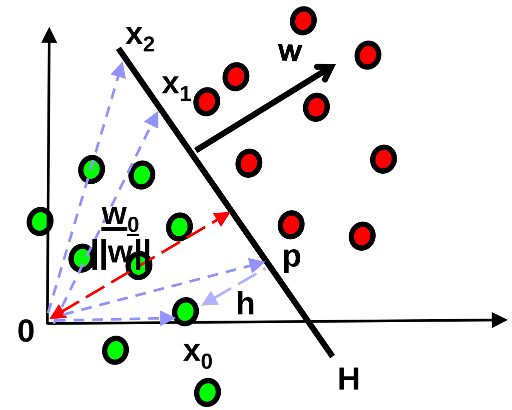
Лекция 11:
Метод опорных векторов (начало)

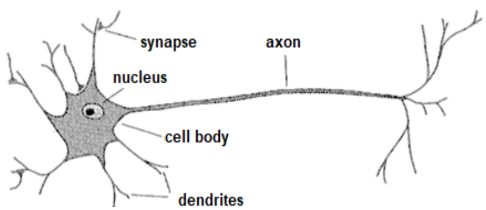
Линейный бинарный классификатор на основе разделяющей гиперплоскости

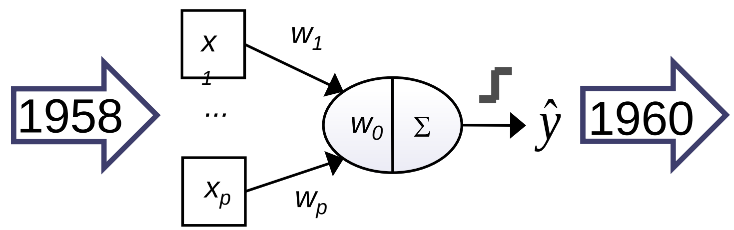


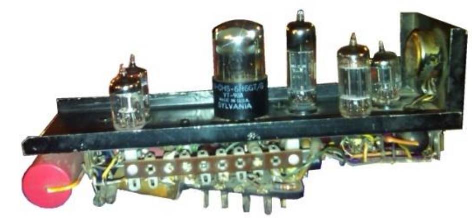
- Основные определения и свойства:
  - Отклик $Y = \{-1, +1\}$
  - Дискр. ф-ция $g(x) = g_+(x) - g_-(x)$
  - Отступ $M(x, y) = yg(x)$
  - Линейность $g(x) = \langle w, x \rangle + w_0$
  - Граница – гиперплоскость $H = \{x | \langle w, x \rangle + w_0 = 0\}$ определяется нормалью $w/||w||$ и смещением $w_0\ /||w||$.
  - То, что w – ортогонально H, доказывается из определения линейного $g(x)$ и равенства $g(x) = 0$ на границе: пусть $x_1, x_2 \in H \Rightarrow \langle w, x_1 \rangle + w_0 = 0$, $\langle w, x_2 \rangle + w_0 = 0 \Rightarrow \langle w, x_2 - x_1 \rangle = 0 \Rightarrow w$ ортогонально любым $x_1, x_2 \in H$
  - Знаковое расстояние от точки x_0 до границы $d(x_0, H) = g(x_0)/||w||$ (подстановка в нормализованное уравнение гиперплоскости).

    Пусть $x_0 = p + h$, $p \in H$ – проекция x_0 на H, h –ортогональное дополнение, тогда $h = d\frac{w}{||w||}$, $x_0 = p + d\frac{w}{||w||}$ домножаем скалярно на w прибавляем w_0, получаем: $\langle w, x_0 \rangle + w_0 = \cancel{\langle w, p \rangle + w_0} + d\frac{\langle w, w \rangle}{||w||} \Rightarrow d = \frac{\langle w, x_0 \rangle + w_0}{||w||}$
  - Прогноз $a(x) = sign(g(x))$ – с какой стороны от H, расстояния от центра координат до H равно $w_0/||w||$

Персептрон Розенблатта







- Модель - разделяющая гиперплоскость:
  - Функция потерь $L_{perc}(M) = -[M]_+$,
  - Обучение - SGD, доказана сходимость за конечное число шагов
  - Для «ошибок» (примеров не с той стороны гиперплоскости):

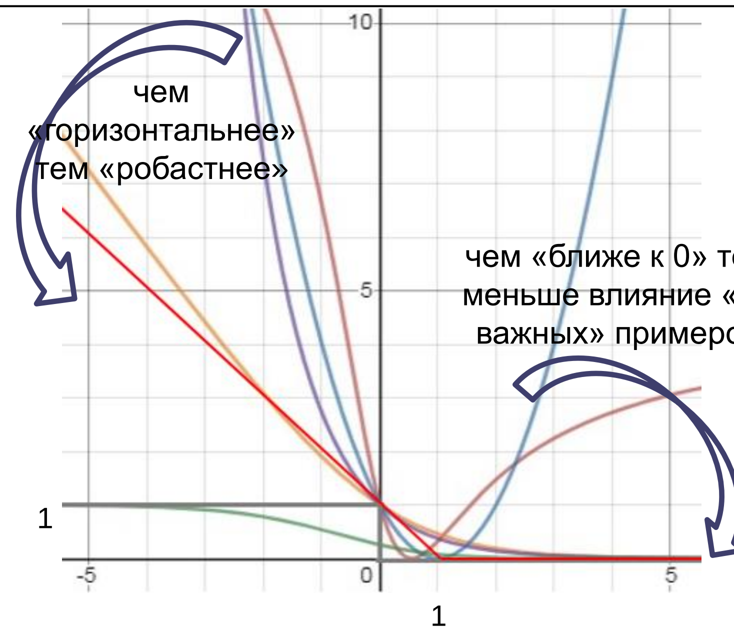
$$\begin{pmatrix} w^{(t)} \\ w_0^{(t)} \end{pmatrix} + \eta \begin{pmatrix} y_i x_i \\ y_i \end{pmatrix} \to \begin{pmatrix} w^{(t+1)} \\ w_0^{(t+1)} \end{pmatrix}$$

- Недостатки (их устранение - достоинства SVM):
  - Несколько возможных решений при линейной разделимости классов (зависит от начального приближения)
  - Не сходится при линейной неразделимости классов, а при линейной разделимости долго сходится (много шагов)

Обучение линейного классификатора

- «Пороговая» (персептрон) функция потерь $L_{perc}(M) = -[M]_+$
  - кусочно-постоянная ⇒ имеет нулевые градиенты
- Можно ограничить ее сверху другой гладкой функцией потерь и искать решение задачи оптимизации с регуляризацией:
  - Логистическая:
    $L_{log}(M) = \log_2(1 + e^{-M})$
  - Квадратичная:
    $L_{sq}(M) = (1 - M)^2$
  - Экспоненциальная:
    $L_{exp}(M) = e^{-M}$
  - Тангесовая:
    $L_{tng}(M) = (2\arctan(M) - 1)^2$
  - Hinge («шарнир»):
    $L_{hinge}(M) = -[1 - M]_+$

$$\min_w \frac{1}{l}\sum_i L_*(y_i(\langle x_i, w\rangle + w_0)) + \boldsymbol{\gamma L_p(w)}$$



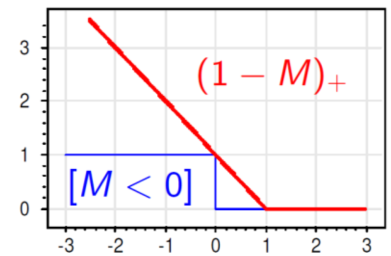
Аппроксимация Hinge функцией потерь с L_2 регуляризацией

- Ограничим сверху эмпирический риск персептрона L_2 - регуляризованным эмпирическим риском с с Hinge функцией потерь:

$$Q_{perc}(w, w_0) = \sum_{i=1}^{l} [M_i(w, w_0) < 0] \le$$

$$\le Q_{hinge}(w, w_0) = \sum_{i=1}^{l} \big(1 - M_i(w, w_0)\big)_+ + \gamma ||w||^2$$

$$Q_{hinge} \to \min_{w, w_0} \quad \Rightarrow Q_{perc} \to \min_{w, w_0}$$



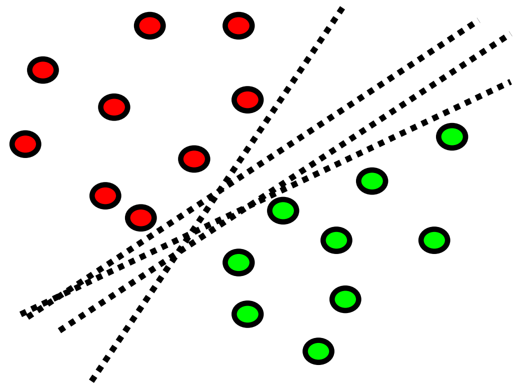
- Первое слагаемое:
  - линейно штрафует за приближение к границе классов с «правильной стороны» ближе чем 1
  - линейно штрафует за удаление от границы с «неправильной стороны»
- Второе слагаемое:
  - штрафует за сложность, не давая переобучаться
  - контролирует стабильность при мультколлинеарности

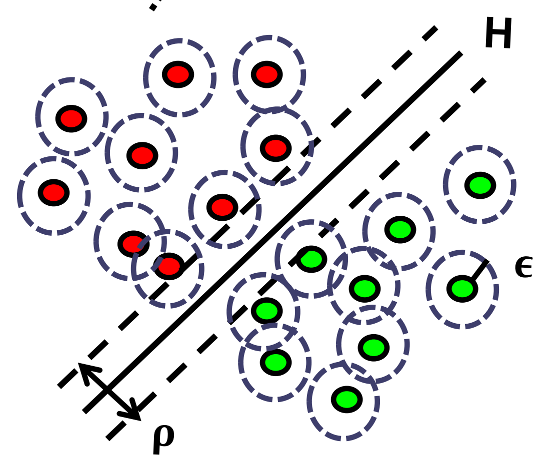
Оптимальная разделяющая гиперплоскость в случае линейно разделимых классов

- В случае линейно разделимости классов:
  - можно провести бесконечно много разделяющих гиперплоскостей.
  - **Какая из них лучше?**
- Определим ширину разделяющей полосы - зазор (марджин) для множества точек как минимум по всем:

$$\rho = \min_{1 \le i \le l} M(x_i, y_i) = \min_{1 \le i \le l} y_i g(x_i)$$

- Т.к. есть случайная составляющая (шум):
  - наблюдения могут лежать в некоторой окрестности неизвестного радиуса ϵ
  - значит чем больше отступ ρ, тем меньше вероятность, что окрестность точек рядом с границей пересечет ее
- Вывод – нужно **максимизировать зазор**





Максимизация отступа в случае линейно разделимых классов

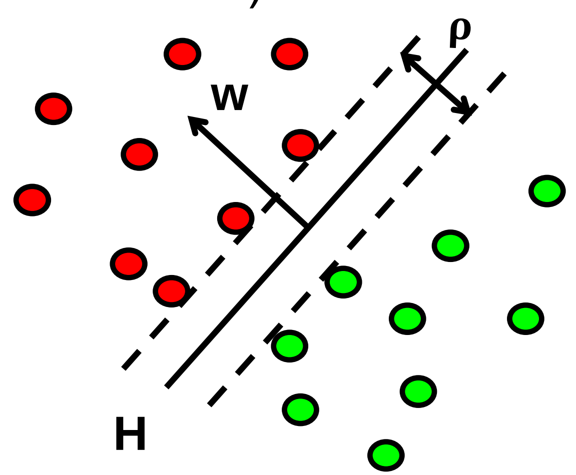
- Каноническое уравнение гиперплоскости:
  - уравнение H определено с точностью до множителя, надо зафиксировать (с точностью до знака)
  - нормируем параметры так, чтобы расстояние $d(x, H) = g(x)/\left|\left|w\right|\right|$ от границы до ближайшего наблюдения каждого класса было равно 1
  - Это приводит к условиям: если $y_i = 1 \Rightarrow \langle w, x_i \rangle + w_0 \geq 1$, а для $y_i = -1 \Rightarrow \langle w, x_i \rangle + w_0 \leq -1$ и в общем виде $\forall \boldsymbol{i}: \boldsymbol{y_i}(\langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + \boldsymbol{w_0}) \geq \mathbf{1}$
- Ширина разделяющей полосы (зазора между классами):

$$\boldsymbol{\rho} = \frac{\mathbf{2}}{||\boldsymbol{w}||} \to \max_{w}$$

- Получаем задачу условной оптимизации:

$$\begin{cases} \min\limits_{w} \frac{1}{2} ||w||^2 \\ \forall i: y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

- Все выпуклое - **единственное решение!**



Решение в случае линейно разделимых классов

- Выпишем лагранжиан:

$$L(w, w_0; \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i [y_i(\langle w, x_i \rangle + w_0) - 1]$$

  - с множителями Лагранжа $\alpha_i \geq 0$ для каждого ограничения
  - с условиями дополняющей нежёсткости (ККТ):

$$\forall i: \alpha_i [y_i(\langle w, x_i \rangle + w_0) - 1] = 0$$

- Из необходимых условий оптимальности следует:

$$\frac{\partial L(w, w_0; \alpha)}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0, \ \frac{\partial L(w, w_0; \alpha)}{\partial w} = 0 \Rightarrow \mathrm{w} = \sum_{i=1}^{l} \alpha_i y_i x_i$$

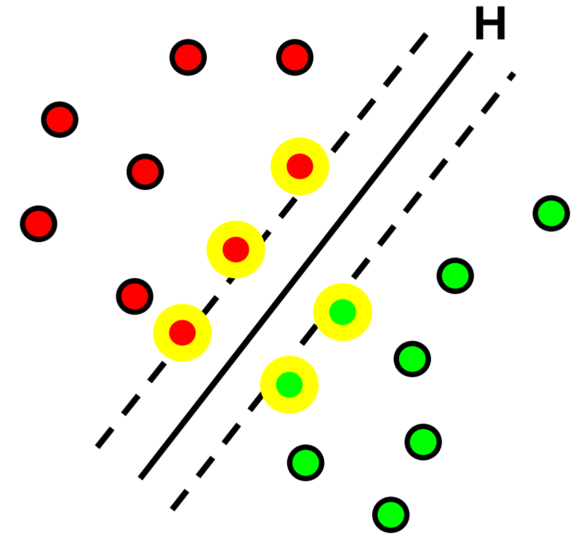
- Дискриминантная функция:

$$g(x) = \langle w, x \rangle + w_0 = \sum_{i=1}^{l} \boldsymbol{\alpha_i y_i \langle x_i, x \rangle + w_0}$$

- Сдвиг w_0 может корректироваться «вручную», обычно инициализируется как: $w_0 = \frac{1}{l} \sum_{j=1}^{l} \left(\boldsymbol{y_j} - \sum_{i=1}^{l} \boldsymbol{\alpha_i y_i \langle x_i, x \rangle} \right)$

Опорные вектора в случае линейно разделимых классов

- По свойствам множителей Лагранжа: $y_i(\langle w, x_i \rangle + w_0) > 1 \Rightarrow \alpha_i = 0$:
  - $\alpha_i \neq 0$ для **опорных векторов** (наблюдения лежат строго на границе, их расстояние до H равно 1)
  - Дискриминантная функция (и модель) зависит **только от опорных векторов**: $a(x) = \mathrm{sign}(\sum_{i \in SV} \alpha_i y_i \langle x_i, x \rangle + w_0)$
  - Результат обучения не зависит от наличия в тренировочном наборе наблюдений, не лежащих на границе, их можно исключить из выборки и получить ту же модель SVM (вот только мы заранее не знаем, какие именно наблюдения лежат на границе)
  - Этим свойством пользуются алгоритмы оптимизации для SVM



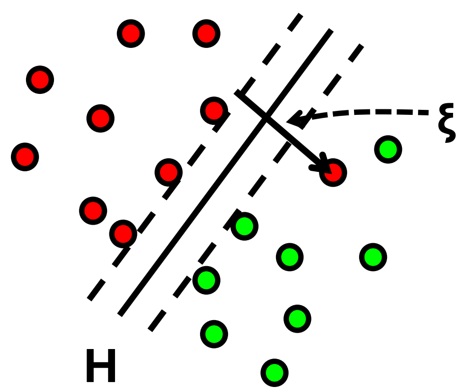
Линейно неразделимые классы

- Классы не обязаны быть линейно разделимы:
  - можно попробовать перебрать оптимальные гиперплоскости, минимизируя число ошибок, но оказалось, что это NP-трудная задача (не найдено не экспоненциальных по сложности методов)
- Основной подход – дополнительно линейно штрафовать модель за «нарушение» неравенств канонической гиперплоскости:

обобщающая способность

ошибка

$$\begin{cases} \min\limits_{w,\xi,w_0} \frac{1}{2}||w||^2 + \frac{C}{l}\sum_{i=1}^{l}\xi_i \\ \forall i: y_i(\langle w, x_i\rangle + w_0) \ \geq \ 1 - \xi_i, \xi_i \geq 0 \end{cases}$$



  - параметр *C* – задает в явном виде компромисс между точностью и сложностью модели
  - Аналогично безусловной минимизации Hinge функции потерь с L_2 регуляризацией:

$$Q_{hinge}(w, w_0) = \sum_{i=1}^{l}\big(1 - M_i(w, w_0)\big)_+ + \gamma||w||^2 \to \min_{w,w_0}$$

Метод множителей Лагранжа для линейно неразделимых классов

- Снова выпишем лагранжиан:

$$L(w, w_0, \xi; \alpha, \eta) = \frac{1}{2}\left||w|\right|^2 - \sum_{i=1}^{l} \alpha_i [y_i(\langle w, x_i\rangle + w_0) - 1] - \sum_{\boldsymbol{i=1}}^{\boldsymbol{l}} \boldsymbol{\xi_i(\alpha_i + \eta_i - C)}$$

  - α_i - двойственные переменные к условиям $y_i(\langle w, x_i\rangle + w_0) \ \geq \ 1 - \boldsymbol{\xi_i}$
  - η_i - двойственные переменные к условиям $\boldsymbol{\xi_i \geq 0}$
  - условия дополняющей нежёсткости ККТ:

$$\forall i\colon \alpha_i[y_i(\langle w, x_i\rangle + w_0) - (1 - \boldsymbol{\xi_i})] = 0, \boldsymbol{\eta_i \xi_i = 0} \qquad (*)$$

- Из необходимых условий седловой точки функции Лагранжа:

$$\frac{\partial L(w,w_0,\alpha,\eta)}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0, \frac{\partial L(w,w_0,\alpha,\eta)}{\partial w} = 0 \Rightarrow \mathrm{w} = \sum_{i=1}^{l} \alpha_i y_i x_i,$$

$$\boldsymbol{\frac{\partial L(w,w_0,\alpha,\eta)}{\partial \xi} = 0 \Rightarrow \eta_i + \alpha_i = C} \qquad (**)$$

- Дискриминантная функция и сдвиг те же, но опорные вектора другие: $g(x) = \sum_{i=1}^{l} \alpha_i y_i \langle x_i, x\rangle + w_0, w_0 = \frac{1}{l}\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{l} \alpha_i y_i \langle x_i, x\rangle\right)$

Опорные вектора для линейно неразделимых классов

- Получаем два типа опорных векторов:
  - **Ошибки** – неравенство со штрафом строго НЕ выполняется:

    $$\boldsymbol{\alpha_i} = \boldsymbol{C}, \eta_i = 0, \xi_i > 0, y_i(\langle w, x_i \rangle + w_0) > 1$$

  - **Граничные** – неравенство выполняется как равенство:

    $$\mathbf{0} < \boldsymbol{\alpha_i} < \boldsymbol{C}, 0 < \eta_i < C, \boldsymbol{\xi_i} = \mathbf{0}, y_i(\langle w, x_i \rangle + w_0) = 1$$

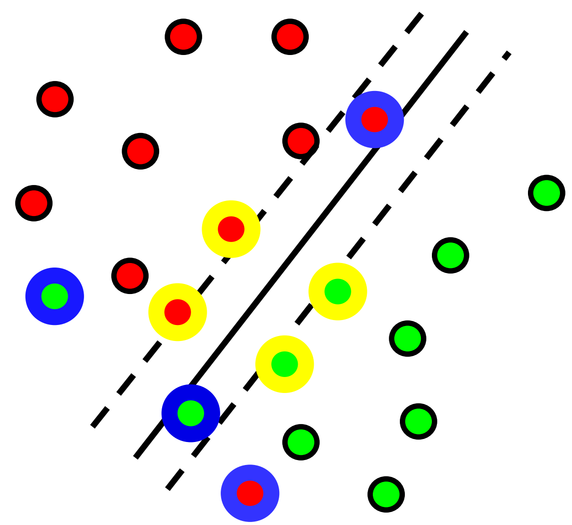
- Остальные (не важные) наблюдения:
  - **Периферийные** - неравенство со штрафом выполняется:

    $$\alpha_i = 0, \eta_i = \mathrm{C}, \xi_i = 0, y_i(\langle w, x_i \rangle + w_0) < 1$$

  - снова от них ничего не зависит

Граничные опорные вектора

опорные вектора - ошибки



Двойственная задача

- Можно решать прямую задачу (есть для этого методы оптимизации), но оказалось, что удобнее решать двойственную
- Подставим равенства, полученные из условий $(*)$ и $(**)$ в $L(w, w_0, \xi; \alpha, \eta)$ и увидим, что Лагранжиан после всех сокращений зависит только от двойственных переменных α_i и имеет простую квадратичную форму:

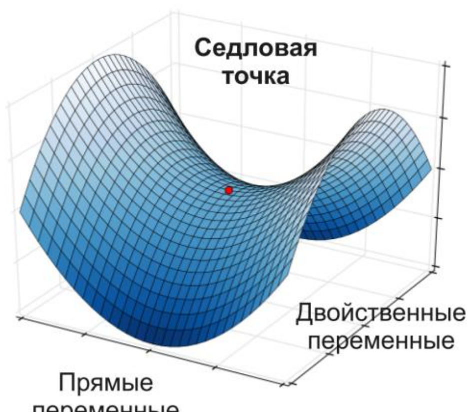
$$L(\cancel{w, w_0, \xi}; \alpha, \cancel{\eta}) = W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_j, x_i \rangle$$

- пользуясь свойством седловой точки Лагранжа:

$$L(w^*, w_0^*, \xi^*; \alpha^*, \eta^*) = \min_{w, w_0, \xi} L(w, w_0, \xi; \alpha^*, \eta^*) = \max_{\alpha, \eta} L(w^*, w_0^*, \xi^*; \alpha, \eta)$$

- перейдем к решению двойственной задачи:

$$\begin{cases} \max\limits_{\alpha} W(\alpha) \\ 0 \le \alpha_i \le C, \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases}$$

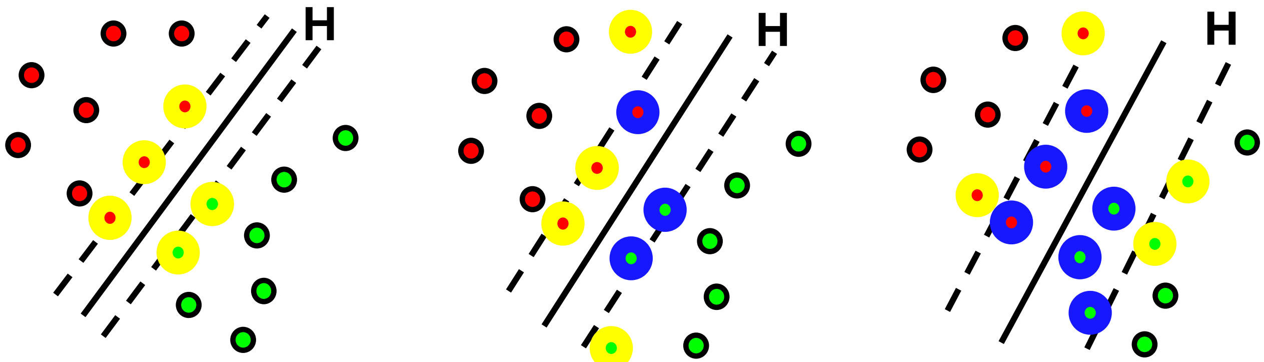


- решение прямой задачи выражается через него как:

$$a(x) = \operatorname{sign}\left(\sum_{i \in SV} \alpha_i y_i \langle x_i, x \rangle + w_0\right)$$

Выбор параметра штрафа C

- Аналогично параметру регуляризации (но наоборот):
  - чем больше *C* тем меньше смещение и больше дисперсия модели
  - чем меньше *C* тем больше обобщающая способность и ошибка подгонки модели ($C_{left} > C_{middle} > C_{right}$)



- На практике:
  - используют стандартные эвристики: C={0.1, 1, 10}
  - подбирают с помощью кросс-валидации (по сетке значений)
- **Не интуитивный** параметр
  - Тяжело: угадать точно, выбрать сетку для перебора, понять смысл

Nu-SVM

- Основная идея – напрямую максимизировать зазор (ширину разделяющей полосы) между классами $\boldsymbol{\rho}$

$$\min_{\xi,\rho,w,w_0} \frac{1}{2}\|w\|^2 + \frac{1}{l}\sum_{i=1}^{N} \xi_i - \boldsymbol{\rho\nu}$$

$$\forall i: (y_i(\langle x_i, w\rangle + w_0) \le \boldsymbol{\rho} - \boldsymbol{\xi_i}, \boldsymbol{\rho} \ge \mathbf{0}, \xi_i \ge 0$$

- Также через преобразование Лагранжиана сводится к задаче квадратичного программирования в двойственных переменных:

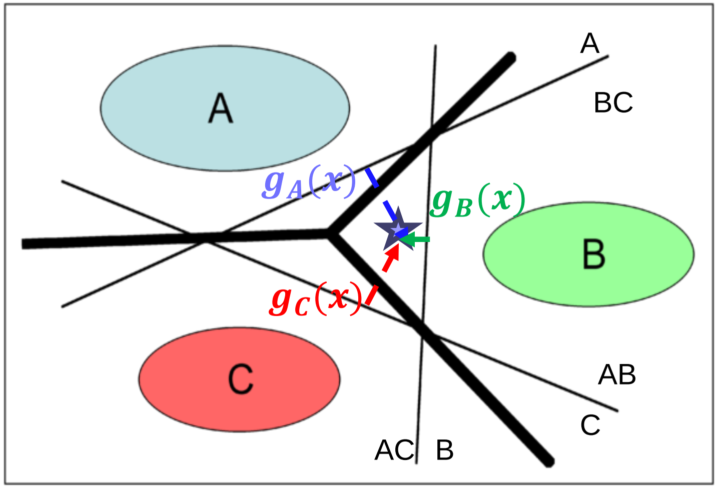
$$\begin{cases} \max\limits_{\alpha} -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j y_i y_j \langle x_j, x_i\rangle \\ \mathbf{0} \le \boldsymbol{\alpha_i} \le \frac{\mathbf{1}}{\boldsymbol{l}}, \sum_{i=1}^{l} \alpha_i y_i = 0, \sum_{\boldsymbol{i=1}}^{\boldsymbol{l}} \boldsymbol{\alpha_i} \ge \boldsymbol{\nu} \end{cases}$$

- Вместо метапараметра C используется ν с важными «ν-свойствами»:
  - ν-верхняя граница пропорции опорных векторов – ошибок
  - ν-нижняя граница пропорции опорных векторов - граничных
  - асимптотически с вероятностью 1 (при определенных условиях) эти границы достигаются

Многоклассовый SVM «каждый против всех»

- Каждый против всех:
  - Строим k моделей (k-число классов), выбираем класс с наиболее уверенным прогнозом – наибольшей дискриминантной функцией:

$$\arg\max_{j=1,\dots,k} g_j(x)$$



- Особенности:
  - Гарантировано есть хотя бы один несбаласированный набор (т.к. 1 класс против всех остальных)
  - Вычислительно сложно при больших наборах данных – *k* бинарных задач с *l* наблюдениями в каждой
  - независимое обучение – независимые $g_j(x)$, надо приводить на близкие шкалы, можно с помощью $\mathrm{softmax}(g_1(x), \dots, g_k(x))$ или более корректно с помощью корректировки Платта

Корректировка Платта

- Преобразуем отклик SVM из $(-\infty, +\infty)$ в **вероятностный** диапазон $[0,1]$с помощью подгонки сигмоиды:

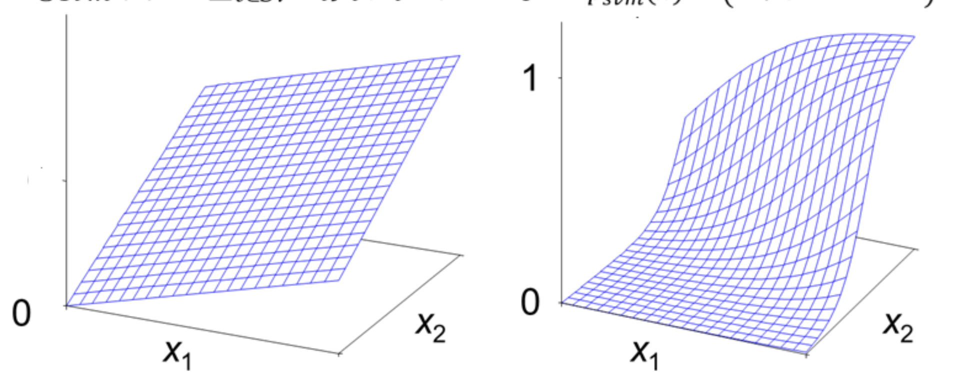
$$p_{svm}(x) = \frac{1}{1+\exp(Ag_{svm}(x)+B)}$$

где $g_{svm}(x) = \sum_{i \in SV} \alpha_i y_i \langle x_i, x \rangle + w_0)$, а A и B – параметры

- Чтобы не переобучиться:
  - параметры A и B подбираются как в логистической регрессии с одним предиктором (откликом SVM) на валидационной выборке (не использовалась для обучения SVM) или с помощью кросс-валидации
  - дополнительно часто используется «регуляризация» откликов:

$$y_i^{Platt} = \begin{cases} \dfrac{l_+ + 1}{l_+ + 2}, y_i = 1 \\ \dfrac{1}{l + 2}, y_i = -1 \end{cases}$$

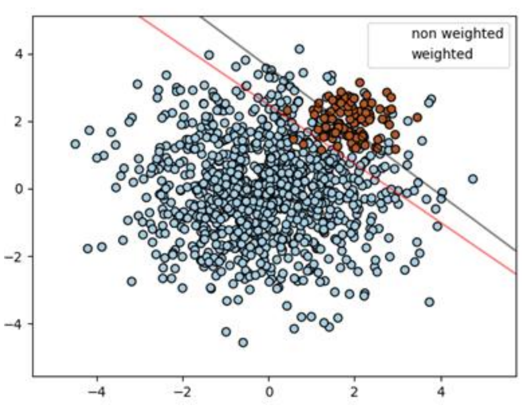
$g_{svm}(x) = \sum_{i \in SV} \alpha_i y_i \langle x_i, x \rangle + w_0$ $\quad$ $p_{svm}(x) = \left(1 + e^{Ag_{svm}(x)+B}\right)^{-1}$



Дисбаланс классов

- Возможные подходы к решению проблемы:
  - В целом SVM менее чувствителен к дисбалансу, чем другие методы, т.к. модель зависит только от опорных векторов
  - SMOTE (oversampling)
  - Undersampling + корректировка сдвига w_0– строим SVM на сбалансированной выборке, а w_0 выбираем с учетом дисбаланса, например $w_0^* = \underset{w_0}{\operatorname{argmin}} F_\beta\left(g(x, w_0), y\right)$
  - Undersampling + корректировка Платта – строим SVM на сбалансированной выборке, а корректировку Платта на несбалансированной
  - Используем веса наблюдений
  - Используем веса классов:

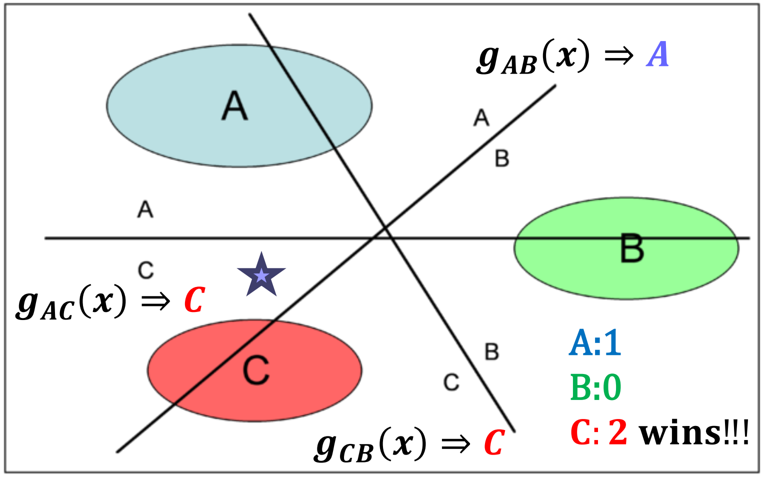
$$\begin{cases} \min\limits_{w,\xi,w_0} \frac{1}{2}\left|\left|w\right|\right|^2 + \frac{C_{-1}}{l_{-1}} \sum\limits_{i:y_i=-1} \xi_i + \frac{C_1}{l_1} \sum\limits_{i:y_i=+1} \xi_i \\ \forall i: y_i(\langle w, x_i\rangle + w_0) \ \geq \ 1 - \xi_i, \xi_i \geq 0 \end{cases}$$



Многоклассовый SVM«каждый против каждого»

- Каждый против каждого
  - Строим k(k-1)/2 моделей (k-число классов), выбираем класс голосованием:

$$\arg \max_{j=1,\dots,k} \sum_{i \neq j} [g_{ij}(x)]_+$$



- Особенности:
  - меньше проблем с дисбалансом классов чем в каждом против всех
  - вычислительно сложно при больших *k*, получаем k(k-1)/2 бинарных задач, правда наблюдений в каждой меньше *l*
  - независимое обучение – независимые $g_{ij}(x)$ не так критично как в каждом против всех (не сравниваем отклики разных моделей друг с другом напрямую)
  - могут быть «ничьи», простое голосование – не лучший подход, надо учитывать «уверенность» в прогнозе, а значит тоже корректировать отклики

Вероятности классов на основе попарных сравнений

- Если по результатам применения подхода «каждый против каждого» необходимо вычислить вероятности принадлежности наблюдения x_0 каждому из k классов $p_1(x_0), \dots, p_k(x_0)$, то можно воспользоваться подходом попарных сравнений:
  - Применяем все *k(k-1)/2* попарных моделей и получаем для каждой пары классов (i, j) значение дискриминантной функции $g_{ij}(x_0)$
  - Делаем корректировку Платта $p_{ij}(x_0)$ (x_0 можно не указывать, т.к. все считается только для него)
  - Принимаем предположение модели Брэдли-Терри для попарных сравнений: $p_{ij} = p_i/(p_i + p_j)$, где p_s неизвестны для $1 \le s \le k$
  - Находим их, минимизируя численным методом дивергенцию Кульбака-Лейблера :

$$\sum_{i,j} p_{ij} \log\left(\frac{p_{ij}(p_i + p_j)}{p_i}\right) + \sum_{i,j} \frac{p_i}{p_i + p_j} \log\left(\frac{p_i}{p_{ij}(p_i + p_j)}\right) \to \min_{p_1,\dots,p_k}$$

Многоклассовый ECOC SVM

- ECOC:
  - Строим кодовую матрицу M с m новыми «суперклассами», каждый объединяет комбинацию исходных классов и
  - Обучаем m моделей $g_1(x), \dots, g_m(x)$
  - При классификации получаем вектор прогнозов и выбираем класс с наиболее близким кодовым словом:

$$\min_{j=1,\dots,k} dist((g_1(x), \dots, g_m(x), M[j])$$

Выходной вектор прогнозов

$\bar{a}(x) = (-1, 1, \dots, -1)$

Ближайший код

	1	2	...	s	...	l
1	+1	+1	...	-1	...	0
2	0	-1	...	0		+1
...	...	...	...	...	...	...
j	0	-1	...	-1	...	-1
...	...	...	...	...	...	...
k	-1	0	...	+1	...	-1

$y = j$

Предсказанный класс

- Особенности:
  - вычислительную сложность можно контролировать числом столбцов
  - можно рассчитывать «уверенность» в прогнозе на основе расстояний до кодовых слов или по модели Брэдли-Терри
  - но качество зависит от M – если не угадали, то начинаем все заново

SVM с многоклассовой целевой функцией

- Постановка задачи:
  - пусть k – число классов
  - вводим k гиперплоскостей и отдельно штрафуем за нарушение каждой границы и отдельно штрафуем каждую за ее сложность (максимизируем ширину каждой разделяющей полосы)
  - все штрафы суммируем в целевой функции:

$$\begin{cases} \min\limits_{w,w_0,\xi} \frac{1}{2}\sum\limits_{j=1}^{k} \left\|w^j\right\|^2 + \frac{C}{l}\sum\limits_{i=1}^{l}\sum\limits_{j\neq y_i} \xi_{ij} \\ \forall i,j: y_i\left(\langle w^{y_i}, x_i\rangle + w_0^{y_i}\right) \ \geq \ \langle w^j, x_i\rangle + w_0^j + 2 - \xi_{ij}, \xi_{ij} \geq 0 \end{cases}$$

- Особенности:
  - менее гибкие настройки по сравнению с остальными методам
  - вычислительно сложно – много двойственных переменных при большом наборе, проблема дисбаланса тоже есть
  - зато дискриминантные функции подгоняются вместе – прогнозы зависимы, не нужны корректировки