1. Опорна Машина Векторів

Опорна Машина Векторів - це алгоритм лінійної класифікації. Що означає лінійна класифікаці? Це означає, що ми зображаємо нашу вибірку у вигляді набору з N точок у просторі $\mathbb{R}^p$. Тобто, як і раніше маємо p - числових регресорів, та N спостережень, тоді i-те спостереження зображається у вигляді точки в $\mathbb{R}^p$: $x_i = (x_i^1, \ldots, x_i^p)$. Для початку розглядаємо бінарну класифікацію, і позначатимемо відгук $y_i \in \{-1, 1\}$. Відповідно, кожна точка з набору x_i належить до якогось класу (а саме до класу y_i). Лінійна класифікація полягає у тому, щоби провести пряму, яка розділить ці два класи.

Припустимо, спочатку, що класи лінійно роздільні, тобто існує така пряма, по одну сторону якої лежать точки з $y_i = 1$ а по іншу з $y_i = -1$.

Визначимо гіперплощину:

$$\{x : \ f(x) = x^T\beta + \beta_0 = 0\}, \text{ де } x \in \mathbb{R}^{p\times 1}, \ \beta \in \mathbb{R}^{p\times 1}, \ \beta_0 \in \mathbb{R}.$$

де β це одиничний вектор, $||\beta|| = 1$. Класифікаційне правило, створене $f(x)$ виглядає таким чином:

$$G(x) = \text{sign}(\sum_{k=1}^{p} x^k\beta_k + \beta_0).$$

Опорна Машина Векторів шукає "найкращу" пряму, тобто таку, яка "максимально" розділяє класи.

За допомогою $f(x)$ можна визначити знакозмінну відстань з точки x до гіперплощини $f(x) = x^T\beta + \beta_0 = 0$. Оскільки класи лінійно роздільні то можливо знайти функцію $f(x) = x^T\beta + \beta_0$, таку що $y_i f(x_i) > 0, \ \forall i$. Таким чином можливо знайти гіперплощину що створює найбільший відступ між тренувальними точками для класів 1 та -1.

Для початку нагадаємо трохи алгебри. Гіперплощина в $\mathbb{R}^p$ задається формулою $\beta^T x + \beta_0 = 0$. Тут β це вектор нормалі, перпендикулярний до гіперплощини. Якщо вздовж вектору β відкласти відстань β_0 отримаємо "новий нуль" - x_0. Для довільної точки $x \in \mathbb{R}^{p\times 1}$ відстань від x до гіперплощини задається формулою:

$$d(L, x) = \beta^T(x - x_0) = \frac{1}{||\beta||}(\beta^T x + \beta_0),$$

для довільного $x' \in L$: $\beta^T x' = -\beta_0$.

Наша задача знайти таку пряму, щоб відстань від усіх тренувальних точок до прямої була максимальною (поки розглядаємо випадок лінійно роздільних даних). Позначимо цю відстань $2M$.

Тоді оптимізаційна проблема має вигляд:

$$\begin{cases} \max_{\beta,\beta_0,||\beta||=1} M \\ y_i(x_i\beta + \beta_0) \geq M, \ i = 1, \ldots, N \end{cases} \tag{1}$$

Відступ з кожного боку гіперплощини до найближчої тренувальної точки буде принаймні M, а отже вся смуга буде мати ширину $2M$.

Зауважимо, що задача полягає у максимізації відстані від усіх точок, тобто $y_i(x_i\beta + \beta_0) \geq ||\beta||M$ ($x_i \in \mathbb{R}^{1\times p}$ - i-те тренувальне спостереження) . Зауважимо, що множення на будь яку константу не змінить цієї задачі оптимізації, тому можна покласти $M = ||\beta||^{-1}$.

Дана проблема може буте перефразована більш зручно:

$$\begin{cases} \min_{\beta,\beta_0} ||\beta|| \\ y_i(x_i\beta + \beta_0) \geq 1, \ i = 1, \ldots, N \end{cases} \tag{2}$$

Зауважимо, що умови $||\beta|| = 1$ більше немає, а $M = ||\beta||^{-1}$. Дана проблема є задачею опуклої оптимізації і тому допускає розв'язання, як це показано у дискримінантному аналізі.

Розв'язучи цю задачу (а це є задача опуклої оптимізації), записавши умови Каруша-Куна-Такера ми побачимо, що значеня матимуть лише точки що потрапляють на край межі. Такі точки і називаються опорними векторами.

Розглянемо ситуацію, коли класи не є лінійно роздільними, або вірніше, є "трішки" лінійно нероздільними.

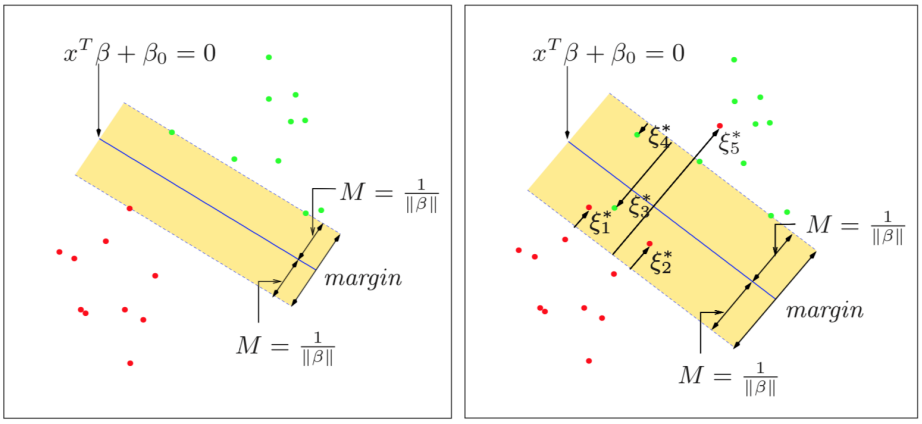

FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong side of their margin by an amount $\xi_j^* = M\xi_j$; points on the correct side have $\xi_j^* = 0$. The margin is maximized subject to a total budget $\sum \xi_i \leq$ constant. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.*

Припустимо зараз що класи перетинаються у фазовому просторі. Один зі способів як працювати з таким перетином, все ще максимізувати M, але дозволити деяким точкам бути на неправильній стороні. Визначимо допоміжні (люфтові - slack) змінні: $\xi = (\xi_1, \ldots, \xi_N)$ (це такі змінні які в задачі оптимізації перетворюють нерівності на рівності). Існує два природніх способи як модифікувати задачу оптимізації:

$$y_i(x_i^T\beta + \beta_0) \geq M - \xi_i, \tag{3}$$

або

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i), \tag{4}$$

$\forall i,\ \xi_i \geq 0,\ \sum_{i=1}^{N} \xi_i \leq$ constant.

Ці два варіанти ведуть до різних розв'язків. Перший варіант виглядає більш природнім, оскільки він вводить вимірює перекриття (overlap) у абсолютній відстані, другий вибір вводить відносну відстань для перекриття, що змінює ширину відступу M. Однак перший варіант веде до неопуклої задачі оптимізації а другої до опуклої. Таким чином ми використаємо другу форму, що приведе нас до "стандартних" опорних векторів.

Ідея полягає в наступному. Величина ξ_i в обмеженнях $y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i)$ - це величина пропорційна "перестрибку" прогнозу $f(x_i) = x_i^T\beta + \beta_0$ на неправильній

стороні відступу. Тому обмежуючи суму $\sum \xi_i$ ми обмежуємо загальну пропорційну величину на яку точкам "дозволяється" перестрибнути на неправильну сторону відступу. Неправильна класифікація відбувається коли $\xi_i > 1$, отже обмеження $\sum \xi_i < K$, обмежує загальну кількість неправильних класифікацій величиною K.

Можемо переформулювати задачу оптимізації позбавившись обмеження на норму β, визначивши $M = ||\beta||^{-1}$:

$$\begin{cases} ||\beta|| \to \min, \\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \ \forall i, \\ \xi \geq 0, \\ \sum \xi_i \leq \text{constant} \end{cases} \tag{5}$$

Зауважимо, що величини ξ_i мають наступну інтерпритацію: якщо $\xi_i = 0$, то точка лежить у своїй півплощині поза смугою, якщо $\xi \in (0, 1)$, то точка лежить у смузі, і якщо $\xi \geq 1$, то точка лежить поза смугою у неправильній півплощині.

Із сутності задачі оптимізації бачимо, що точки які лежать у правильних півплощинах не відіграють великої ролі, що є однією з відмінностей від дискримінантного аналізу, де вирішуючі межі (межі смуги) визначені коваріацією розподілів класів та позиціями центроїдів. Такі точки, з $\xi > 0$, що впилвають на класифікатор і називаються опорними векторами.

Задачу (5) можна переформулювати у наступному вигляді:

$$\begin{cases} \min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum\limits_{i=1}^{N} \xi_i, \\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi \ \forall i \xi \geq 0 \end{cases} \tag{6}$$

Тут "ціновий" параметр C виступає в ролі константи з (5). Лінійно роздільний випадок відповідає $C = \infty$, (тобто всі $\xi_i = 0$).

Функція Лагранжа має вигляд:

$$L = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left(y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)\right) - \sum_{i=1}^{N} \mu_i\xi_i,$$

яку слід мінімізувати по β, β_0 та ξ_i. Прирівнявши відповідні похідні до нуля, отримаємо:

$$\begin{aligned} \beta &= \sum_{i=1}^{N} \alpha_i y_i x_i, \\ 0 &= \sum_{i=1}^{N} \alpha_i y_i, \\ \alpha_i &= C - \mu_i, \ \forall i, \end{aligned} \tag{7}$$

а також невід'ємність обмежень $\alpha_i, \mu_i, \xi_i \geq 0, \ \forall i$.

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] = 0, \tag{8}$$

$$y_i(x_i^T\beta + \beta_0) - (1 - \xi_i) \geq 0, \tag{9}$$

Роз'язок бути мати вигляд:

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i,$$

де ненульові коефіцієнти $\hat{\alpha}_i$ будуть лише для тих спостережень i, для яких (9) в точності виконано (в силу (8)). Ці спостереження і називають опорними векторами, оскільки $\hat{\beta}$ зображено у вигляді лише цих векторів.

Серед цих точок, деякі будуть лежати на межі смуги ($\hat{\xi}_i = 0$), а тому будуть характеризуватися $0 < \hat{\alpha}_i < C$, решта $\hat{\xi}_i > 0$ матимуть $\hat{\alpha} = C$.

Отримавши вектори $\hat{\beta}_0$ та $\hat{\beta}$, класифікатор має вигляд:

$$\hat{G}(x) = \text{sign}(\hat{f}(x)) = \text{sign}(x^T\hat{\beta} + \hat{\beta}_0).$$

Параметром налаштування (тюнінгу) моделі є константа C.

1.1. **Використання ядер у SVM..** Розглянуті вище моделі є лінійними. Як і в задачах регресії їх можна зробити нелінійними додавши додаткові ознаки, такі як степені початкових даних, або сплайни. Так, для кожного x_i можна розглянути функцію $h\colon \mathbb{R}^p \to \mathbb{R}^m$:

$$h(x_i) = (h_1(x_i), \ldots, h_m(x_i)),$$

і звичайний класифікатор:

$$G(x) = \text{sign}(h(x)^T\beta + \beta_0).$$

Зауважимо, що задачу оптимізації яку ми розглядали раніше, можна переформулювати лише у вигляді скалярних добутків, так функція L_D може бути зображена у вигляді:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j < h(x_i), h(x_j) >,$$

розв'язок цієї задачі може бути представлений у вигляді:

$$f(x) = h(x)^T\beta + \beta_0 = \sum_{i=1}^{N}\alpha_i y_i < h(x), h(x_i) > +\beta_0.$$

Таким чином бачимо, що знати всю функцію $h(x)$ непотрібно, а потрібно знати лише скалярні добутки, або ядерну функцію:

$$K(x, x') =< h(x), h(x') > .$$

Щоб функція $K(x, y)$ могла зображатись у вигляді скалярного добутку вона повинна бути симетричною, додатньо напів-визначеною.

Популярними ядерними функціями є:

$$\begin{array}{l} \text{Поліноміальне ядро, степені d } K(x, x') = (1+ < x, x' >)^d, \\ \text{Радіальне (Гаусове) ядро } K(x, x') = e^{-\gamma||x-x'||^2}, \\ \text{Тангенс гіперболічний } K(x, x') = tanh(k_1 < x, x' > +k_2 > . \end{array} \tag{10}$$

Наприклад для поліноміального ядра другого степеня:

$$K(x, x') = (1+ < x, x' >)^2 = 1 + 2x_1x_1' + 2x_2x_2' + (x_1x_1')^2 + (x_2x_2')^2 + 2x_1x_2x_1'x_2'.$$

Таким чином $m = 6$ і $h(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$. Тоді має місце зображення: $K(x, x') =< h(x), h(x') >$.

В багатовимірних просторах роль параметра C полягає у згладжуванні розділяючої повернхні. Справді, якщо C - велике, це приведе до малих значень ξ_i тобто до малої кількості неправильно класифікованих тренувальних даних, що як правило веде до перенавчання та нерегулярної розділячої кривої.