Курс “Data Science”

Бонусное занятие по SVM

SVM Bonus

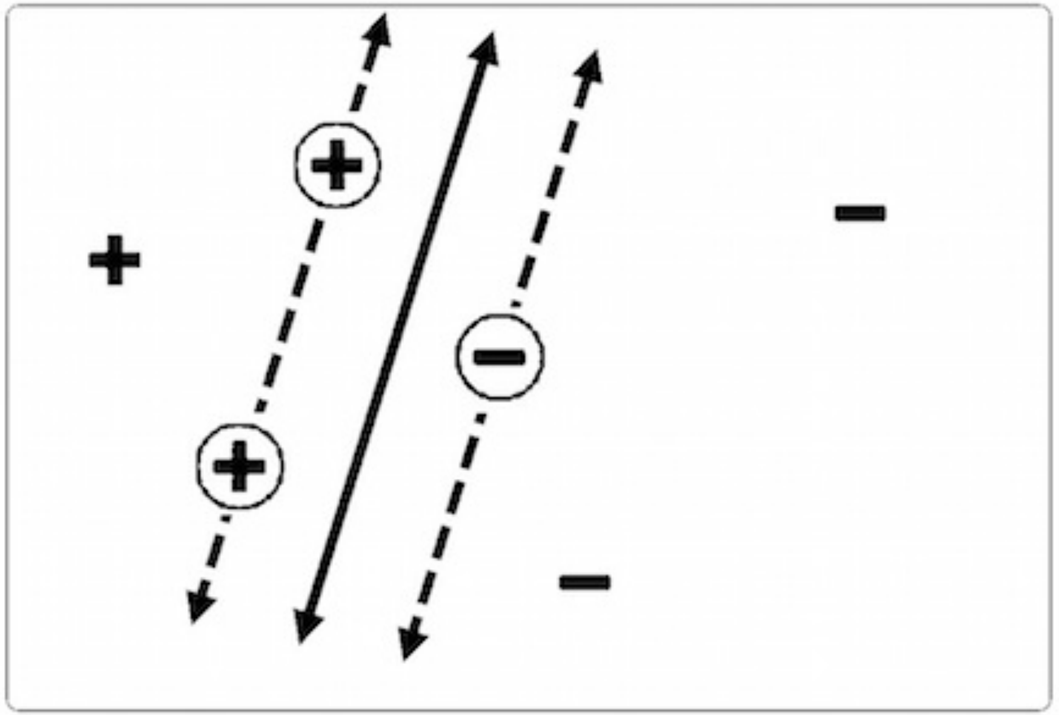
План

- Определение границы, зазора и опорные вектора
- Функция потерь в SVM
- Слабые переменные
- Двойственная задача
- SVM с ядром

Введение.
Определение границы, зазора и опорные вектора

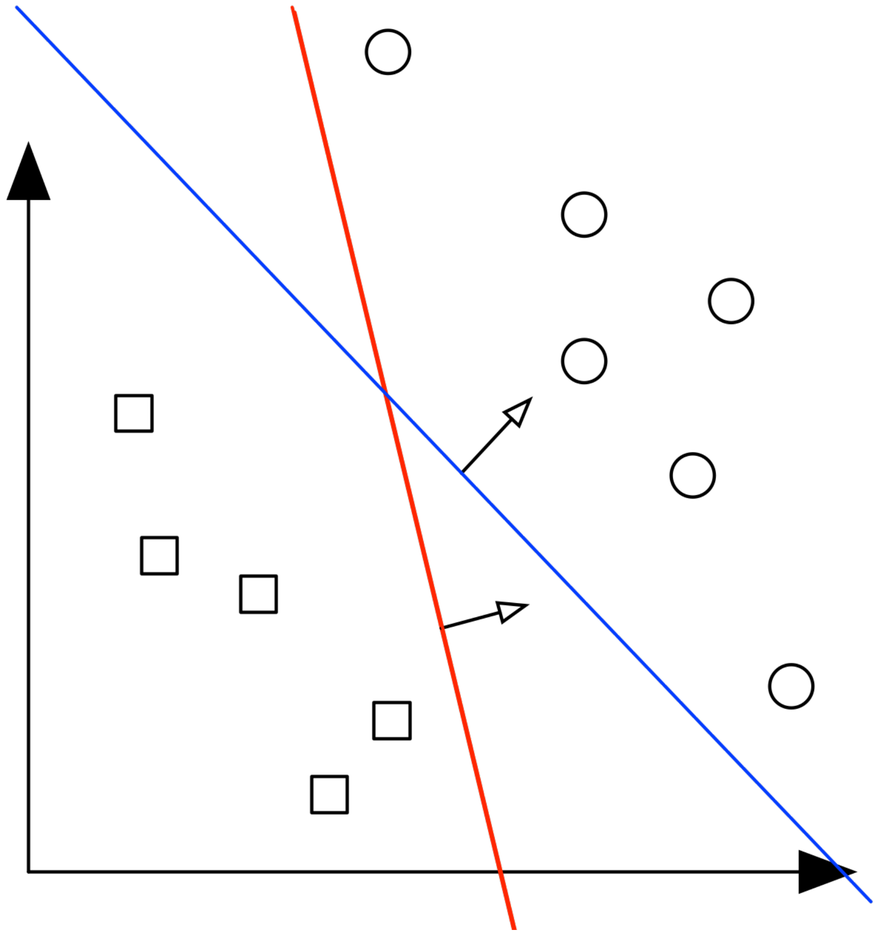
Введение

- **Машины опорных векторов (SVM)**
  - мощная и универсальная техника машинного обучения, способная выполнять линейную или нелинейную классификацию, регрессию.
- **Обучающее множество:**
  - $\{\mathbf{x}^{(i)}, y^{(i)}\}, i = 1,N$
  - $\mathbf{x}^{(i)}$ - i-й обучающий пример
  - $y^{(i)} \in \{-1,1\}$ - метка класса для i-го обучающего примера



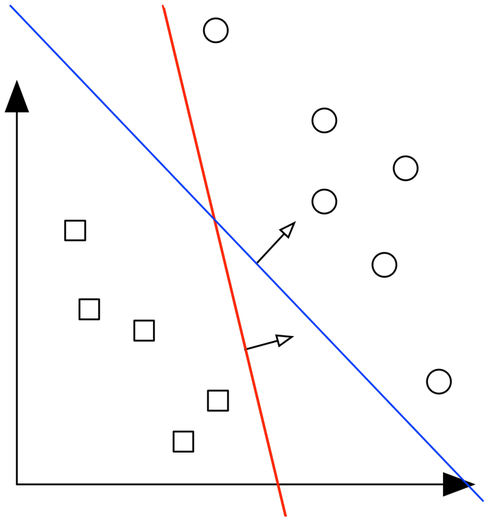
Классификатор как гиперплоскость (граница)

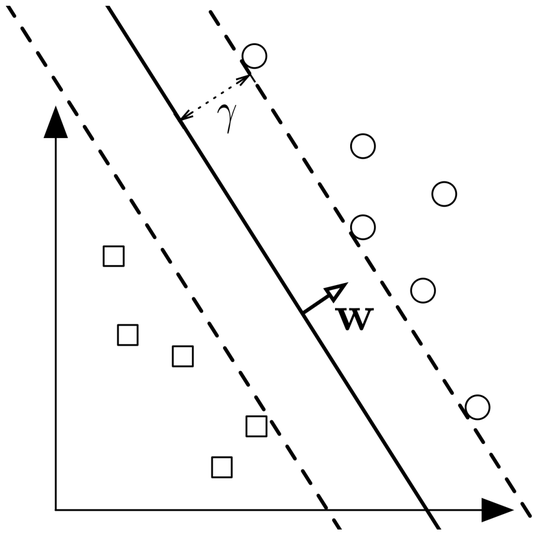
- **разделяющая гиперплоскость определяется уравнением**
  - $b + w_1 x_1 + \ldots + w_N x_N = 0$
  - или в векторном виде:
  - $b + \mathbf{w} \cdot \mathbf{x} = 0$
  - где
  - b - число
  - $\mathbf{w}$ - вектор нормали к границе



Классификатор как гиперплоскость (граница)

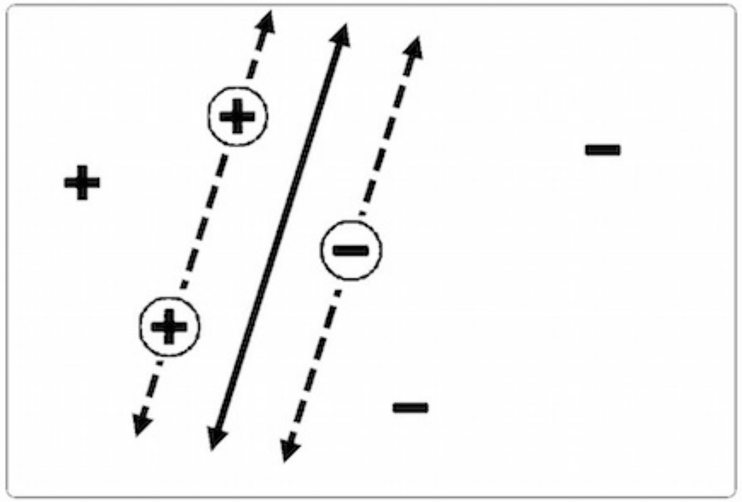
- **разделяющая гиперплоскость определяется уравнением**
  - $b + \mathbf{w} \cdot \mathbf{x} = 0$
  - $b + \mathbf{w} \cdot \mathbf{x} > 0$, по одну сторону границы
  - $b + \mathbf{w} \cdot \mathbf{x} < 0$, по **другую** сторону границы





Классификатор как гиперплоскость (граница)

- **разделяющая гиперплоскость определяется уравнением**
  - $b + \mathbf{w} \cdot \mathbf{x} = 0$
  - $b + \mathbf{w} \cdot \mathbf{x} > 0$, по одну сторону границы
  - $b + \mathbf{w} \cdot \mathbf{x} < 0$, по **другую** сторону границы
- **Корректное предсказание:**
  - $b + \mathbf{w} \cdot \mathbf{x}^{(i)} > 0$, если $y^{(i)} = +1$
  - $b + \mathbf{w} \cdot \mathbf{x}^{(i)} < 0$, если $y^{(i)} = -1$
  - или, что то же самое
  - $y^{(i)}(b + \mathbf{w} \cdot \mathbf{x}^{(i)}) > 0$
  - Обозначим через $\hat{\gamma}^{(i)} = y^{(i)}(b + \mathbf{w} \cdot \mathbf{x}^{(i)})$ степень уверенности при классификации i-го примера

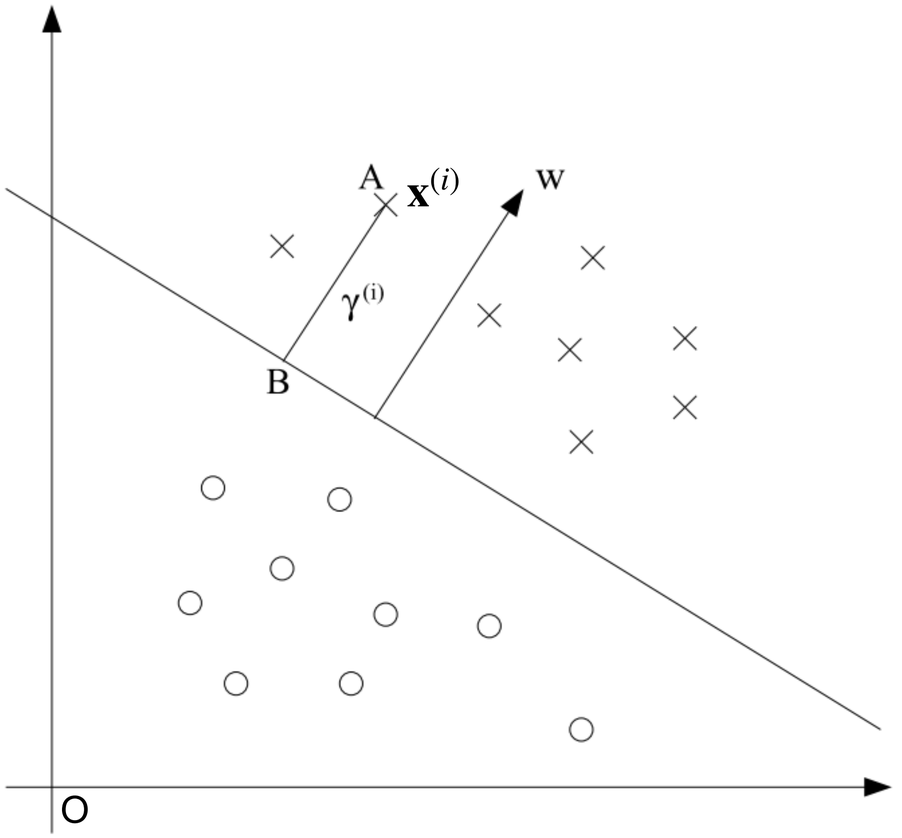


Функциональный зазор

- степень уверенности при классификации i-го примера:
  - $\hat{\gamma}^{(i)} = y^{(i)}(b + \mathbf{w} \cdot \mathbf{x}^{(i)})$
- было бы неплохо выбрать w и b так, чтобы максимизировать минимальную степень уверенности
  - $\max\limits_{w,b} J(w, b) = \max\limits_{w,b} \quad (\min\limits_{i=1,N} \hat{\gamma}^{(i)}) = \max\limits_{w,b} \quad [\min\limits_{i=1,N} y^{(i)}(b + \mathbf{w} \cdot \mathbf{x}^{(i)})]$
- Очевидно, что просто увеличивая норму вектора w, можно бесконечно увеличивать J
- Поэтому целевую функцию надо определять через расстояние от границы до точек обучающего множества
  - Пусть $\gamma^{(i)}$ - расстояние от i-го примера до границы

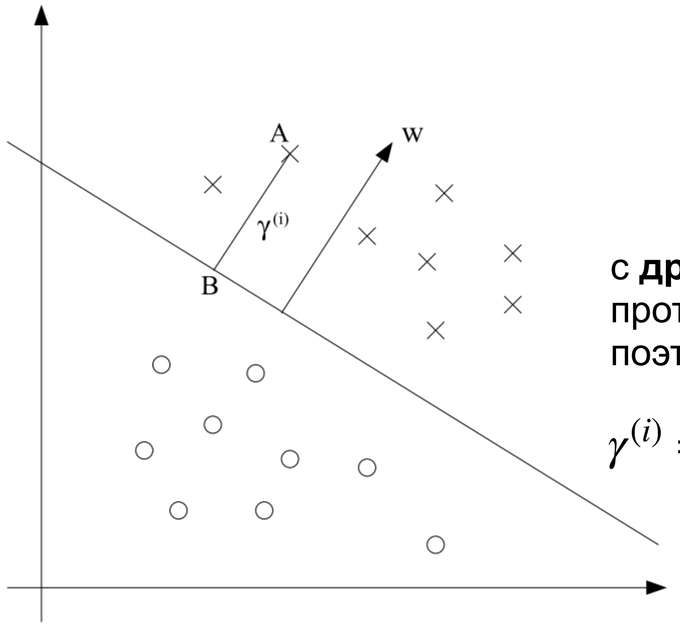
Определение геометрического зазора

- **Пусть $\gamma^{(i)}$ - расстояние от $\mathbf{x}^{(i)}$ до границы**
  - Пусть $\overline{OA} = \mathbf{x}^{(i)}$
  - Тогда
  - $\overline{OB} = \overline{OA} - \overline{BA} = \mathbf{x}^{(i)} - \gamma^{(i)}\frac{\mathbf{w}}{\|\mathbf{w}\|}$
  - $\mathbf{w} \cdot \overline{OB} + b = \mathbf{w} \cdot (\mathbf{x}^{(i)} - \gamma^{(i)}\frac{\mathbf{w}}{\|\mathbf{w}\|}) + b = 0$
  - следовательно $\mathbf{w} \cdot \mathbf{x}^{(i)} - \gamma^{(i)}\frac{\mathbf{w} \cdot \mathbf{w}}{\|\mathbf{w}\|} + b = 0$
  - наконец, $\mathbf{w} \cdot \mathbf{x}^{(i)} - \gamma^{(i)}\|\mathbf{w}\| + b = 0$



Определение геометрического зазора

- **Пусть $\gamma^{(i)}$ - расстояние от $\mathbf{x}^{(i)}$ до границы**
  - Пусть $\overline{OA} = \mathbf{x}^{(i)}$
  - Тогда
  - $\overline{OB} = \overline{OA} - \overline{BA} = \mathbf{x}^{(i)} - \gamma^{(i)}\frac{\mathbf{w}}{\|\mathbf{w}\|}$
  - $\mathbf{w} \cdot \overline{OB} + b = \mathbf{w} \cdot (\mathbf{x}^{(i)} - \gamma^{(i)}\frac{\mathbf{w}}{\|\mathbf{w}\|}) + b = 0$
  - следовательно $\mathbf{w} \cdot \mathbf{x}^{(i)} - \gamma^{(i)}\frac{\mathbf{w} \cdot \mathbf{w}}{\|\mathbf{w}\|} + b = 0$
  - наконец, $\mathbf{w} \cdot \mathbf{x}^{(i)} - \gamma^{(i)}\|\mathbf{w}\| + b = 0$



$$\gamma^{(i)} = \frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|}$$

с **другой** стороны знак будет противоположным.
поэтому в общем случае:

$$\gamma^{(i)} = y^{(i)}[\frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|}]$$

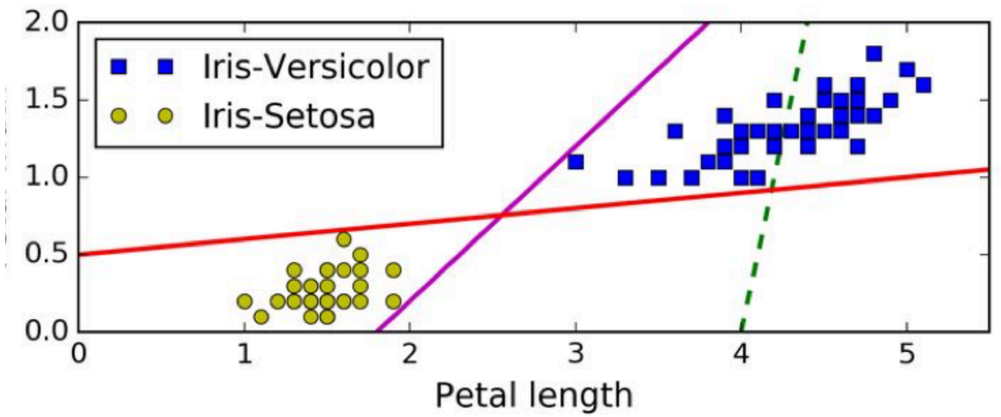
$$\gamma^{(i)} = y^{(i)}[\frac{\mathbf{w} \cdot \mathbf{x}^{(i)}}{\|\mathbf{w}\|} + \frac{b}{\|\mathbf{w}\|}] = \frac{1}{\|\mathbf{w}\|}y^{(i)}[\mathbf{w} \cdot \mathbf{x}^{(i)} + b]$$

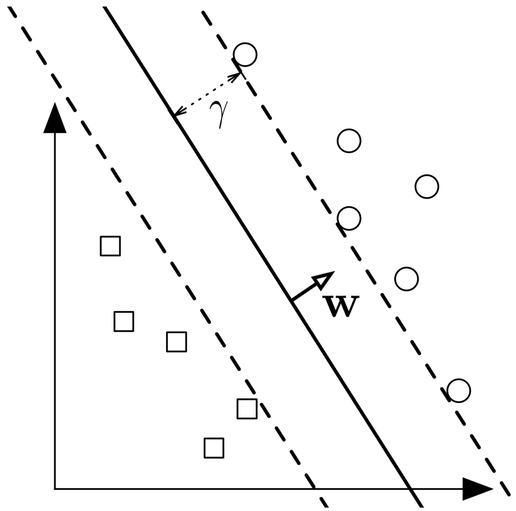
$$\hat{\gamma}^{(i)} = y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$$

$$\gamma^{(i)} = \frac{1}{\|\mathbf{w}\|}\hat{\gamma}^{(i)}$$

Определение геометрического зазора

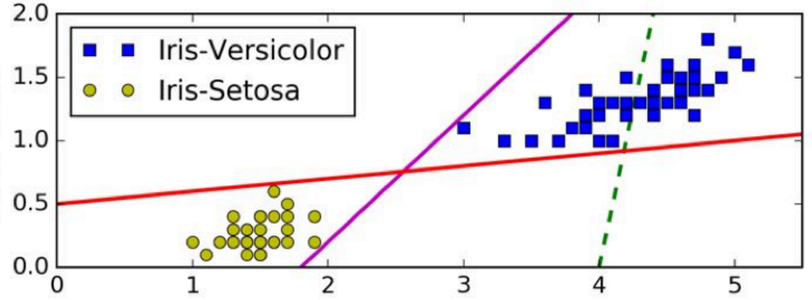
- расстояние от i-го примера до границы
  - $\gamma^{(i)}$
- рассмотрим минимальное расстояне от границы до обучающих примеров
  - $\gamma = \min_{i=1,N} \gamma^{(i)}$
- γ - геометрический зазор



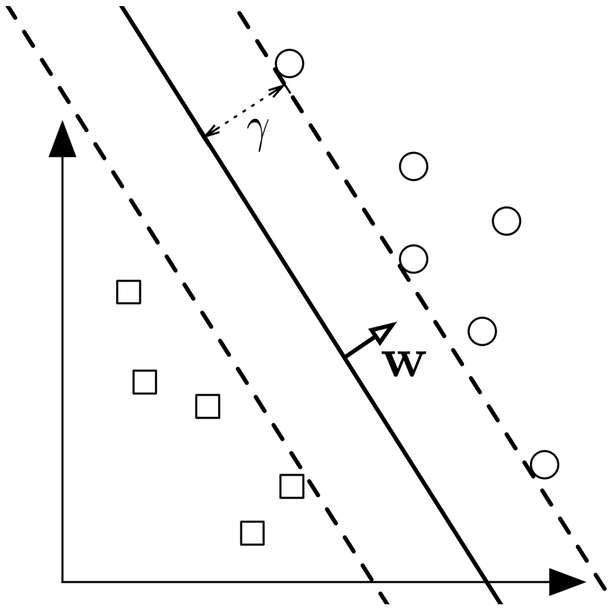


Максимизация зазора

- поскольку классы линейно разделимы, то минимальный зазор существует



- Цель состоит в максимизации зазора:
  - $$\max_{w,b} \gamma = \max_{w,b} \; [\min_{i=1,N} \gamma^{(i)}] = \max_{w,b} \min_{i=1,N} [\frac{1}{\|\mathbf{w}\|} \hat{\gamma}^{(i)}]$$
- минимум считается по всем обучающим примерам, а результат зависит и от w, и от b
  - Кроме этого при корректной классификации всех примеров должны выполняться еще и условия:
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq \min_{i=1,N} \hat{\gamma}^{(i)}$,
  - или что то же самое:
  - $$y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq \hat{\gamma}$$



Задача оптимизации и функция потерь в SVM

- **Итак, нам необходимо решить задачу максимизации**
  - $\max\limits_{w,b} \gamma$ **или** $\max\limits_{w,b} \frac{1}{\|\mathbf{w}\|}\hat{\gamma}$
- **с ограничениями**
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq \hat{\gamma}$
  - **для всех** $i = 1,...,N$
- **Решение не зависит от выбора** $\hat{\gamma}$
  - Пусть $\hat{\gamma} = 1$
- **Получим эквивалентную задачу минимизации**

$$\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq 1$$

- **Это задача квадратичного программирования**

Что значит (геометрически) $\hat{\gamma} = 1$?

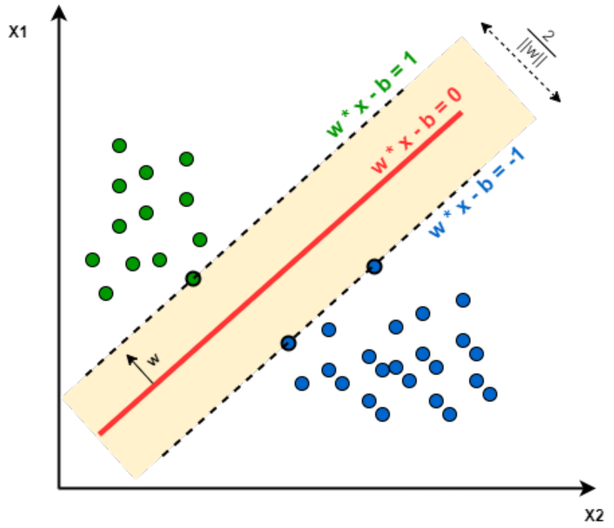
- **Для любых точек в датасете: $\hat{\gamma}^{(i)} = y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$**
  - Рассмотрим 2 точки по разные стороны, но на границе зазора
  - для $\mathbf{x}_+ : \mathbf{w} \cdot \mathbf{x}_+ + b = \hat{\gamma}_+$
  - для $\mathbf{x}_- : \mathbf{w} \cdot \mathbf{x}_- + b = -\hat{\gamma}_-$
  - для них $\hat{\gamma}_+ = \hat{\gamma}_- = \hat{\gamma} = 1$

- **то есть ограничения**
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq 1$
  - в виде равенств выполняются только для опорных векторов

- **а вот строгие неравенства**
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) > 1$
  - для всех правильно классифицированных точек вне зазора

$$\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2$$

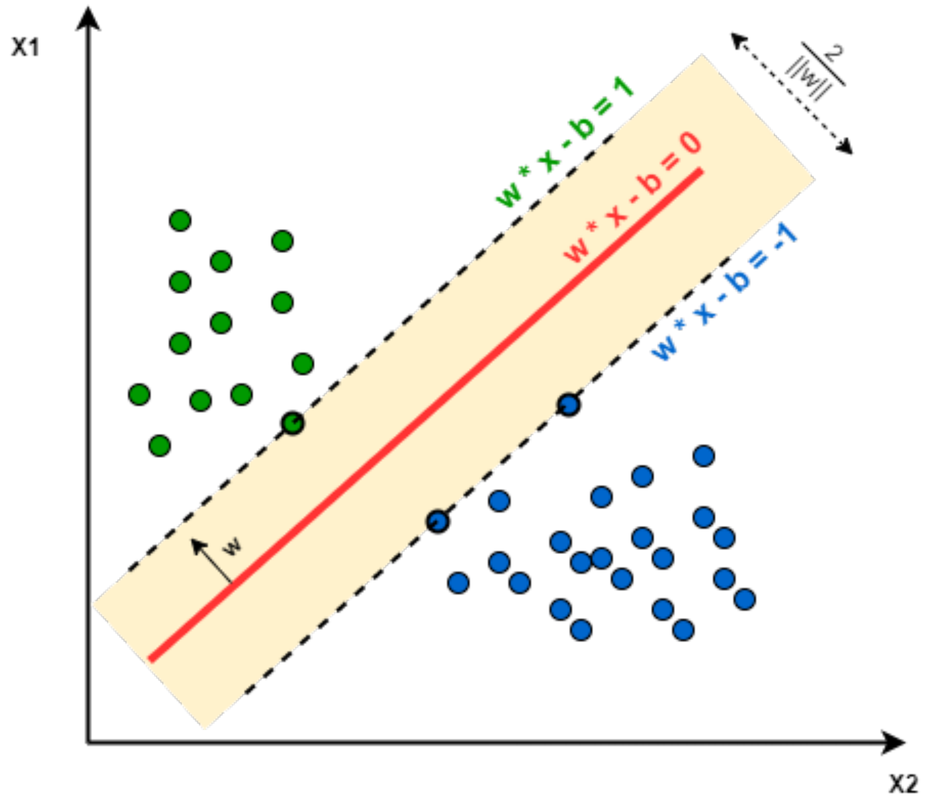
$$y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq 1$$



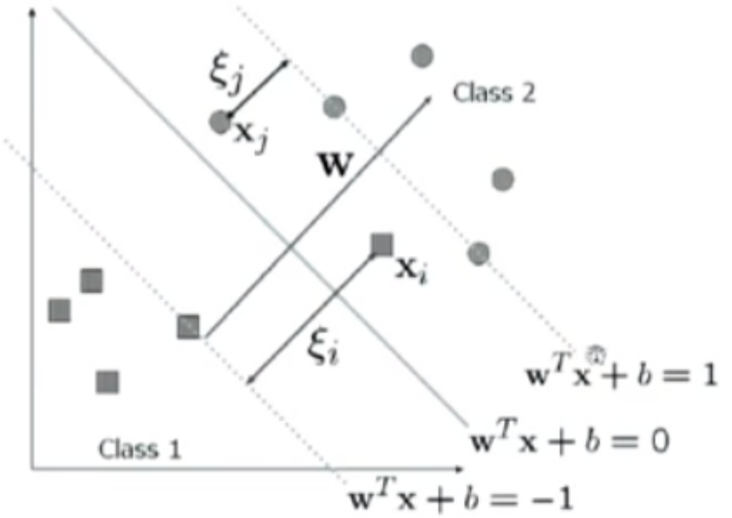
Слабые переменные

- **равенства**
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) = 1$
  - для опорных векторов
- **строгие неравенства**
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) > 1$
  - для всех правильно классифицированных точек **вне** зазора
- **Ошибки классификации**
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) < 0$

- **для точек попавших внутрь полосы**
  - $0 \leq y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) < 1$



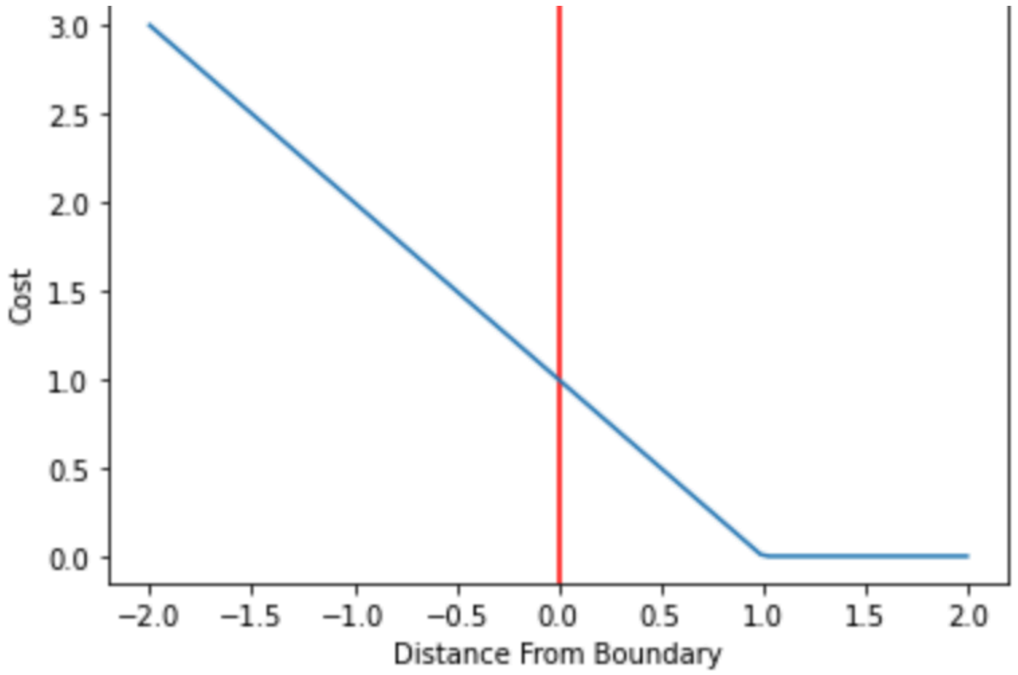
Слабые переменные



- **Определение**
  - $\xi_i = 0$, если нет ошибки
  - $0 < \xi_i \leq 1$, если x_i внутри зазора и с правильной стороны от границы
  - $\xi_i > 1$, если x_i внутри зазора с неправильной стороны от границы (ошибка классификатора)
- **Итоговая целевая функция:**
  - $\min\limits_{w,b}[\frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{N}\xi_i]$
  - C — гиперпараметр
- **Итоговые ограничения для Soft SVM:**
  - $y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)} + b) \geq 1 - \xi_i$
  - $\xi_i \geq 0$

Hinge Loss (Функция потерь для SVM)

- **Как ведет себя функция потерь для отдельного примера?**
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq 1 - \xi_i$
  - $\xi_i \geq 0$ (по сути, ξ_i — доп. штраф от $\mathbf{x}^{(i)}$)
  - перепишем первое неравенство:
  - $\xi_i \geq 1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$
  - или что тоже самое: $\xi_i = \max(0, 1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b))$
  - штраф зависит от дистанции до границы,
  - которая вычисляется как $\mathbf{w} \cdot \mathbf{x}^{(i)} + b$



- **Обратите внимание не схожесть с функцией потерь для логистической регрессии**

Функция потерь для SVM

- $$J(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
- Перепишем ее
- $$J(\mathbf{w}, b) = \frac{1}{2}\sum_{k=1}^{D} w_k^2 + C\sum_{i=1}^{N}\max(0, 1 - y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)} + b))$$
- D - число признаков
- N - число примеров
- Что теперь можно считать целевой функцией, а что параметром регуляризации?

Двойственная задача

Прямая задача

$$\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq 1$$

Двойственная задача

- $$\max_{\alpha} [-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)}\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} + \sum_{i}\alpha_i]$$
- $$\sum_{i=1}^{N}\alpha_i y^{(i)} = 0$$
- $$\alpha_i \geq 0$$

Двойственная задача

- Решение исходной задачи
  - $\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2$
  - $y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) \geq 1$
- Можно заменить на решение двойственной задачи (которую проще решать)
- Рассмотрим шаги, необходимые для этого

Преобразование ограничений

- **Вводим новую целевую функцию:**
  - $$L(\mathbf{w}, b, \alpha) = f(\mathbf{w}, b) + \sum_{i=1}^{N} \alpha_i g_i(\mathbf{w}, b)$$
- **где**
  - α_i — множители Лагранжа ($\alpha_i \geq 0$)
  - $g_i(\mathbf{w}, b) \leq 0$
  - $f(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2$ — исходная целевая функция
  - $g_i(\mathbf{w}, b) = 1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$ — результат преобразования неравенств
  - если для $\mathbf{w}^*, b^*$ — достигается минимум f, то там же достигается и минимум L
  - $\alpha_i g_i(\mathbf{w}^*, b^*) = 0$

Дифференцируем

- приравниваем производную функции Лагранжа к нулю
  - $\nabla L(\mathbf{w}, b, \alpha) = 0$
  - $\nabla_{\mathbf{w}} L(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)} = 0 \quad => \quad \mathbf{w} = \sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)}$
  - $\nabla_b L(\mathbf{w}, b, \alpha) = -\sum_{i=1}^{N} \alpha_i y^{(i)} = 0 \quad => \quad \sum_{i=1}^{N} \alpha_i y^{(i)} = 0$
- Преобразуем целевую функцию:
  - $L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{N} \alpha_i (1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)) = -\frac{1}{2}\|\mathbf{w}\|^2 + \sum_i \alpha_i =$
  - $= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j (y^{(i)} \mathbf{x}^{(i)}) \cdot (y^{(j)} \mathbf{x}^{(j)}) + \sum_i \alpha_i$
  - зависит только от данных и множителей α_i

Итоговая двойственная задача

- **Задача:**
  - $$\max_{\alpha} [-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} + \sum_{i} \alpha_i]$$
  - $$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$
  - $\alpha_i \geq 0$ (часть из условий Каруша-Куна-Таккера)
- **Решая эту задачу квадратичного программирования мы можем**
  - найти решение α_i
  - вычислить $\mathbf{w}$,
  - затем найти b, поскольку $b = y^{(i)} - \mathbf{w} \cdot \mathbf{x}^{(i)}$ для любого опорного вектора, но на практике усредняют по нескольким (всем) опорным векторам

Краткий итог

- Двойственную задачу решать быстрее, чем основную, когда количество обучающих примеров меньше, чем количество признаков.

- Что еще более важно, двойственная задача делает возможным трюк с ядром, а прямая - нет.

Predict в SVM

- **Для нового примера $\mathbf{x}$:**
  - predict дает значение sign($\mathbf{w} \cdot \mathbf{x} + b$) = sign ($\sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)} \mathbf{x} + b$)

- **Из ККТ следует, что для решения:**
  - $\alpha_i g_i(\mathbf{w}^*, b^*) = 0$
  - $g_i(\mathbf{w}, b) = 1 - y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$ — то есть $g_i(\mathbf{w}, b)$=0 только если $\mathbf{x}^{(i)}$ лежит на границе зазора
  - и только в этом случае α_i может быть ненулевым (!)
  - примеры на границе зазора называются опорными векторами

- **то есть для предсказания не нужно знать координаты вектора нормали, нужны только опорные вектора из датасета !**

Трюк с ядром

- **Для нового примера x:**
  - predict дает значение sign($\mathbf{w} \cdot \mathbf{x} + b$) = sign ($\sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)} \mathbf{x} + b$)
  - Поскольку целевая функция зависит тоже только от скалярных произведений $\mathbf{x}^{(i)} \mathbf{x}^{(j)}$
  - (а не от самих векторов), то и для тренировки достаточно знать только произведения $\mathbf{x}^{(i)} \mathbf{x}^{(j)}$
- **При использовании трюка с ядром это важно:**
  - max [$-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \sum_{i} \alpha_i$]
  - где $K(a, b)$ — ядро

Трюк с ядром

- **Для нового примера x:**
  - predict дает значение sign($\mathbf{w} \cdot \mathbf{x} + b$) = sign ($\sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^{(i)} \mathbf{x} + b$)
  - Поскольку целевая функция зависит тоже только от скалярных произведений $\mathbf{x}^{(i)} \mathbf{x}^{(j)}$
  - (а не от самих векторов), то и для тренировки достаточно знать только произведения $\mathbf{x}^{(i)} \mathbf{x}^{(j)}$

- **При использовании трюка с ядром это важно:**
  - max [$-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \sum_{i} \alpha_i$]
  - где $K(a, b)$ — ядро

$$K(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T \mathbf{b}$$

$$K(\mathbf{a}, \mathbf{b}) = (\gamma \mathbf{a}^T \mathbf{b} + r)^d$$

$$K(\mathbf{a}, \mathbf{b}) = \exp(-\gamma \| \mathbf{a} - \mathbf{b} \|^2)$$

$$K(\mathbf{a}, \mathbf{b}) = \tanh(\gamma \mathbf{a}^T \mathbf{b} + r)$$

Заключение

- Функция потерь в SVM
- Слабые переменные
- Двойственная задача
- SVM с ядром

Ссылки

- https://programmathically.com/understanding-hinge-loss-and-the-svm-cost-function/
- https://en.wikipedia.org/wiki/Mercer%27s_theorem