МАШИНА ОПОРНЫХ ВЕКТОРОВ

Постановка задачи

Дан датасет $D = \{(x_i \quad y_i) \mid x_i \in R^n, y_i \in \{+1 \quad -1\}\}_{i=1}^{N}$ и его геометрическая интерпретация:

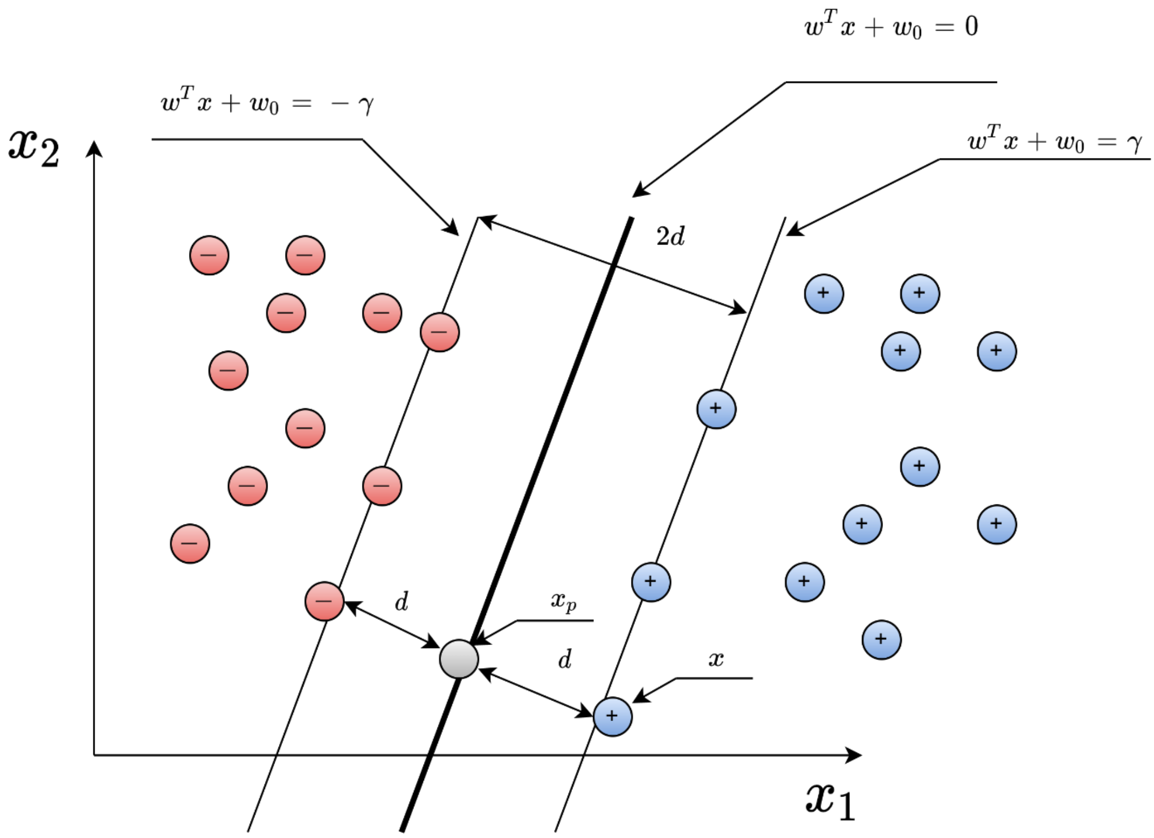

Рисунок 1 - Геометрическая интерпретация постановки задачи

Где:

- $H = \{x \mid w^T x + w_0 = 0\}$ - разделяющая гиперплоскость
- $H_{+1} = \{x \mid w^T x + w_0 = \gamma\}$ - гиперплоскость, проходящая через ближайшую точку x_i при условии, что для всех возможных точек расстояние от нее до H минимально, также $y_i = +1$ и $H \parallel H_{+1}$
- $H_{-1} = \{x \mid w^T x + w_0 = -\gamma\}$ - гиперплоскость, проходящая через ближайшую точку x_i при условии, что для всех возможных точек расстояние d от нее до H минимально, также $y_i = -1$ и $H \parallel H_{-1}$
- d - минимальное расстояние между H и H_{-1}, H и H_{-1}, $d \perp H$
- $x_p \in H$
- $x \in H_{+1}$

Необходимо найти такие w^T, w_0 при которых $d \to \min$

Поиск минимального расстояния d

Координаты точки $x_p \in H$ выражаются через расстояние как $x_p = x - d$.

По определению для плоскости $Ax + By + Cz + D = 0$ нормальный вектор имеет координаты $\vec{n} = \{A \quad B \quad C\}$. Следовательно, расстояние d равно взвешенному нормальному вектору $\vec{w}$. То есть $d = \alpha w$.

Точка $x_p \in H$ выражается в терминах гиперплоскости: $w^T x_p + w_0 = 0$. Эту точку можно выразить через $x_p = x - d$ и $d = \alpha w$:

$$
\begin{aligned}
&w^T x_p + w_0 = 0 \\
&w^T (x - d) + w_0 = 0 \\
&w^T (x - \alpha w) + w_0 = 0 \\
&w^T x - \alpha w^T w + w_0 = 0 \\
&w^T x + w_0 = \alpha w^T w \\
&\alpha = \frac{w^T x + w_0}{w^T w}
\end{aligned}
$$

Длину вектора можно выразить через норму Евклида:

$$
\begin{aligned}
&\|d\| = \sqrt{d^T d} \\
&\|d\| = \sqrt{\alpha^2 w^T w} \\
&\|d\| = \sqrt{\left(\frac{w^T x + w_0}{w^T w}\right)^2 w^T w} \\
&\|d\| = \sqrt{\frac{\left(w^T x + w_0\right)^2}{w^T w}} \\
&\|d\| = \frac{\left|w^T x + w_0\right|}{\|w\|}
\end{aligned}
$$

Классификатор с максимальным зазором

Определим зазор как $M = 2\|d\|$. Таким образом, максимизируя $\|d\|$ мы максимизируем M. И нужно найти такие коэффициенты гиперплоскости H, которые будут максимизировать зазор для тех точек, которые ближе всего находятся к H, при условии, что ошибок нет. То есть $y_i\left(w^T x_i + w_0\right) \geq 0, \forall i$

$$\begin{cases} \max\limits_{w,w_0} \frac{1}{\|w\|} \min\limits_{x_i \in D} \left| w^T x + w_0 \right| \\ y_i \left(w^T x_i + w_0 \right) \geq 0, \forall i \end{cases}$$

Уравнение 1 - Первая форма постановки задачи оптимизации

Поскольку H масштабно-инвариантна: $c\left| w^T x + w_0 \right| = 0$, то, что минимизация не взвешенной, что взвешенной функции, эту гиперплоскость не меняет и также приводит к требуемому решению. Таким образом, положим, что $\left| w^T x + w_0 \right| = 1$. Стоит так же отметить, что $\|w\|$ - не дифференцируема, но можем подобрать квадратичную сверху, то есть такую функцию, которая является дифференцируемой: $\|w\|^2$. Также максимизация d ведет к минимизации $\|w\|^2$.

$$\begin{cases} \min\limits_{w,w_0} \|w\|^2 \\ y_i \left(w^T x_i + w_0 \right) \geq 0, \forall i \\ \min \left| w^T x + w_0 \right| = 1 \end{cases}$$

Уравнение 2 - Вторая форма оптимизационной задачи

Объединив условия, получим другую постановку задачи выпуклого программирования:

$$\begin{cases} \min\limits_{w,w_0} \|w\|^2 \\ y_i \left(w^T x_i + w_0 \right) \geq 1, \forall i \end{cases}$$

Уравнение 3 - Третья форма постановки задачи

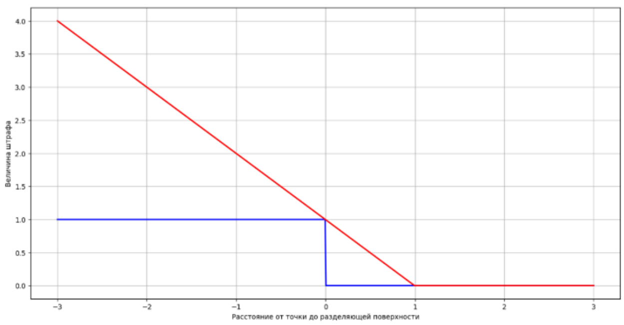

Рисунок 2 - Функция условие или функция риска hinge-loss $\max\left\{0, y_i \left(x_i^T w + w_0 \right)\right\}$

Модификация задачи для линейной несепарабельных данных

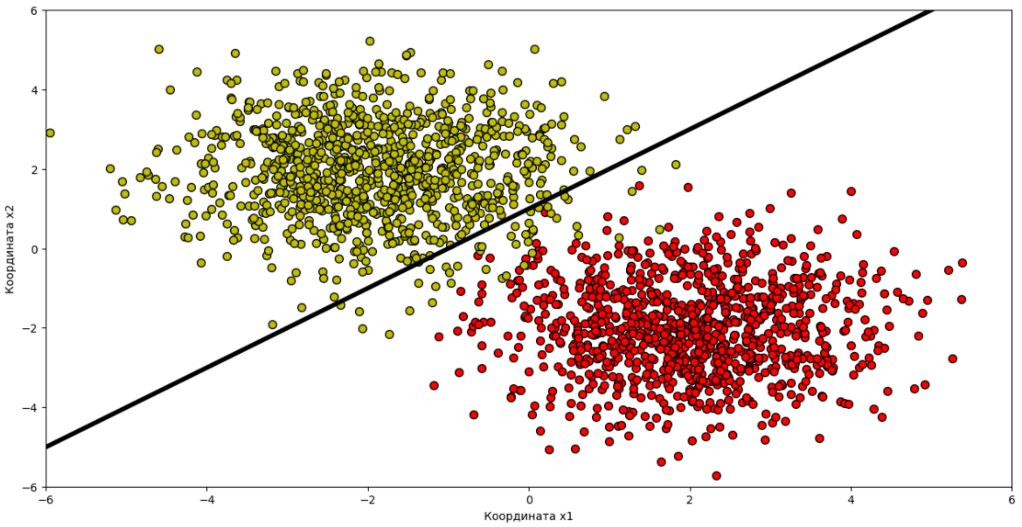

Рисунок 3- Пример линейно неразделимой выборки

Если мы введем вектор фиктивных переменных $\xi = \begin{pmatrix} \xi_1 & \xi_2 & \cdots & \xi_N \end{pmatrix}$ как величину штрафа за выход значений в пространство другого класса, то можно использовать $y_i\left(w^T x_i + w_0\right) \geq M\left(1-\xi_i\right)$ при $i = \overline{1,m}$. Здесь $M\left(1-\xi_i\right)$ - величина пропорциональная расстоянию, при котором прогноз находится на неправильной стороне своего зазора. Поскольку $M = 1$ то условие упрощается до: $y_i\left(w^T x_i + w_0\right) \geq 1-\xi_i, \forall i$. При этом мы будем решать оптимизационную задачу для общего случая и потребуем, чтобы: $\xi_i \geq 0, \forall i$. Чтобы учесть эти фиктивные переменные, необходимо внести их в постановку задачи. Таким образом, постановка задачи поиска оптимальной разделяющей поверхности в общем случае будет выглядеть так:

$$\begin{cases} \min\limits_{w,w_0} \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{N}\xi_i \\ y_i\left(w^T x_i + w_0\right) \geq 1-\xi_i, \forall i \\ \xi_i \geq 0 \end{cases}$$

Уравнение 4 - Постановка задачи оптимизации для линейно несепарабельных данных

Решение задачи оптимизации

Для решения задачи квадратичного программирования согласно уравнению 4 необходимо перейти к двойственной задаче Вульфа, что позволит нам использовать ядерный трюк. Так называемый возможный переход от одного пространства к другому, если нет такой прямой, которая может быть решена на основе уравнения 4. Двойственная задача Вульфа также. подразумевает условия Каруша-Куна-Таккера.

$$\begin{cases} f(x)+\sum_{j=1}^{m} u_j g_j(x) \to \max \\ \nabla f(x)+\sum_{j=1}^{m} u_j \nabla g_j(x) = 0 \forall j \\ u_j \geq 0 \forall j \\ u_j \nabla g_j(x) = 0 \forall j \end{cases}$$

Уравнение 5 - Двойственная задача Вульфа

Составим Лагранжиан:

$$\max_{\lambda,\xi} \min_{w,w_0,\xi} L\left(w, w_0, \xi_i, \lambda_i, \xi_i\right) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\lambda_i\left[y_i\left(w^T x_i + w_0\right) - 1 + \xi_i\right]$$

$$\max_{\lambda,\xi} \min_{w,w_0,\xi} L\left(w, w_0, \xi_i, \lambda_i, \xi_i\right) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\lambda_i y_i w^T x_i - \sum_{i=1}^{N}\lambda_i y_i w_0 - \sum_{i=1}^{N}\lambda_i\xi_i + \sum_{i=1}^{N}\lambda_i$$

Найдем минимум:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N}\lambda_i y_i x_i = 0 => w = \sum_{i=1}^{N}\lambda_i y_i x_i$$

$$\frac{\partial L}{\partial w_0} = \sum_{i=1}^{N}\lambda_i y_i = 0$$

$$\frac{\partial L}{\partial \xi} = C - \sum_{i=1}^{N}\mu_i - \sum_{i=1}^{N}\lambda_i = 0$$

При условии, что $\lambda_i, w, \xi_i \geq 0, \forall i$

Выполним подстановку и получим двойственную задачу Вульфа:

- $L\left(w, w_0, \xi_i, \lambda_i, \xi_i\right) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N}\lambda_i y_i w^T x_i - \sum_{i=1}^{N}\lambda_i y_i w_0 + \sum_{i=1}^{N}\lambda_i + \sum_{i=1}^{N}\xi_i\left(C - \lambda_i - \mu_i\right)$ (Группировка множителей)
- $\frac{1}{2}\|w\|^2 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_i$ (подстановка $\frac{\partial L}{\partial w}$)
- $\sum_{i=1}^{N}\lambda_i y_i x_i^T w = \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_i$ (подстановка $\frac{\partial L}{\partial w}$)
- $\frac{1}{2}\|w\|^2 - \sum_{i=1}^{N}\lambda_i y_i x_i^T w = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_i$ (разность двух предыдущих)

- $\sum_{i=1}^{N} \xi_i \left(C - \lambda_i - \mu_i \right) = 0$ (равенство из $\frac{\partial L}{\partial \xi}$)

Таким образом, переходим к двойственной задаче квадратичного программирования при условии ККТ:

$$\begin{cases} L\left(\lambda_i\right) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_i + \sum_{i=1}^{N} \lambda_i \to \max\limits_{\lambda_i} \\ \frac{\partial L}{\partial w} = w = \sum_{i=1}^{N} \lambda_i y_i x_i \\ \frac{\partial L}{\partial w_0} = \sum_{i=1}^{N} \lambda_i y_i = 0 \\ \frac{\partial L}{\partial \xi} = C - \sum_{i=1}^{N} \mu_i - \sum_{i=1}^{N} \lambda_i = 0 \\ \lambda_i \left(y_i \left(w^T x_i + w_0 \right) - \left(1 - \xi_i \right) \right) = 0 \\ \mu_i \xi_i = 0 \\ y_i \left(w^T x_i + w_0 \right) - \left(1 - \xi_i \right) \geq 0 \\ i = \overline{1, N} \end{cases}$$

Уравнение 6 - Постановка задачи для SVM

Такая задача решается вычислительными методами с помощью компьютера

Виды опорных точек

- Если $\lambda_i = 0 \Rightarrow C = \mu_i$ и из Лагранжиана сокращается штраф за выход i-ой точки в чужой класс. Точка лежит в своей классе.
- Если $0 < \lambda_i < C$, то $0 < \mu_i < C$ (штраф за выход в чужую область $\xi_i = 0$), штраф $y_i \left(w^T x_i + w_0 \right) = 1$ - В этом случае объект находятся на границе классов и является опорным.
- Если $\lambda_i = C$, то $\mu_i = 0$ и $y_i \left(x_i^T w + w_0 \right) + \xi_i - 1 = 0$, что дает нам $\xi_i \geq 0$. Такие объекты находятся в толще чужого класса и также являются опорными нарушителями.

Таким образом точка называется опорной, когда $\lambda_i \neq 0$