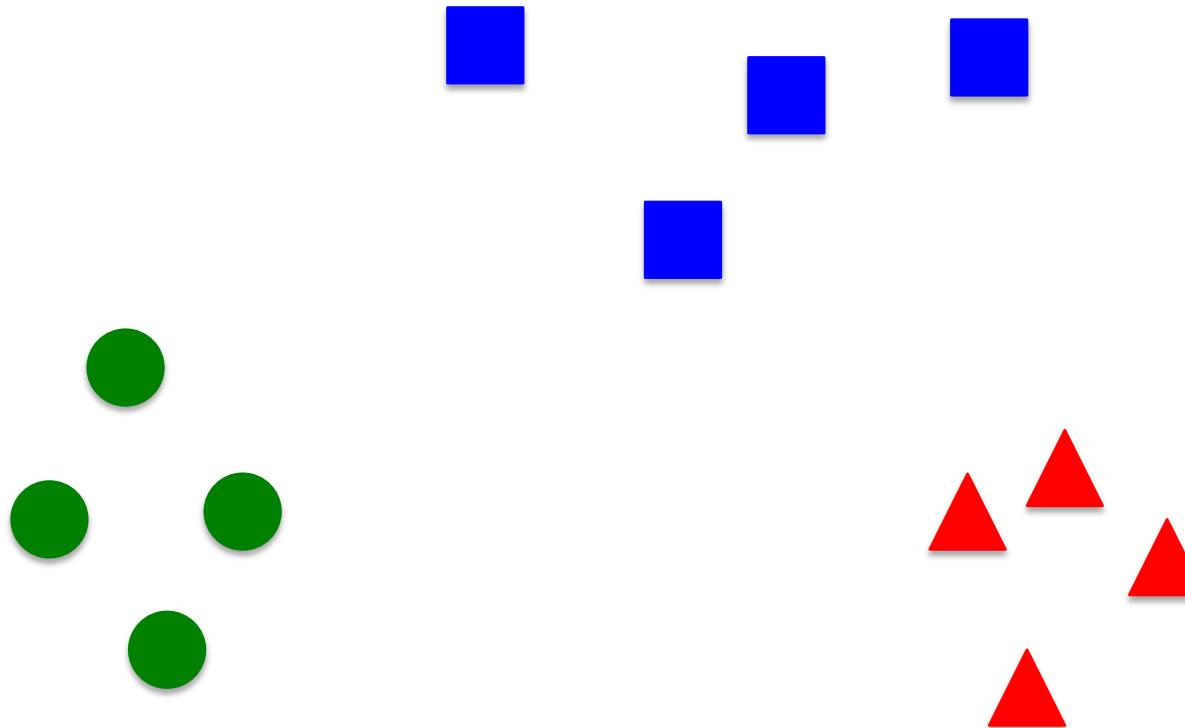
На прошлой лекции

- Разобрали логистическую регрессию
- Рассмотрели метрики оценки качества классификаторов
- Научились подбирать порог для классификаторов

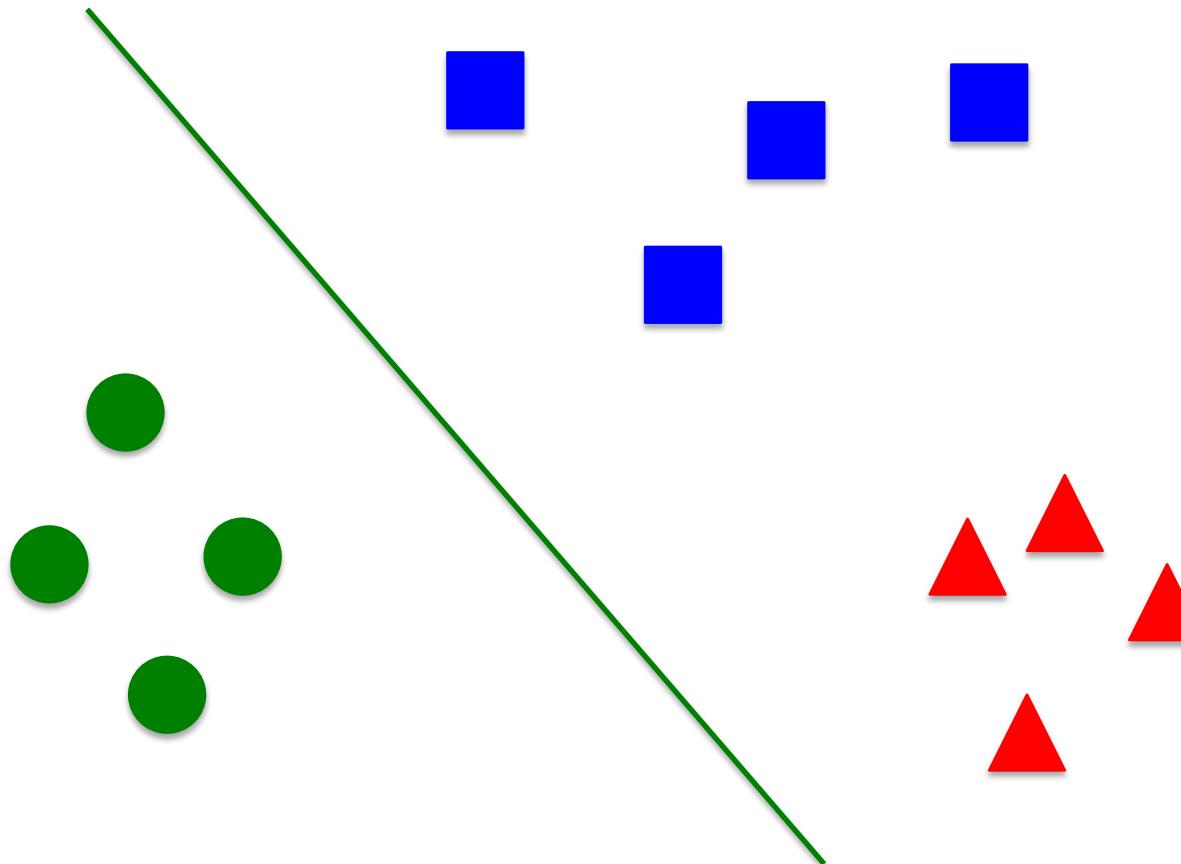
Multiclass классификация. One vs all



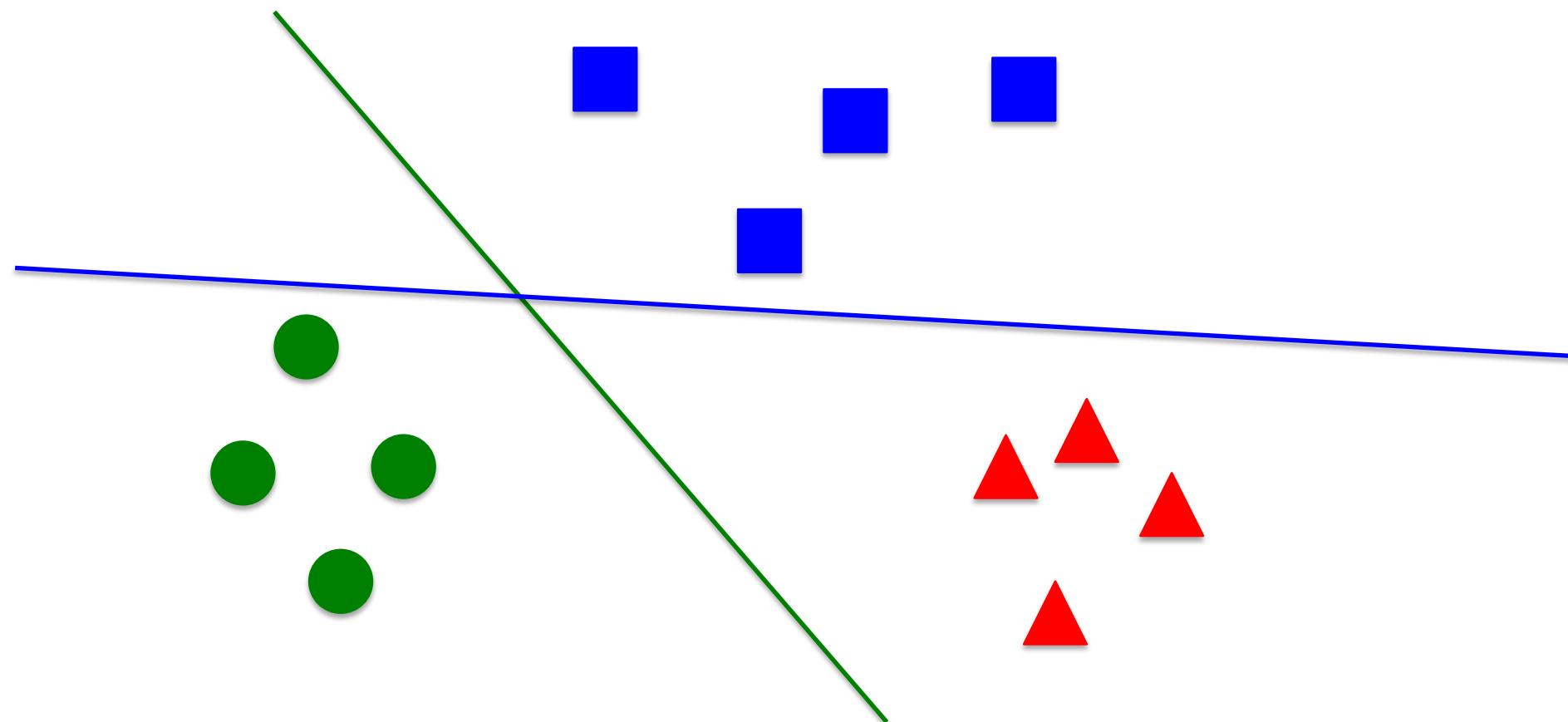
Многоклассовая классификация: one-vs-all



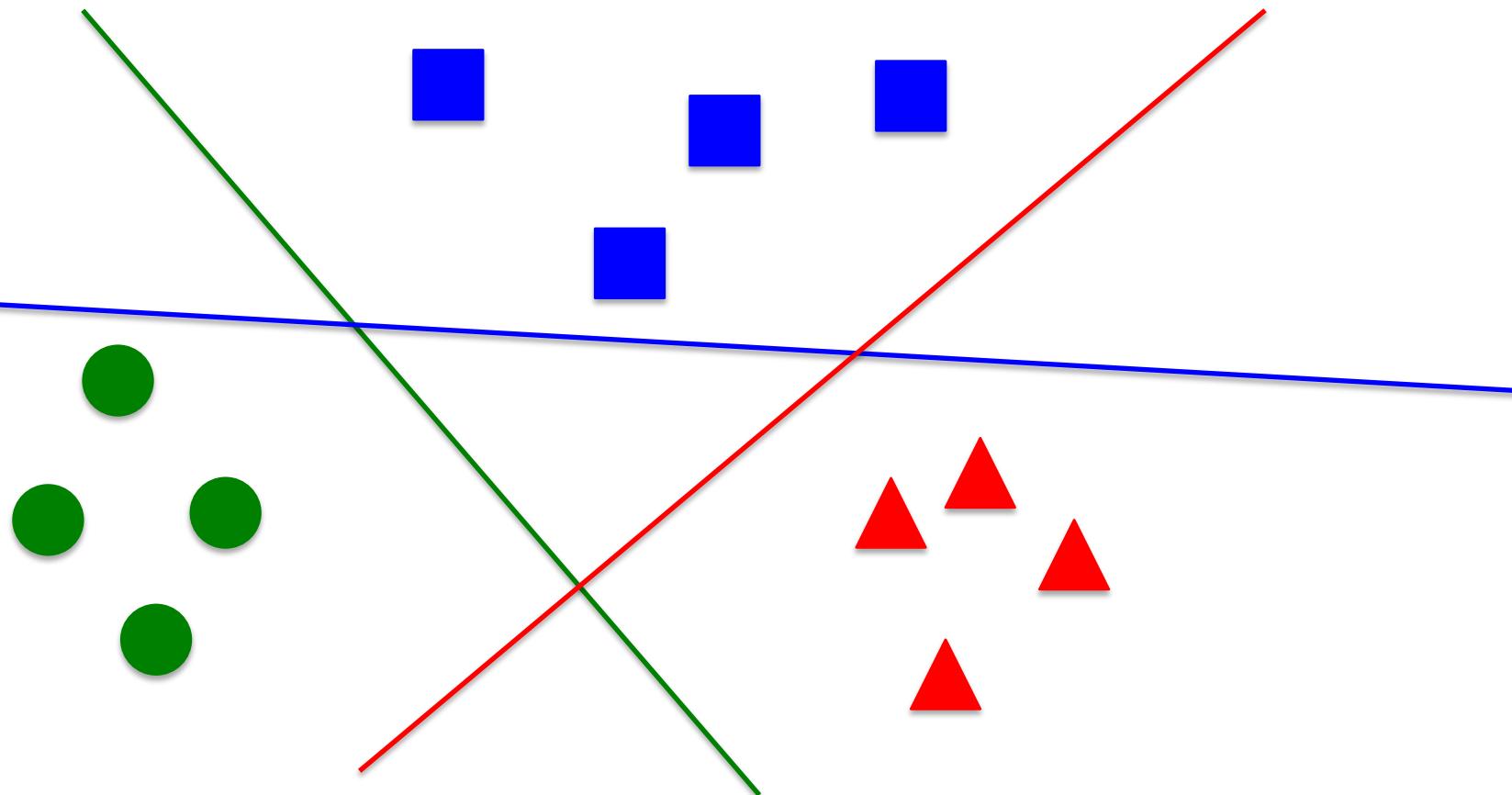
Многоклассовая классификация: one-vs-all



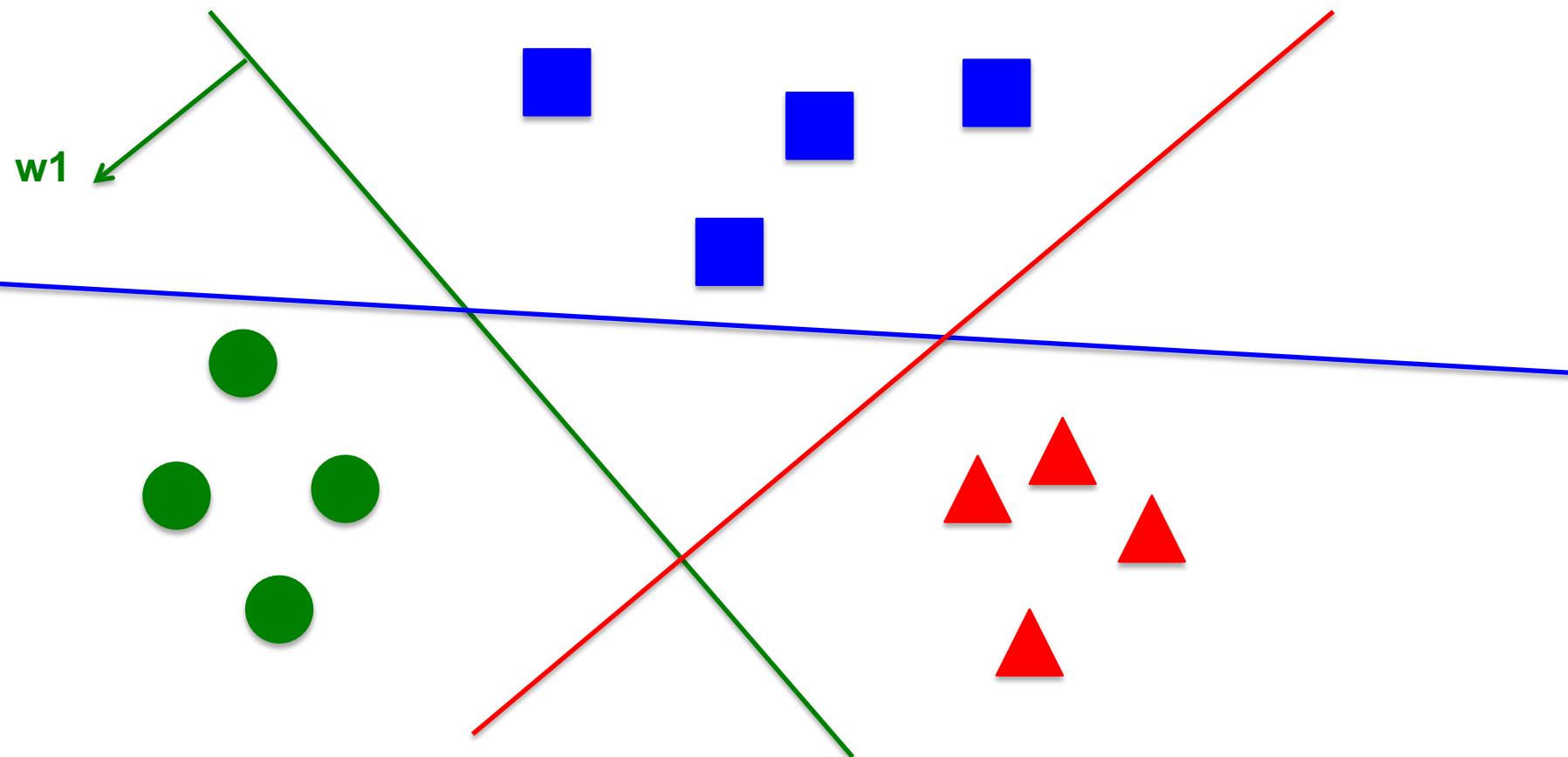
Многоклассовая классификация: one-vs-all



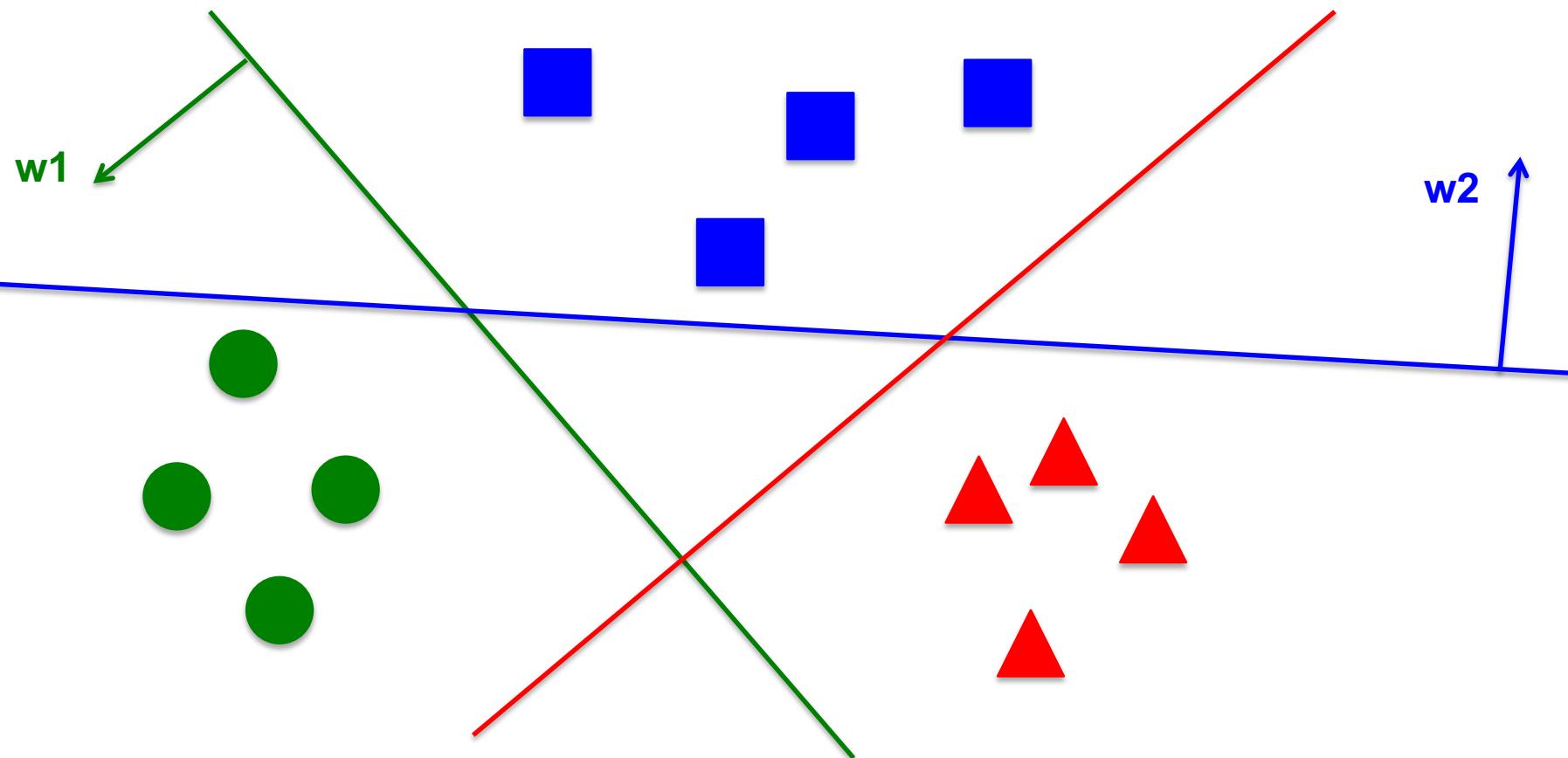
Многоклассовая классификация: one-vs-all



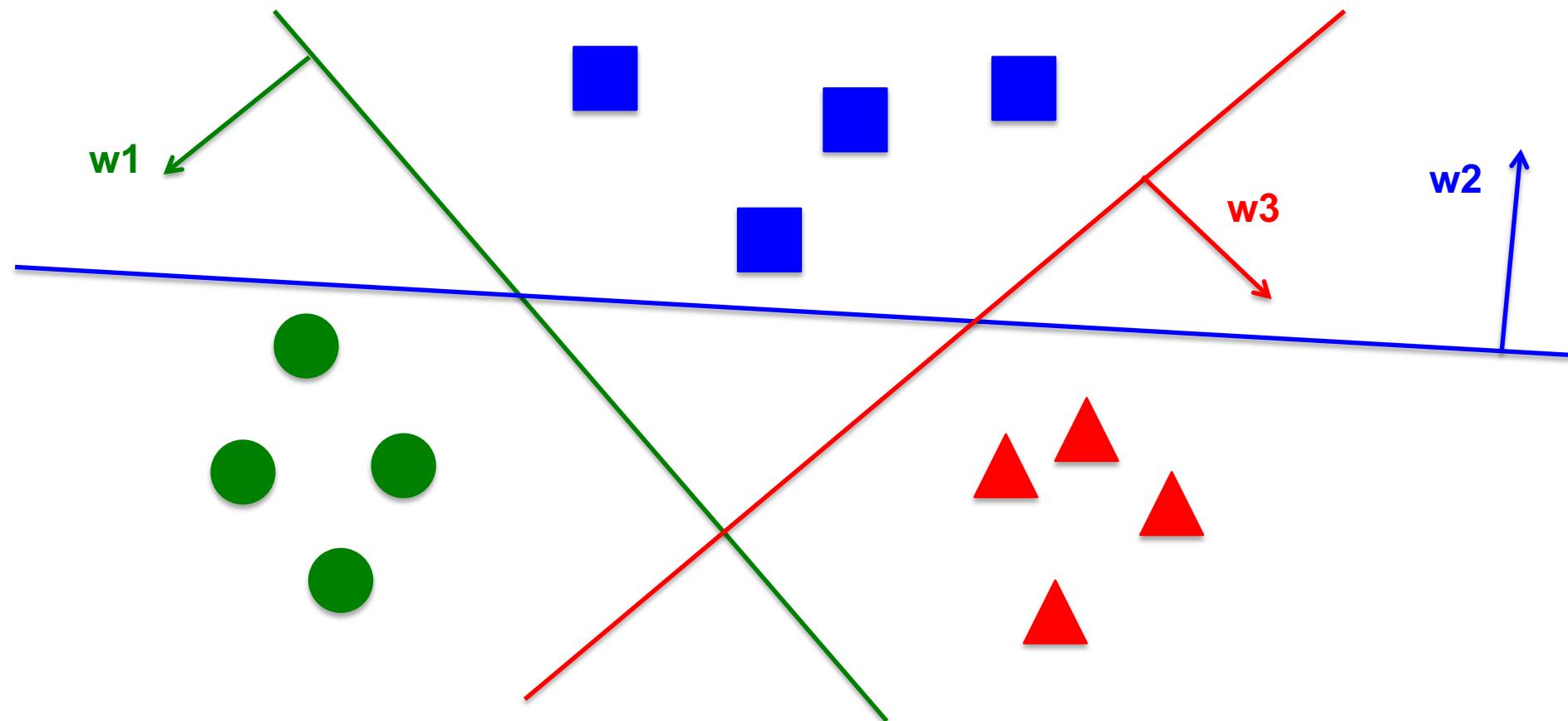
Многоклассовая классификация: one-vs-all



Многоклассовая классификация: one-vs-all



Многоклассовая классификация: one-vs-all



Как предсказать класс?



У какого класса больше вероятность, тот и предсказываем

Как предсказать класс?

У какого класса больше вероятность, тот и предсказываем

w1 → 0.3

w2 → 0.8

w3 → 0.15

Как предсказать класс?

У какого класса больше вероятность, тот и предсказываем

w1 → 0.3

w2 → 0.8

w3 → 0.15

Как предсказать класс?



Какой классификатор сработал, такую метку и ставим

Как предсказать класс?

Какой классификатор сработал, такую метку и ставим

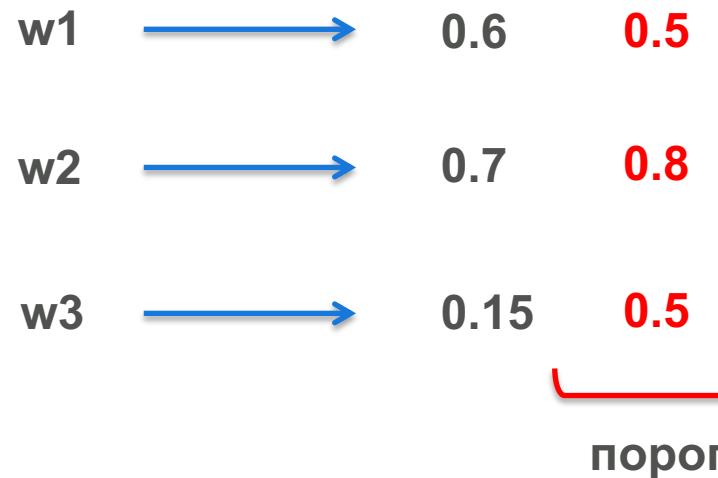
w1 → 0.6

w2 → 0.7

w3 → 0.15

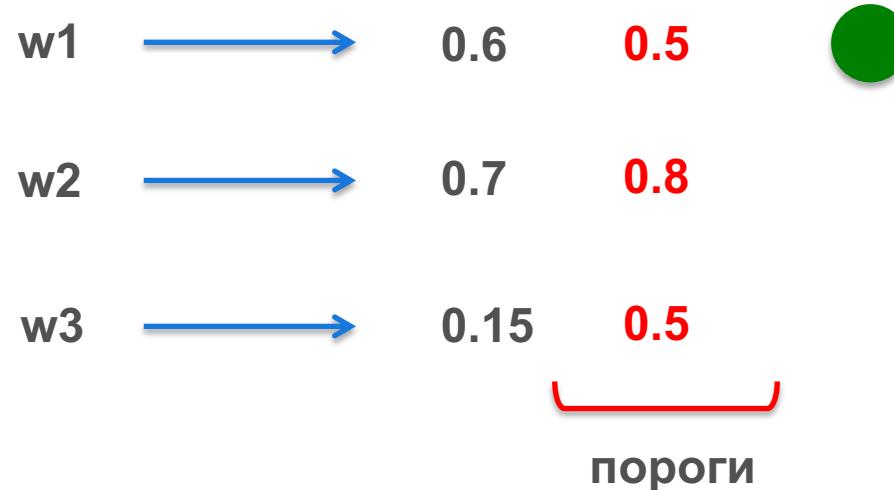
Как предсказать класс?

Какой классификатор сработал, такую метку и ставим



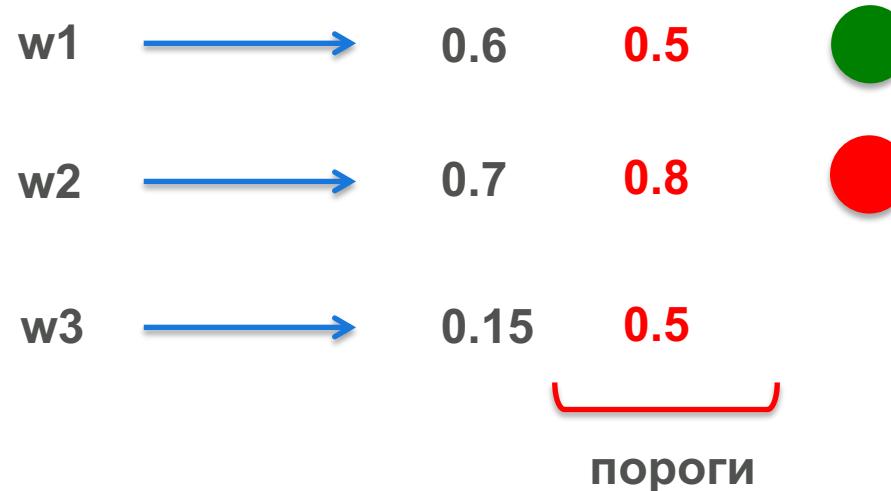
Как предсказать класс?

Какой классификатор сработал, такую метку и ставим



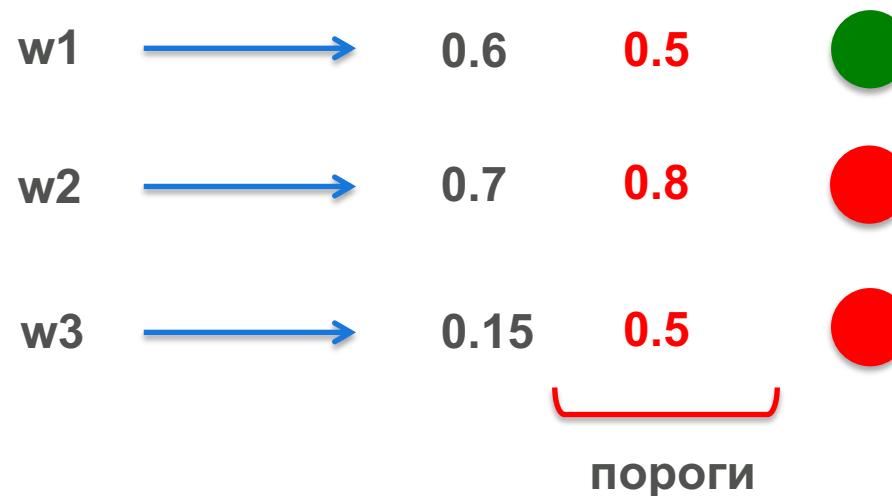
Как предсказать класс?

Какой классификатор сработал, такую метку и ставим



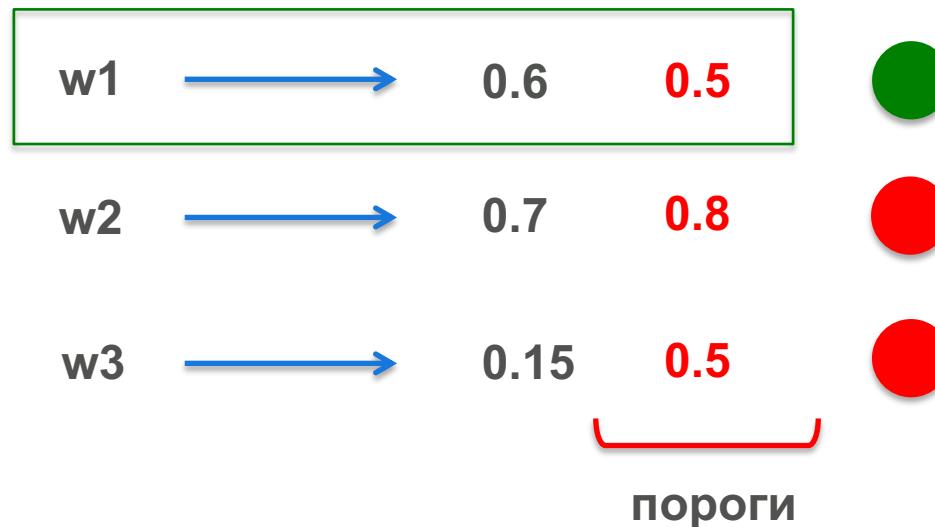
Как предсказать класс?

Какой классификатор сработал, такую метку и ставим



Как предсказать класс?

Какой классификатор сработал, такую метку и ставим



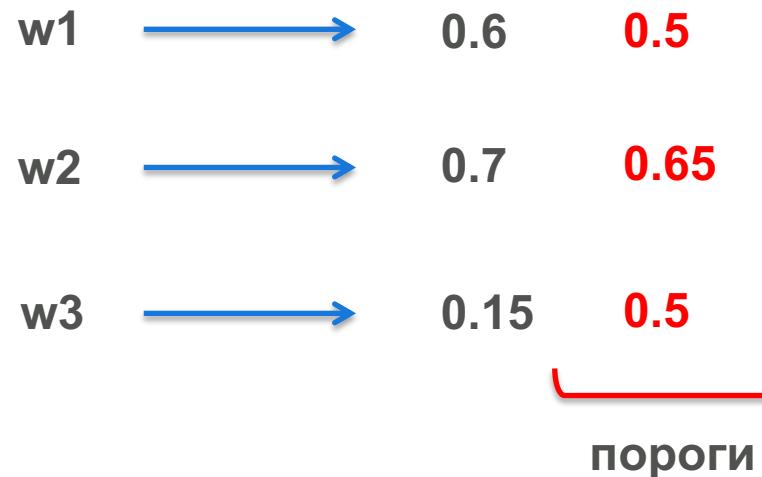
Как предсказать класс?



Если сработало несколько классификаторов, можно выбрать метку того, который дает максимальную вероятность

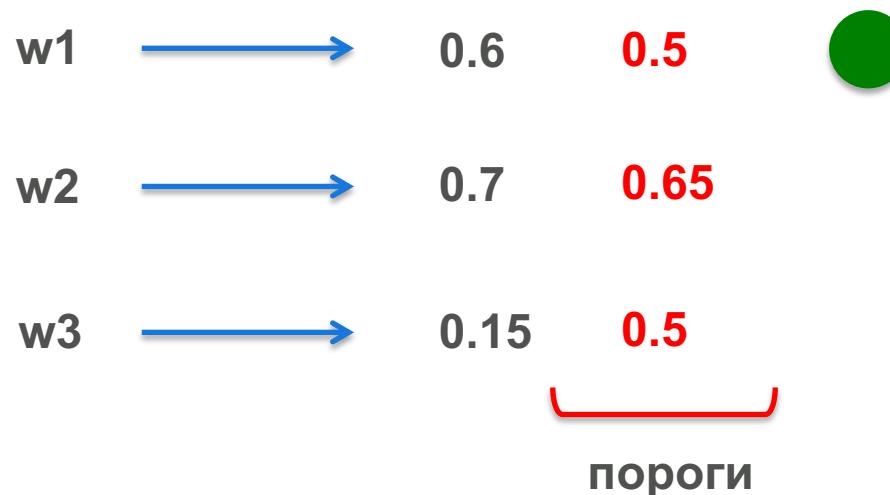
Как предсказать класс?

Если сработало несколько классификаторов, можно выбрать метку того, который дает максимальную вероятность



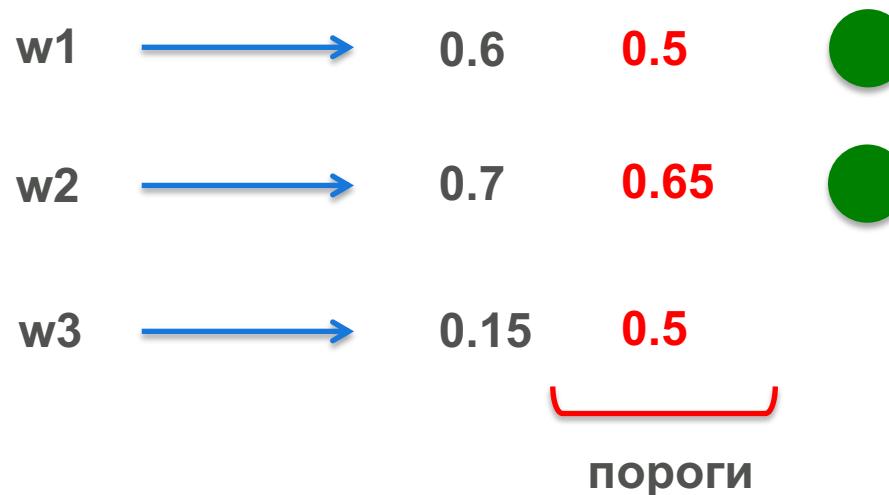
Как предсказать класс?

Если сработало несколько классификаторов, можно выбрать метку того, который дает максимальную вероятность



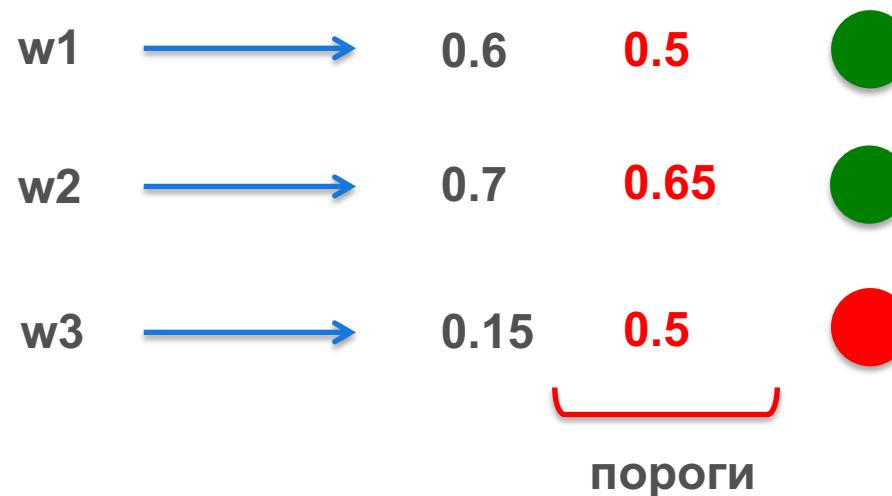
Как предсказать класс?

Если сработало несколько классификаторов, можно выбрать метку того, который дает максимальную вероятность



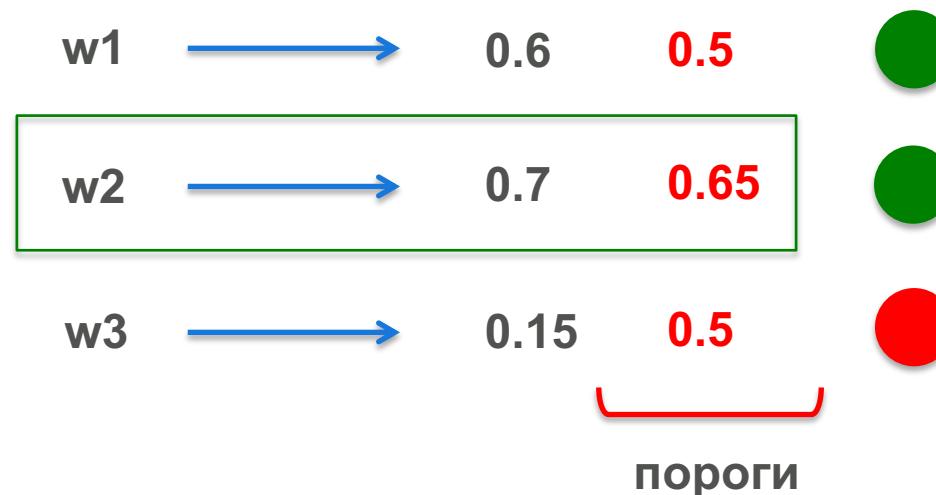
Как предсказать класс?

Если сработало несколько классификаторов, можно выбрать метку того, который дает максимальную вероятность



Как предсказать класс?

Если сработало несколько классификаторов, можно выбрать метку того, который дает максимальную вероятность

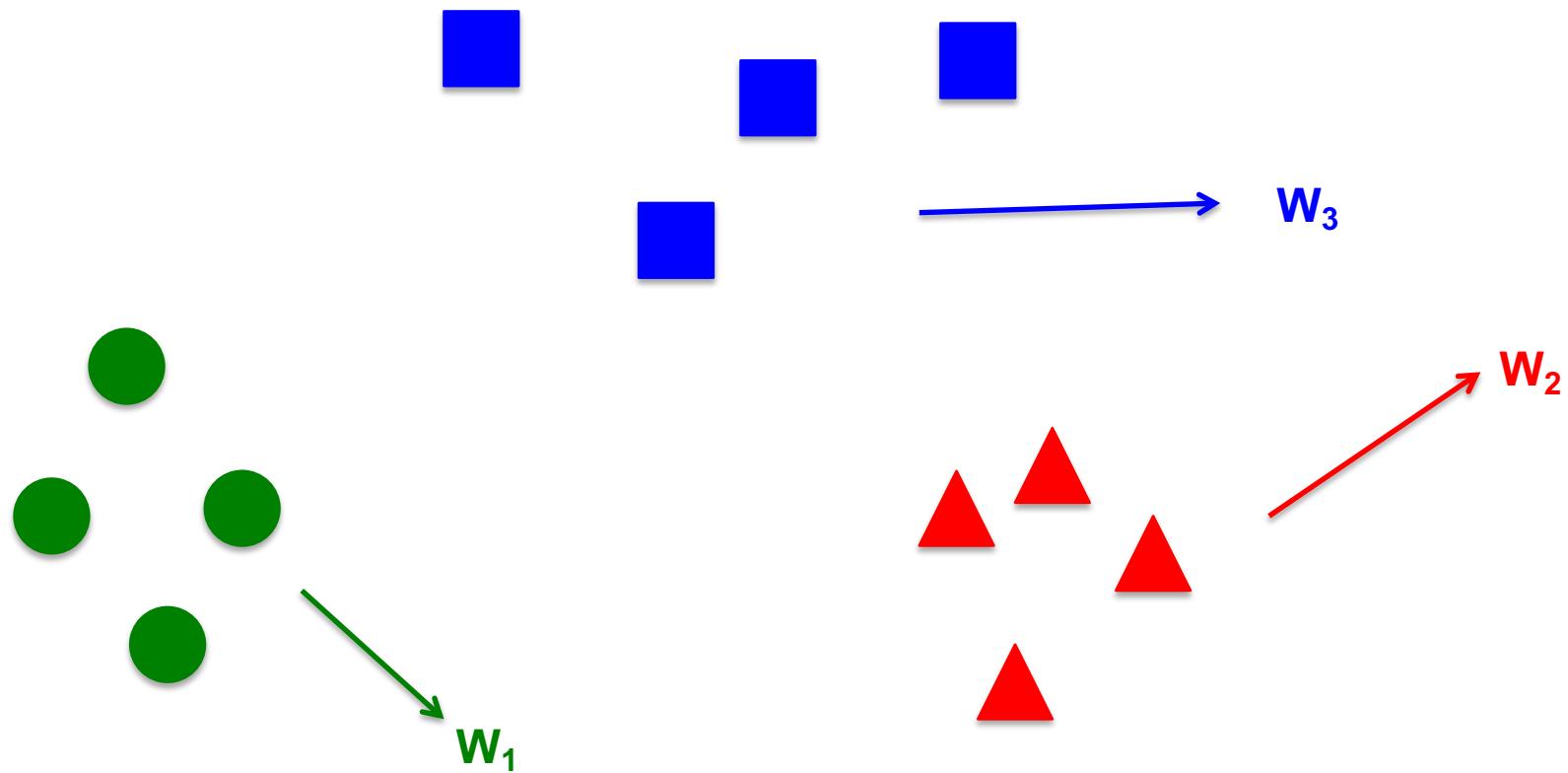


- Число классификаторов = числу классов N
- Итоговое число параметров для оптимизации $N^* \{X\}$
- Возможны варианты по окончательной классификации
 - Нельзя напрямую упорядочивать по скорам N классификаторов

Multiclass классификация. Softmax



Многоклассовая классификация: softmax



Многоклассовая классификация: softmax

w1

w2

w3

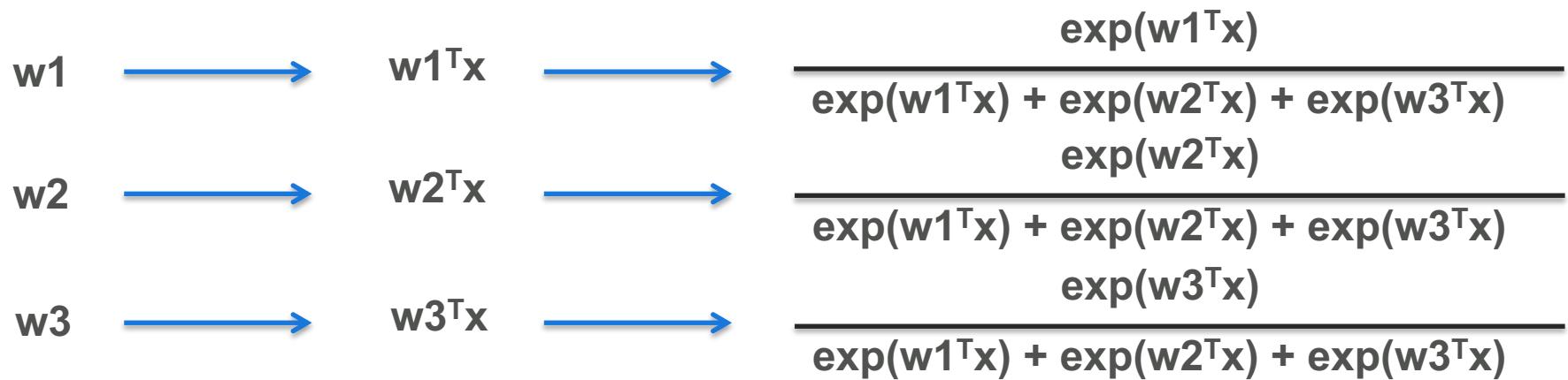
Многоклассовая классификация: softmax

$$w_1 \longrightarrow w_1^T x$$

$$w_2 \longrightarrow w_2^T x$$

$$w_3 \longrightarrow w_3^T x$$

Многоклассовая классификация: softmax



Многоклассовая классификация: softmax

$$\frac{\exp(w_1^T x)}{\frac{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)}{\frac{\exp(w_2^T x)}{\frac{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)}{\frac{\exp(w_3^T x)}{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)}}}}$$

Многоклассовая классификация: softmax

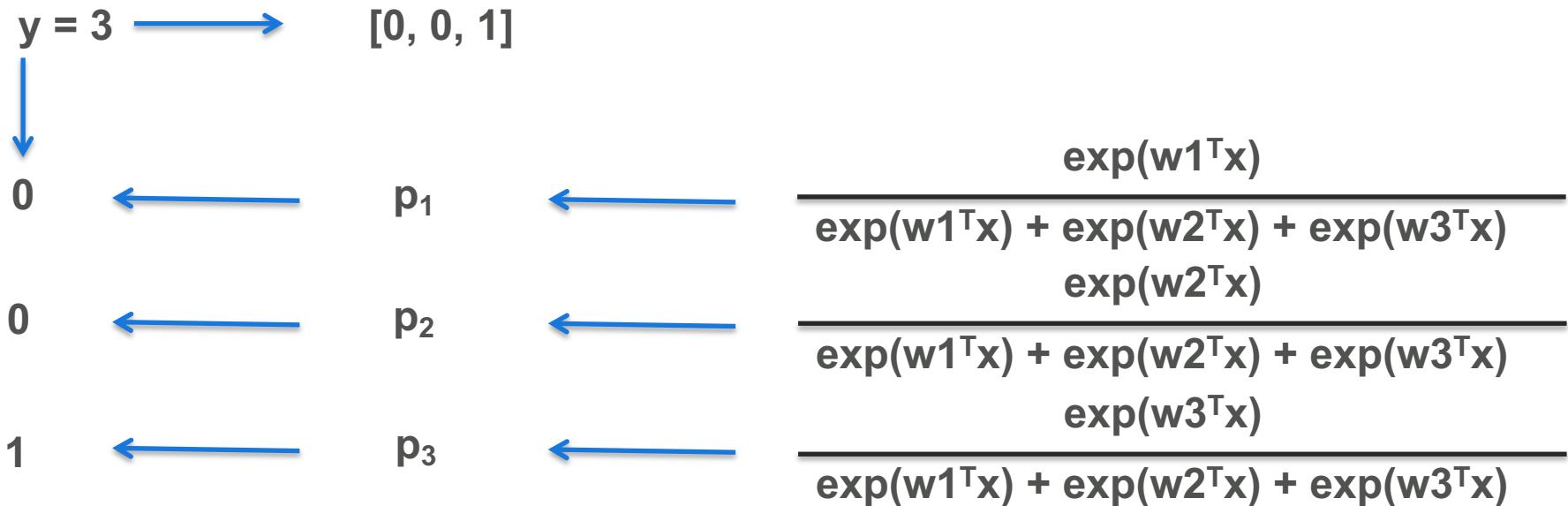
$$\begin{array}{ll} p_1 & \xleftarrow{\text{---}} \frac{\exp(w_1^T x)}{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)} \\ p_2 & \xleftarrow{\text{---}} \frac{\exp(w_2^T x)}{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)} \\ p_3 & \xleftarrow{\text{---}} \frac{\exp(w_3^T x)}{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)} \end{array}$$

Многоклассовая классификация: softmax

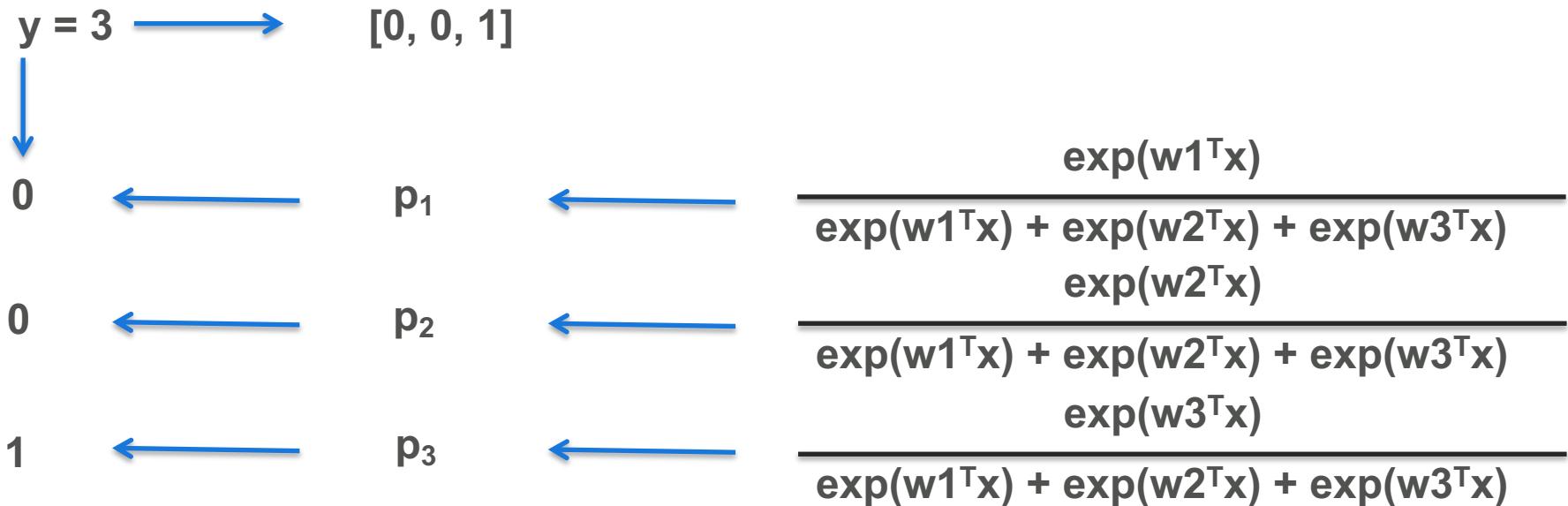
$y = 3 \longrightarrow [0, 0, 1]$

$$\begin{array}{ll} p_1 & \xleftarrow{\quad} \\ \frac{\exp(w_1^T x)}{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)} & \\ p_2 & \xleftarrow{\quad} \\ \frac{\exp(w_2^T x)}{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)} & \\ p_3 & \xleftarrow{\quad} \\ \frac{\exp(w_3^T x)}{\exp(w_1^T x) + \exp(w_2^T x) + \exp(w_3^T x)} & \end{array}$$

Многоклассовая классификация: softmax



Многоклассовая классификация: softmax



$$\text{Error}(x, y, w1, w2, w3) = -(y_1 \log p_1 + y_2 \log p_2 + y_3 \log p_3)$$

Многоклассовая классификация: softmax

Бинарный log-loss: $L(X, Y, W) = -\sum_i^L \log p_i(x_i, w)$

Кросс-энтропия:

$$L(X, Y, W) = -\sum_i^L \sum_j^N [j = \text{true}] \log p_{ij}(x_i, w_j)$$

L – количество рекордов

N – количество классов

Softmax Classifier (Multinomial Logistic Regression)



$$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

unnormalized probabilities

cat

3.2

car

5.1

frog

-1.7

exp

24.5

164.0

0.18

normalize

0.13

0.87

0.00

$$\begin{aligned} L_i &= -\log(0.13) \\ &= 0.89 \end{aligned}$$

unnormalized log probabilities

probabilities

- Один классификатор – 'multinomial logistic regression'
 - *Примерно та же интуиция*
- Итоговое число параметров для оптимизации $N^*\{\lambda\}$
- Напрямую упорядочиваем по N аутпутам, выбираем max

Метрики многоклассовой классификации



Метрики многоклассовой оценки

- Accuracy = $\frac{1}{L} \sum_L [y_i' = y_{true}]$
 - Доля правильных ответов классификатора
 - 'Двуклассовый precision'

Метрики многоклассовой оценки

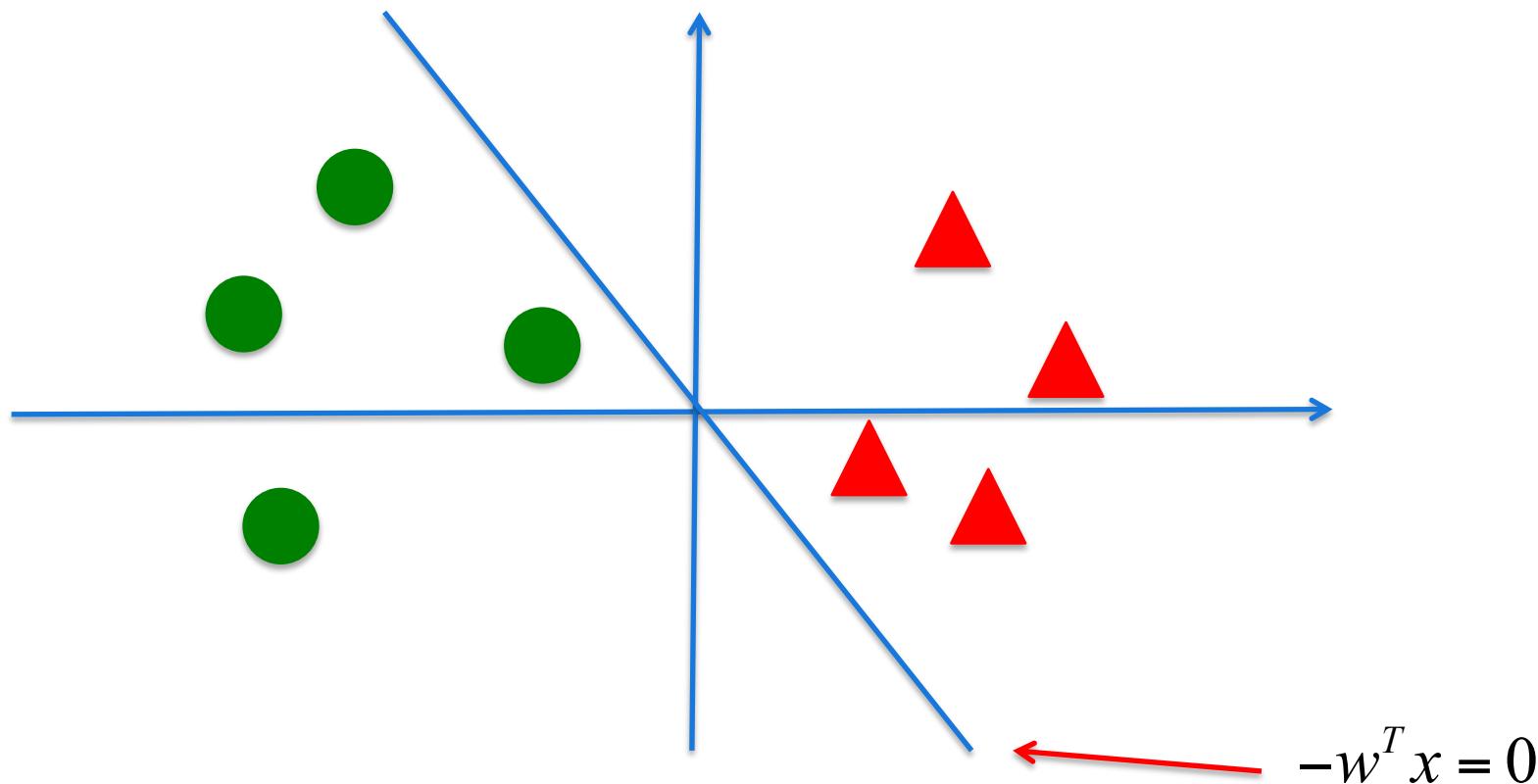
- Accuracy = $\frac{1}{L} \sum_L [y_i' = y_{true}]$
 - Доля правильных ответов классификатора
 - 'Двуклассовый precision'
- Micro-averaging
 - Усреднение confusion matrix по классам One Vs All
 - Нужная метрика – по усредненной матрице

- Accuracy = $\frac{1}{L} \sum_L [y_i' = y_{true}]$
 - Доля правильных ответов классификатора
 - 'Двуклассовый precision'
- Micro-averaging
 - Усреднение confusion matrix по классам One Vs All
 - Нужная метрика – по усредненной матрице
- Macro-averaging
 - Нужная метрика – для каждого класса в отдельности
 - Усреднение метрики

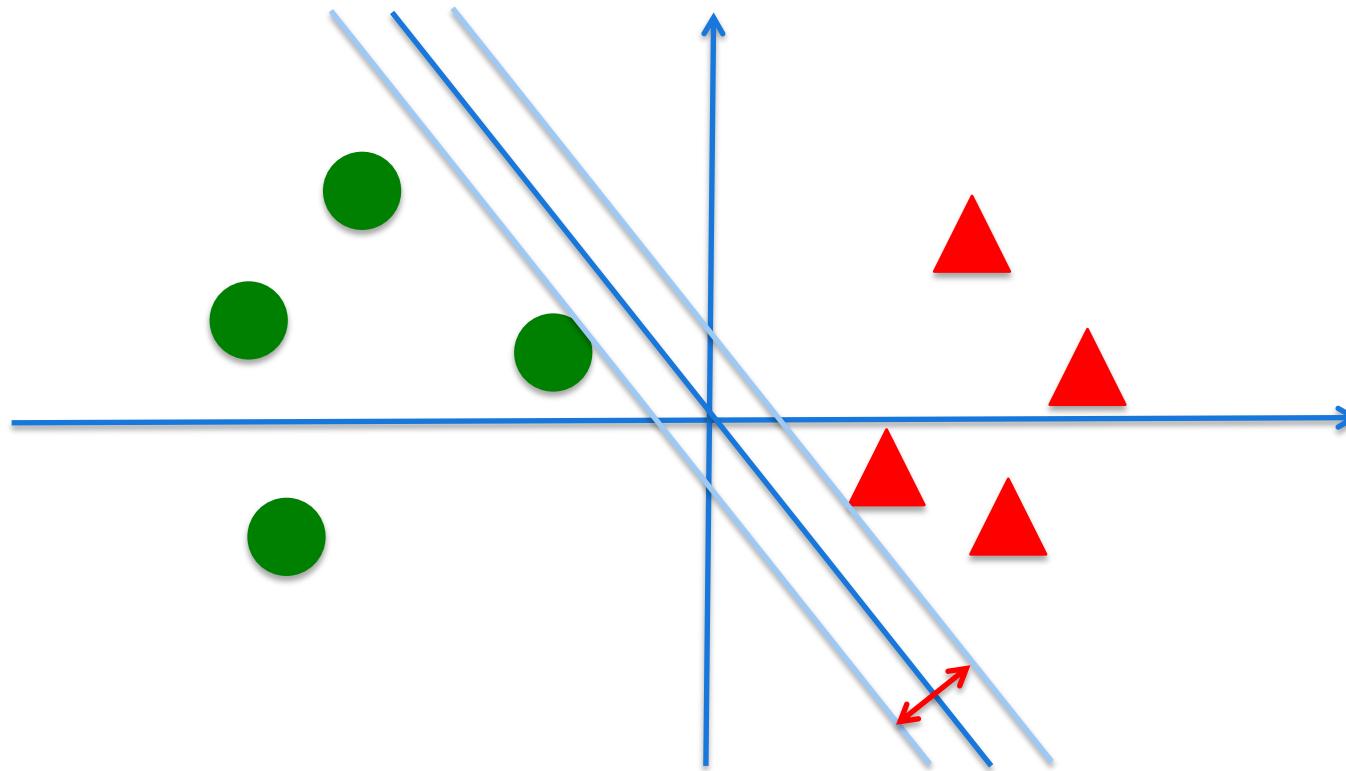
Линейная классификация. Support Vector Machine



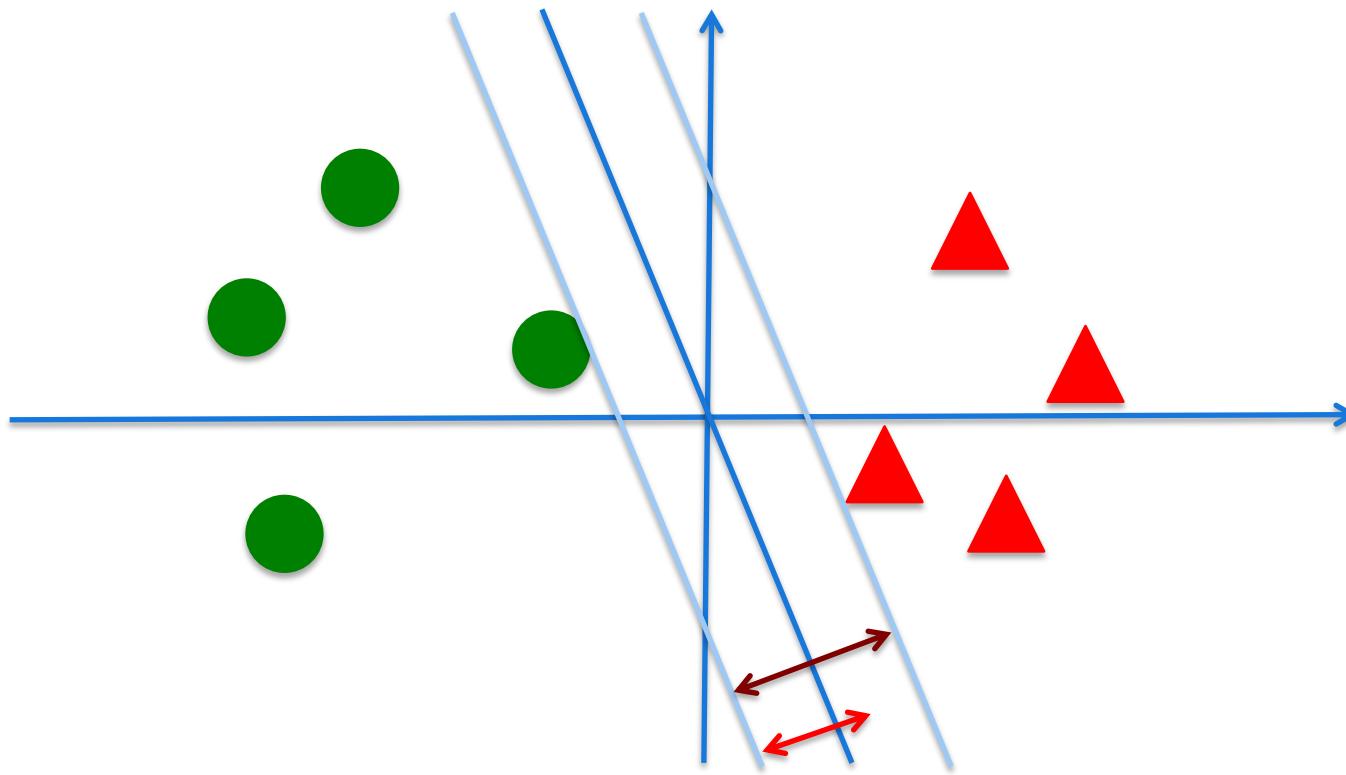
Геометрическая интерпретация



Попробуем найти максимальный зазор



Попробуем найти максимальный зазор



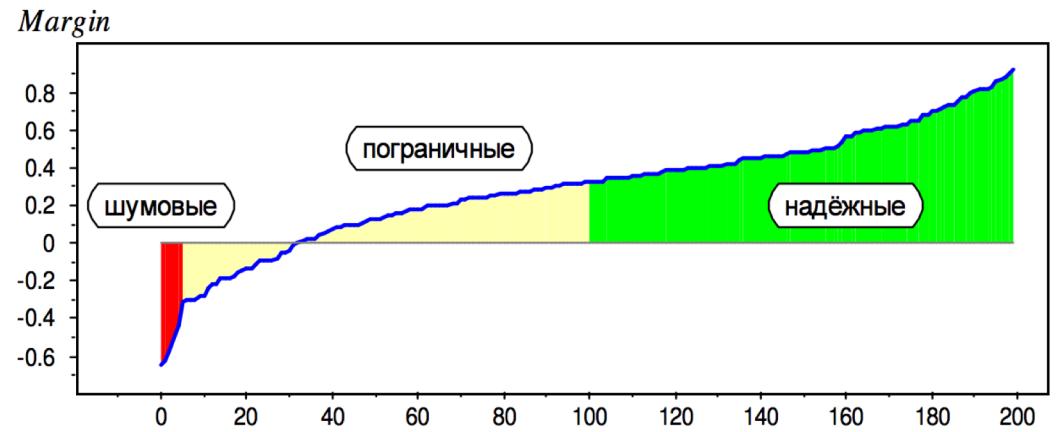
Построение разделяющей поверхности

- Задача классификации на 2 класса $Y=\{1, -1\}$
- Обучающая выборка $X=(x_i, y_i), i=1, L$
- Построить алгоритм классификации $a(x, w) = \text{sign } f(x, w)$

$f(x, w)=0$ – разделяющая поверхность

Отступ

- $f(x, w) = 0$ – разделяющая поверхность
- $M_i(w) = y_i f(x_i, w)$ – отступ объекта i (Margin)
- $M_i(w) < 0 \Rightarrow$ ошибка алгоритма а на объекте i

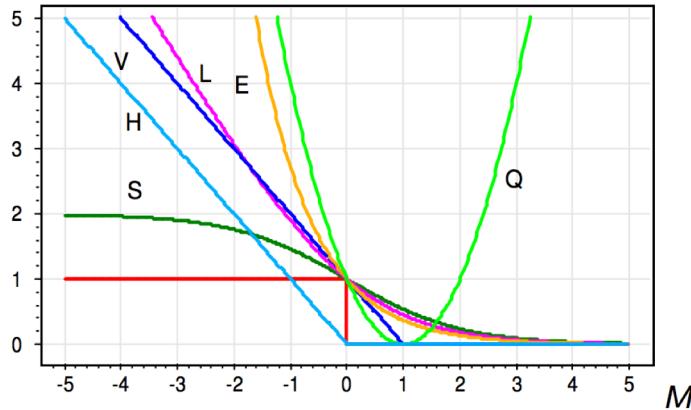


Функционал эмпирического риска

- $M_i(w) < 0 \Rightarrow$ ошибка алгоритма а на объекте i
- Эмпирический риск
 - $Q(w) = \sum_i^L [M_i(w) < 0]$
 - Гладкая аппроксимация $Q \leq Q'(w) = \sum_i^L L(M_i(w))$
 - $Q \rightarrow \min$

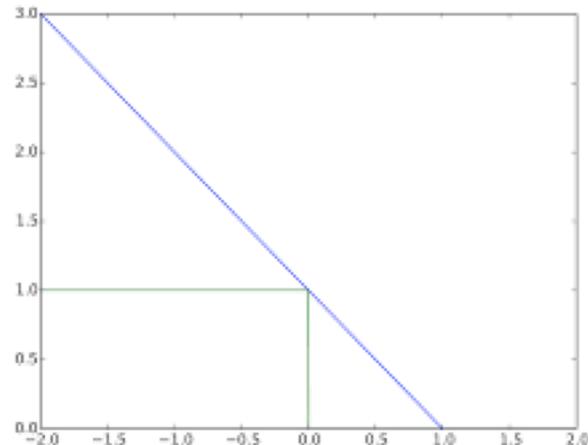
Аппроксимации

Часто используемые непрерывные функции потерь $\mathcal{L}(M)$:



- | | |
|-----------------------------|-----------------------------------|
| $V(M) = (1 - M)_+$ | — кусочно-линейная (SVM); |
| $H(M) = (-M)_+$ | — кусочно-линейная (Hebb's rule); |
| $L(M) = \log_2(1 + e^{-M})$ | — логарифмическая (LR); |
| $Q(M) = (1 - M)^2$ | — квадратичная (FLD); |
| $S(M) = 2(1 + e^M)^{-1}$ | — сигмоидная (ANN); |
| $E(M) = e^{-M}$ | — экспоненциальная (AdaBoost); |

Апроксимация SVM. Hinge Loss



$$L(M) = \max(0, 1-M)$$

Linear SVM на 1 слайде

- $L = \max(0, 1 - M)$
- $Q'(w) = \sum_i^L \max(0, 1 - M_i) + \frac{1}{2C} \|w\|^2 \rightarrow \min_w$
- Оптимизация градиентным спуском (но решается иначе)
- $a(x_i, w_i) = \text{sign}(x_i, w_i)$

Мотивация

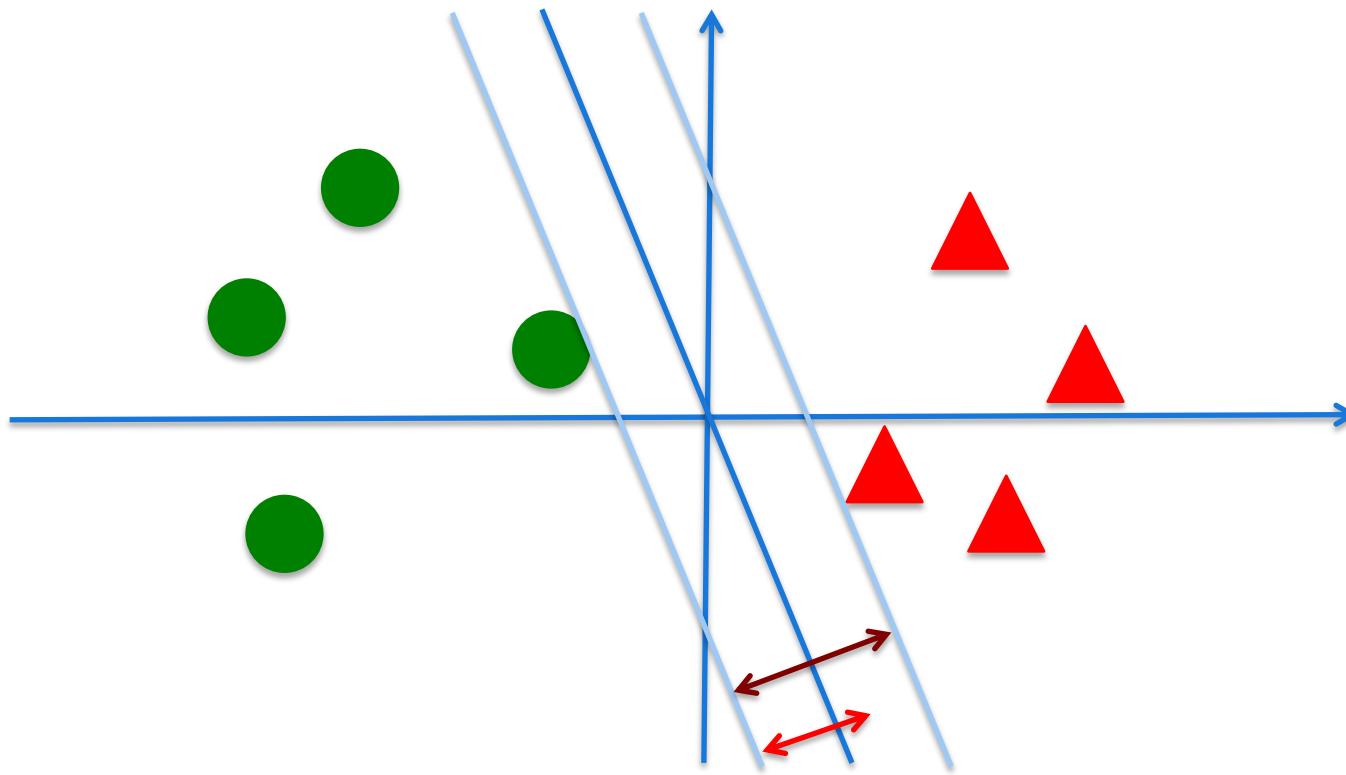
- В исходной постановке – классическая задача минимизации эмпирического риска с регуляризацией
- *Но откуда такой Loss?*
- *Почему такой регуляризатор?*



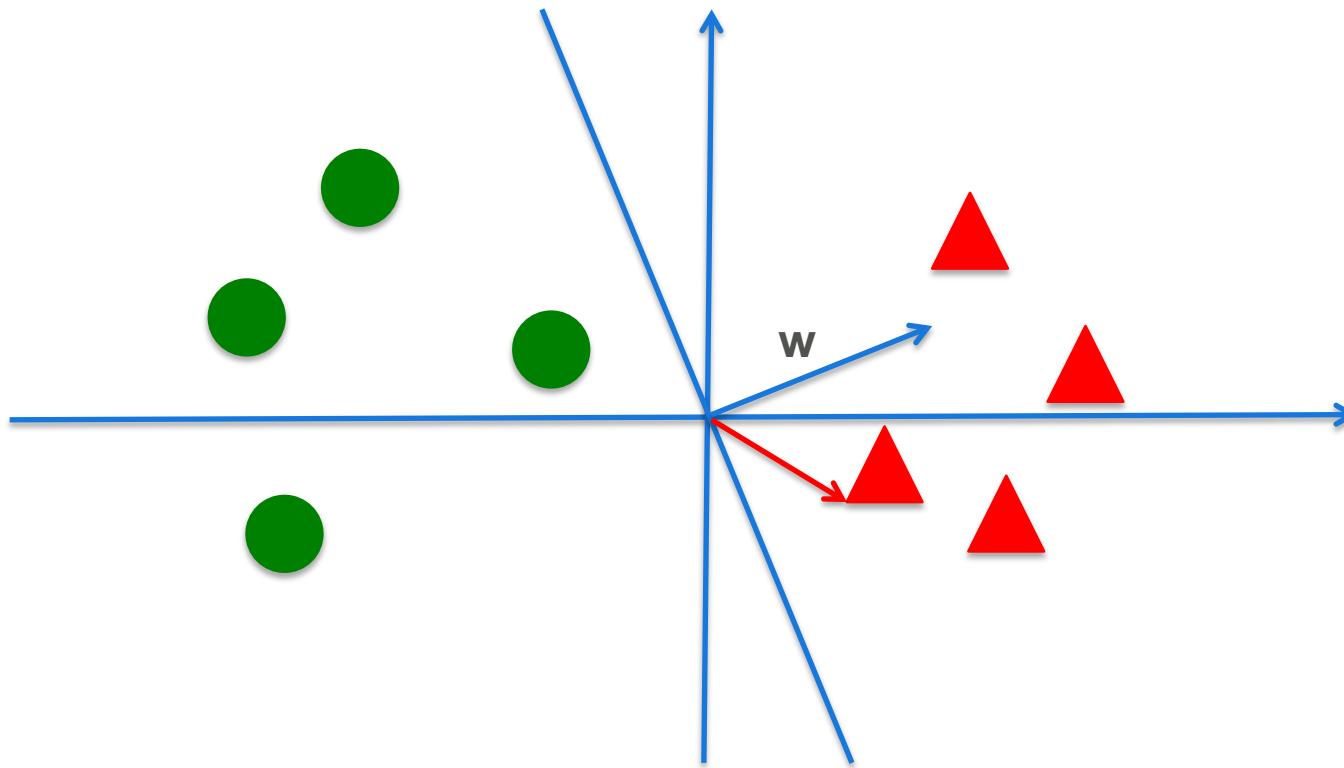
SVM. Линейно разделимый случай



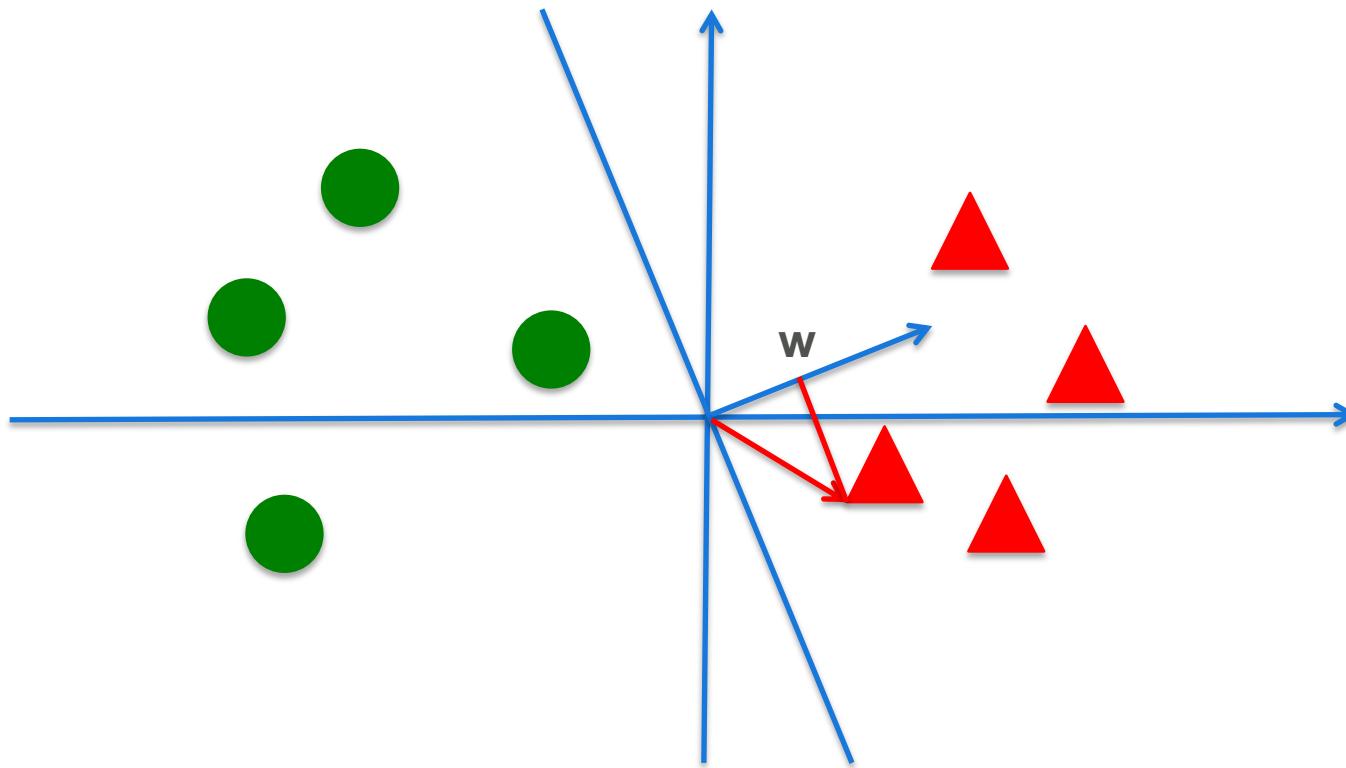
Попробуем найти максимальный зазор



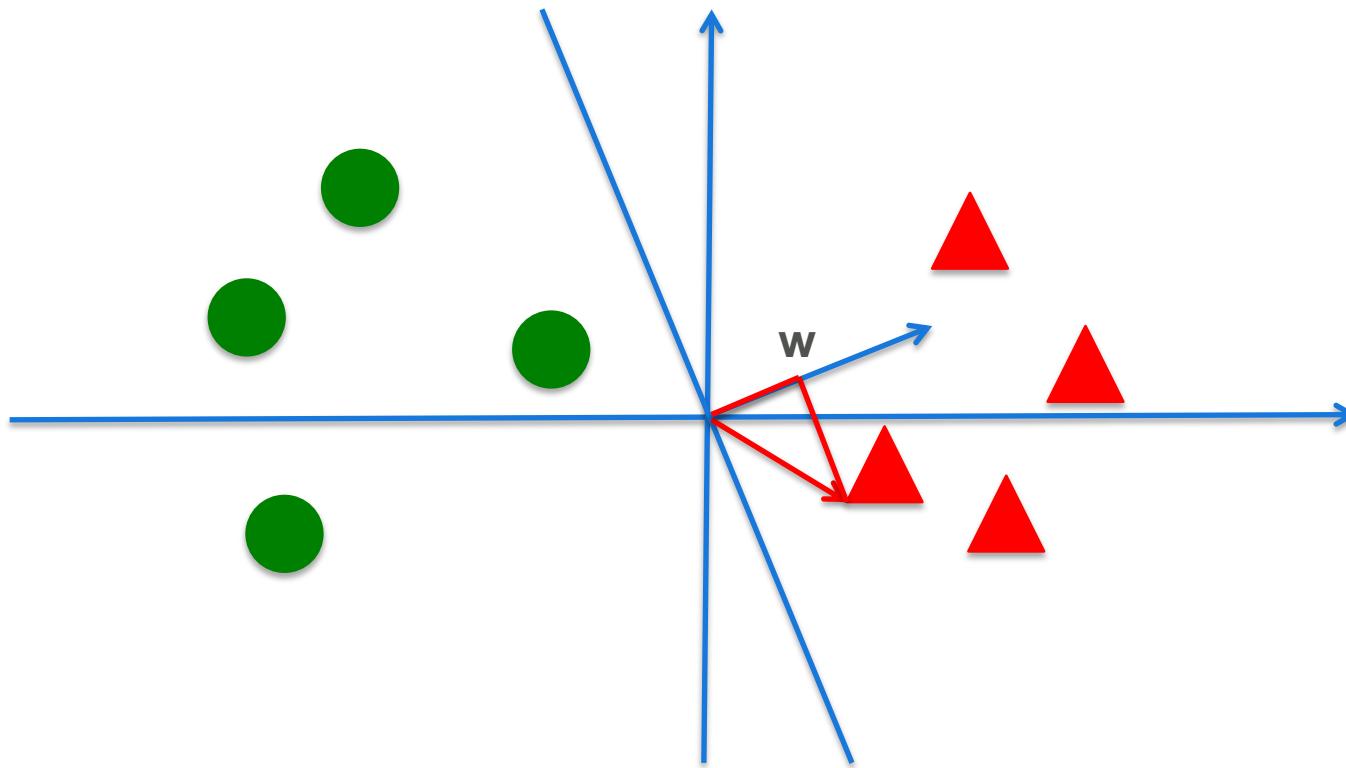
Попробуем найти максимальный зазор



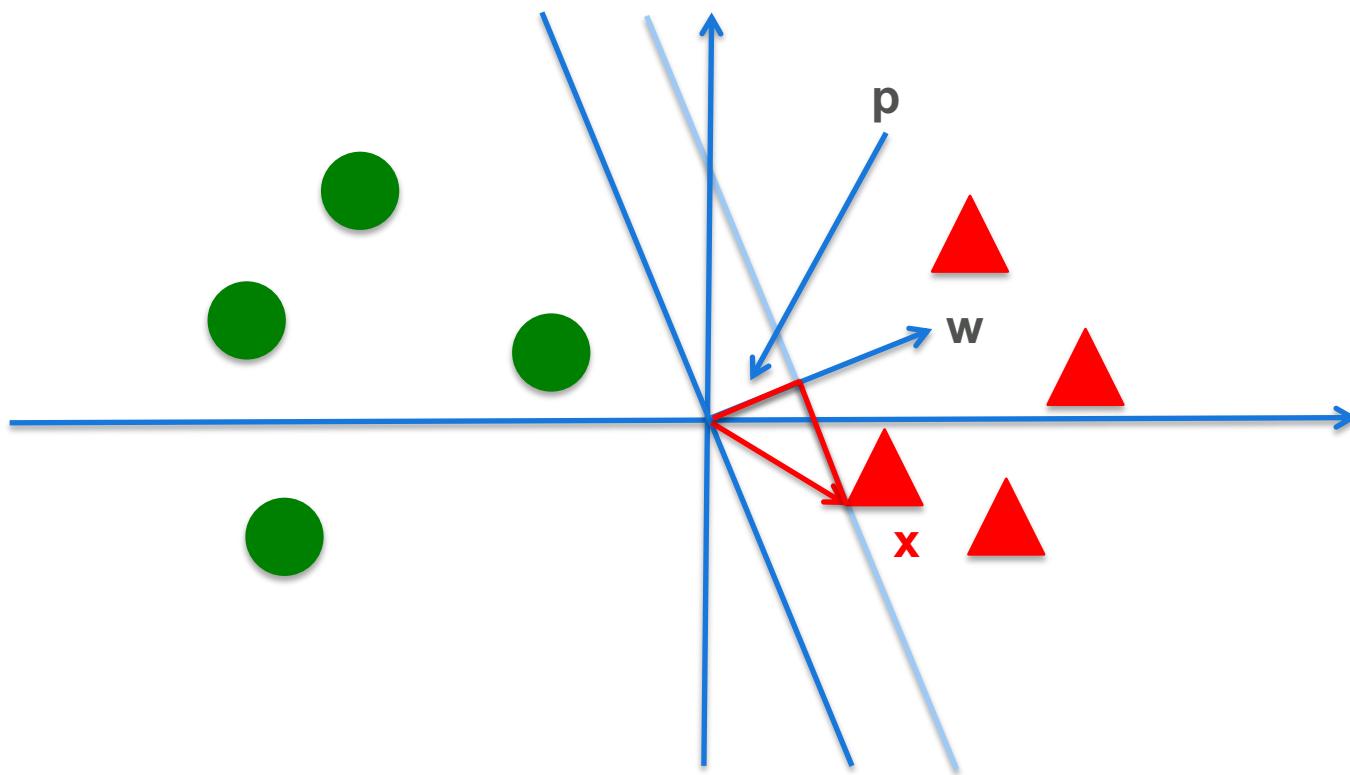
Попробуем найти максимальный зазор



Попробуем найти максимальный зазор



Попробуем найти максимальный зазор



$$w^T x = p \|w\|_2$$

$$w^T x \geq \alpha$$



$$p \geq \frac{\alpha}{\|w\|_2}$$

$$w^T x \geq \alpha$$

Попробуем найти максимальный зазор

$$w^T x = p \|w\|_2$$

$$w^T x \geq \alpha$$



$$\max_w \frac{\alpha}{\|w\|_2}$$

$$w^T x \geq \alpha$$



$$p \geq \frac{\alpha}{\|w\|_2}$$

$$w^T x \geq \alpha$$

Попробуем найти максимальный зазор

$$\max_w \frac{1}{\|w\|_2}$$

$$w^T x = p \|w\|_2$$
$$w^T x \geq \alpha$$

$$w^T x \geq 1$$

$$\max_w \frac{\alpha}{\|w\|_2}$$

$$w^T x \geq \alpha$$

$$p \geq \frac{\alpha}{\|w\|_2}$$



$$w^T x \geq 1$$

Попробуем найти максимальный зазор

$$\min_w \left\| w \right\|_2$$
$$w^T x \geq 1$$



$$\max_w \frac{1}{\left\| w \right\|_2}$$
$$w^T x \geq 1$$



$$\max_w \frac{\alpha}{\left\| w \right\|_2}$$
$$w^T x \geq \alpha$$

$$w^T x = p \left\| w \right\|_2$$
$$w^T x \geq \alpha$$
$$p \geq \frac{\alpha}{\left\| w \right\|_2}$$
$$w^T x \geq \alpha$$



Попробуем найти максимальный зазор

$$\min_w \|w\|_2$$

$$w^T x \geq 1$$



$$\min_w \frac{1}{2} \|w\|_2$$

$$w^T x \geq 1$$



$$\max_w \frac{1}{\|w\|_2}$$

$$w^T x \geq 1$$



$$\max_w \frac{\alpha}{\|w\|_2}$$

$$w^T x \geq \alpha$$



$$w^T x = p \|w\|_2$$

$$w^T x \geq \alpha$$



$$p \geq \frac{\alpha}{\|w\|_2}$$

$$w^T x \geq \alpha$$

Попробуем найти максимальный зазор

$$\min_w \left\| w \right\|_2$$
$$w^T x \geq 1$$

$$\max_w \frac{1}{\left\| w \right\|_2}$$
$$w^T x \geq 1$$

$$w^T x = p \left\| w \right\|_2$$
$$w^T x \geq \alpha$$

$$\min_w \frac{1}{2} \left\| w \right\|_2$$
$$w^T x \geq 1$$

$$\max_w \frac{\alpha}{\left\| w \right\|_2}$$
$$w^T x \geq \alpha$$

$$p \geq \frac{\alpha}{\left\| w \right\|_2}$$
$$w^T x \geq \alpha$$

оптимизационная задача для одного примера

$$\frac{1}{2} \left| \left| w \right| \right|^2 \rightarrow \min_w$$

$$M_i(w) \geq 1, i=1, L$$

- Исключительно из соображений максимизации зазора

SVM. Неразделимый случай



Штрафы за ошибки

- Разделимая постановка

$$\left. \begin{array}{l} \frac{1}{2} \|w\|^2 \rightarrow \min_w \\ M_i(w) \geq 1, i=1, L \end{array} \right\}$$

- Вводим штрафы за неправильную сторону

$$\left. \begin{array}{l} Q'(w) = \frac{1}{2} \|w\|^2 + C \sum_i^L \varepsilon_i \rightarrow \min \\ M_i(w) \geq 1 - \varepsilon_i, i=1, L \\ \varepsilon_i \geq 0, i=1, L \end{array} \right\}$$

Штрафы за ошибки

- Вводим штрафы за неправильную сторону

$$\left\{ \begin{array}{l} Q'(w) = \frac{1}{2} \|w\|^2 + C \sum_i^L \varepsilon_i \rightarrow \min \\ M_i(w) \geq 1 - \varepsilon_i, i=1, L \\ \varepsilon_i \geq 0, i=1, L \end{array} \right.$$



$$Q'(w) = \frac{1}{2} \|w\|^2 + C \sum_i^L \max(0, 1 - M_i) \rightarrow \min$$

SVM. Безусловная постановка

- $Q'(w) = \frac{1}{2} \|w\|^2 + C \sum_i^L \max(0, 1 - M_i) \rightarrow \min$
- Классический функционал для SVM
- С как копромисс между разделением классов и зазором

SVM. Безусловная постановка

- $Q'(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i^L \max(0, 1 - M_i) \rightarrow \min$
- Классический функционал для SVM
- Обосновали Лосс ! Обосновали регуляризацию !
- И да, его можно решать SGD
 - Но люди пошли дальше
 - И получили профит



SVM. Двойственная задача



SVM: переход к двойственной задаче

- Вернемся к условной постановке

$$Q'(w) = \frac{1}{2} \|w\|^2 + C \sum_i^L \varepsilon_i \rightarrow \min$$

$$M_i(w) \geq 1 - \varepsilon_i, i=1, L$$

$$\varepsilon_i \geq 0, i=1, L$$

Условия Куна-Таккера (ККТ)

Задача математического программирования:

$$\begin{cases} f(x) \rightarrow \min_x; \\ g_i(x) \leq 0, \quad i = 1, \dots, m; \\ h_j(x) = 0, \quad j = 1, \dots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители $\mu_i, i = 1, \dots, m, \lambda_j, j = 1, \dots, k$:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, \quad \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \lambda_j h_j(x); \\ g_i(x) \leq 0; \quad h_j(x) = 0; \quad (\text{исходные ограничения}) \\ \mu_i \geq 0; \quad (\text{двойственные ограничения}) \\ \mu_i g_i(x) = 0; \quad (\text{условие дополняющей нежёсткости}) \end{cases}$$

Двойственная задача

$$\begin{cases} -\mathcal{L}(\lambda) = -\sum_{i=1}^{\ell} \lambda_i + \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \rightarrow \min_{\lambda}; \\ 0 \leq \lambda_i \leq C, \quad i = 1, \dots, \ell; \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases}$$

Решение прямой задачи выражается через решение двойственной:

$$\begin{cases} w = \sum_{i=1}^{\ell} \lambda_i y_i \mathbf{x}_i; \\ w_0 = \langle w, \mathbf{x}_i \rangle - y_i, \quad \text{для любого } i: \lambda_i > 0, \ M_i = 1. \end{cases}$$

Линейный классификатор:

$$a(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^{\ell} \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle - w_0 \right).$$

Профиты перехода

$$L(\lambda) = - \sum_i^L \lambda_i + \frac{1}{2} \sum_i^L \sum_j^L \lambda_i \lambda_j y_i y_j (x_i, x_j) \rightarrow \min$$

$$a(x) = sign \left(\sum_i^L \lambda_j y_i (x_i, x) - w_0 \right)$$

- В обучении не участвуют сами объекты, только матрица Грамма на выборке
 - На инференсе сами объекты также не нужны
- В переходе использовались свойства скалярного произведения, но без привязки к пространству
 - Подойдет ск.п. в любом пространстве = путь в нелинейность

Нелинейная классификация. SVM ядра



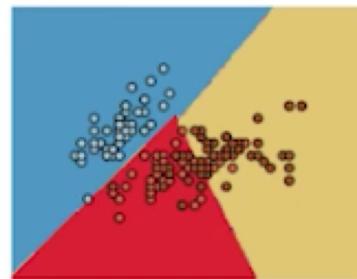
- Переходим к пространству более высоких размерностей – за счет ядер
- Определение – функция $k: X \times X \rightarrow R$ – ядро, если существует ψ :
 - $\psi: X \rightarrow H$
 - $K(x, x') = (\psi(x), \psi(x'))$
- Симметричная, неотрицательно определенная функция

- Линейное
 - $k(x, z) = (x, z)$
- Полиномиальное
 - $k(x, z) = ((x, z) + 1)^d$
- Radial Basis Function
 - $k(x, z) = \exp(-\gamma|x - z|^2)$
- Остальные – через конструктивные методы

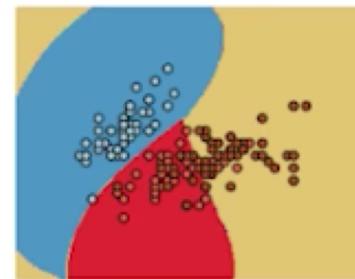
- Линейное
 - $k(x, z) = (x, z)$
- Полиномиальное
 - $k(x, z) = ((x, z) + 1)^d$
- Radial Basis Function
 - $k(x, z) = \exp(-\gamma|x - z|^2)$
- Остальные – через конструктивные методы
- *Пространство для квадратичного ядра (x, z) ?*

Примеры ядер

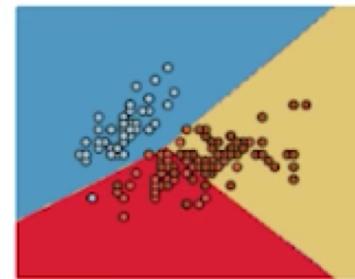
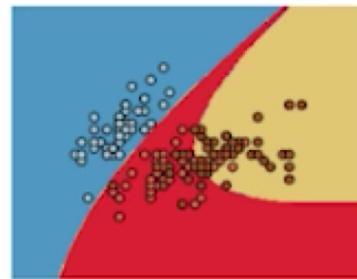
SVC with linear kernel



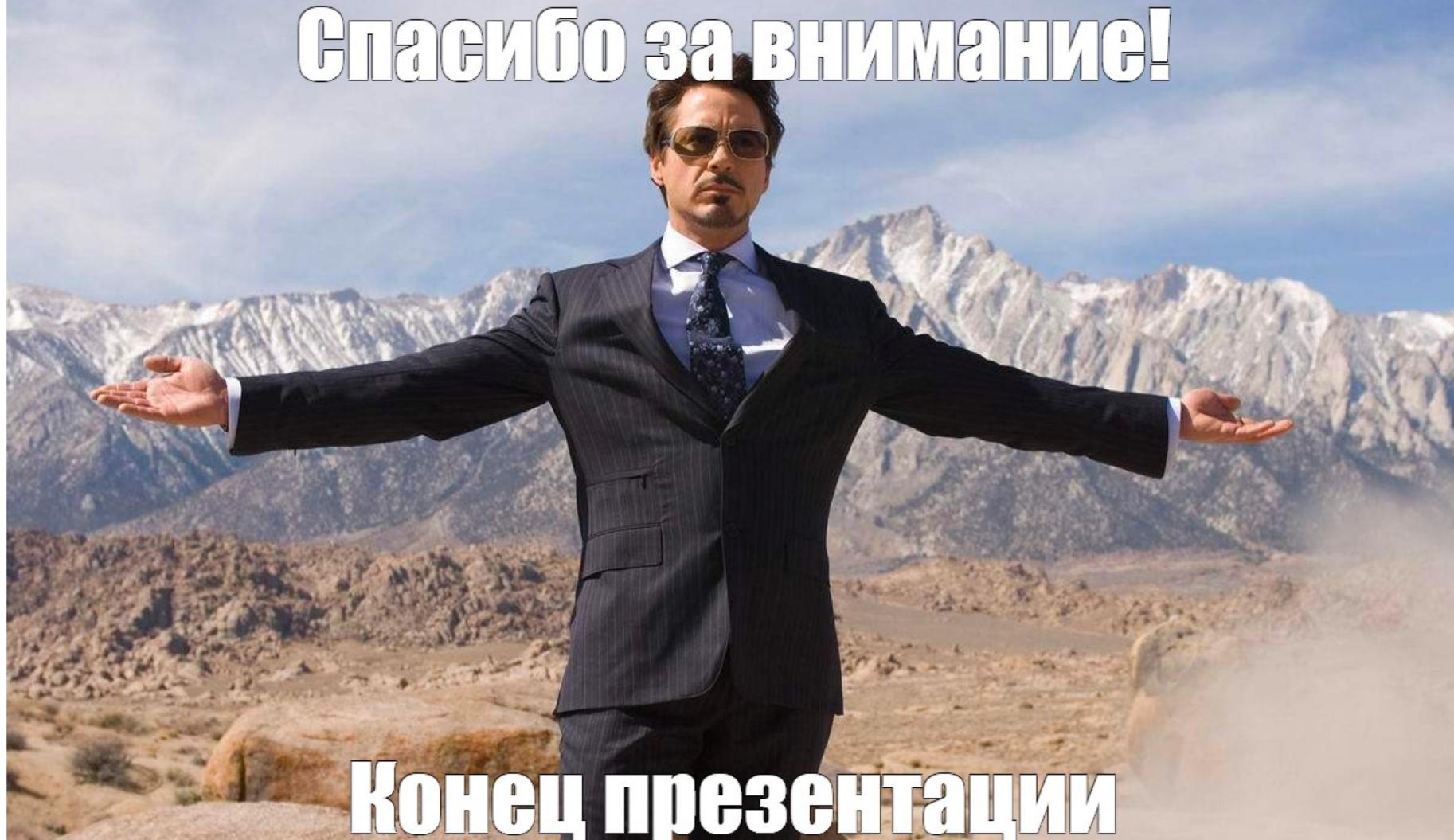
SVC with RBF kernel



SVC with polynomial (degree 3) kernel LinearSVC (linear kernel)



Спасибо за внимание!



Конец презентации