

Lecture 16: Learning: Support Vector Machines

Введение

Мы уже изучили простые методы обучения, рассмотрели идею, основанную на строении мозга человека, обратились к проблеме фонологии и изучению концепций. Но сегодня мы вернёмся в самое начало и подумаем о том, как можно разделить пространство с помощью границ решений.

Можно разделить пространство с помощью нейронной сети, метода ближайших соседей или распознающих деревьев. Все эти идеи очень просты и часто срабатывают. Сегодня мы обратимся к изощрённой идее, у которой, тем не менее, существует простая реализация: **метод опорных векторов**. На примере этой идеи мы посмотрим, как она постепенно развивалась, потому что такие алгоритмы не могут прийти в голову за один вечер.

Есть несколько способов нарисовать границу решений для одной и той же задачи. В методе ближайших соседей она строится по срединным перпендикулярам, для распознающего дерева части границы всегда будут параллельны одной из осей, а нейронная сеть может провести практически любую линию в зависимости от того, как она была обучена. На рисунке 1 приведён пример того, как эти алгоритмы решают одну и ту же задачу.

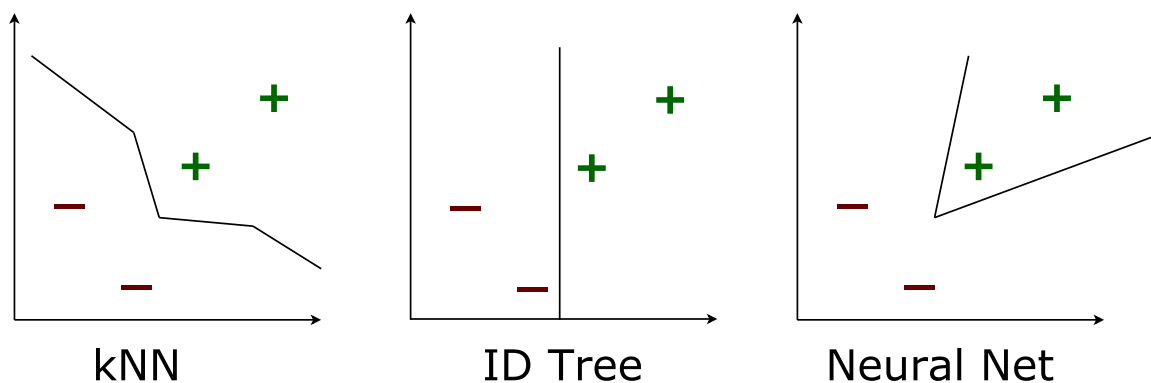


Рис. 1: Разделение пространства с использованием различных алгоритмов

Основная идея метода

Работа над разделением пространства с помощью границы решений велась на протяжении 50-75 лет, и казалось, что ничего нового уже нельзя придумать. Но, ко всеобщему удивлению, в начале 1990-х годов Владимир Вапник предложил новый метод.

Рассмотрим пространство с положительными и отрицательными примерами. Мы хотим их отделить друг от друга прямой линией. Но какую именно линию нужно провести? Это можно сделать по-разному, но линия не должна лежать слишком близко к примерам из одной группы. Линия, показанная на рисунке 2, проведена так, чтобы между положительными и отрицательными примерами был максимальный зазор.

Теперь нам нужно понять, какое формальное правило соответствует этой границе. Пусть имеется вектор $\bar{w}$ произвольной ненулевой длины, перпендикулярный разделяющей прямой, и вектор $\bar{u}$, обозначающий объект, подлежащий классификации. Мы хотим понять, с какой стороны от разделяющей прямой находится этот вектор. Для этого нам необходимо построить его проекцию на вектор $\bar{w}$, потому что длина этой проекции будет показывать, насколько близко находится неизвестный объект к положительным или отрицательным примерам.

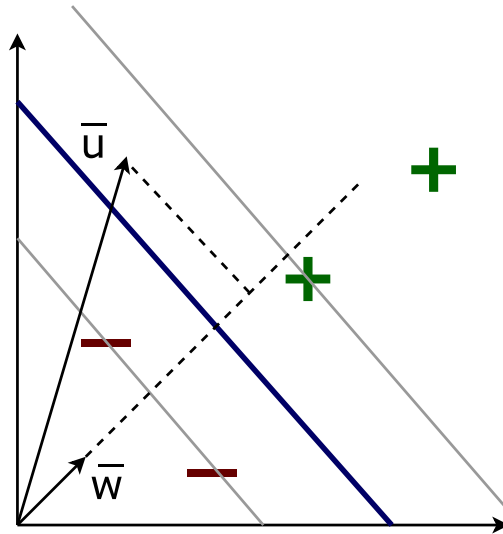


Рис. 2: Классификация методом опорных векторов

Таким образом, нужно проверить условие $\bar{w} \cdot \bar{u} \geq c$, где c — некоторая константа. Мы получили такое условие, потому что скалярное произведение двух векторов равно длине проекции одного вектора на другой. Если это условие выполняется, объект классифицируется как положительный.

Вывод уравнений

Для удобства преобразуем полученное условие к виду:

$$\bar{w} \cdot \bar{u} - b \geq 0 \quad (1)$$

Это и будет правилом, согласно которому будут классифицироваться объекты. Здесь $\bar{w}$ и b неизвестны, но мы знаем, что вектор $\bar{w}$ должен быть перпендикулярным к разделяющей прямой. Пока что у нас недостаточно ограничений, чтобы узнать точные значения этих параметров.

Для того, чтобы найти неизвестные, $\bar{w}$ и b , наложим дополнительные ограничения в следующем виде:

$$\begin{aligned} \bar{w} \cdot \bar{x}_+ - b &\geq 1 \\ \bar{w} \cdot \bar{x}_- - b &\leq -1 \end{aligned}$$

Здесь $\bar{x}_+$ — любой положительный пример, а $\bar{x}_-$ — отрицательный. Эти уравнения означают, что объекты из разных классов должны быть отделены друг от друга зазором.

Для удобства записи введём новую переменную y_i , равную 1, если i -й объект является положительным, и -1 в противном случае. Умножив оба уравнения на y_i , получим для

каждого примера одинаковое уравнение вида $y_i(\bar{w} \cdot \bar{x}_i - b) - 1 \geq 0$. Далее добавим дополнительные ограничения для объектов, которые лежат на границах зазора. Для них значение этого выражения должно быть равно нулю. Получим следующую систему уравнений:

$$\begin{cases} y_i(\bar{w} \cdot \bar{x}_i - b) - 1 \geq 0 \\ y_i(\bar{w} \cdot \bar{x}_i - b) - 1 = 0, \bar{x}_i \in gutter \end{cases} \quad (2)$$

Нашей финальной целью является нахождение прямой, разделяющей пространство таким образом, чтобы между положительными и отрицательными примерами был максимальный зазор. Попробуем выразить размер этого зазора. Если бы мы знали единичную нормаль к разделяющей прямой, то это расстояние можно было бы выразить как скалярное произведение единичной нормали и разности векторов $\bar{x}_+$ и $\bar{x}_-$ (см. рис. 3). Но, как было сказано выше, вектор $\bar{w}$ является нормалью, поэтому мы можем получить и единичную нормаль. Тогда выражение для ширины зазора принимает вид:

$$Width = (\bar{x}_+ - \bar{x}_-) \cdot \frac{\bar{w}}{\|\bar{w}\|}$$

Раскрывая скобки и подставляя равенство из системы (2), получим следующие преобразования:

$$Width = (\bar{x}_+ - \bar{x}_-) \cdot \frac{\bar{w}}{\|\bar{w}\|} = (\bar{x}_+ \cdot \bar{w} - \bar{x}_- \cdot \bar{w}) \frac{1}{\|\bar{w}\|} = (1 + b - (b - 1)) \frac{1}{\|\bar{w}\|} = \frac{2}{\|\bar{w}\|}$$

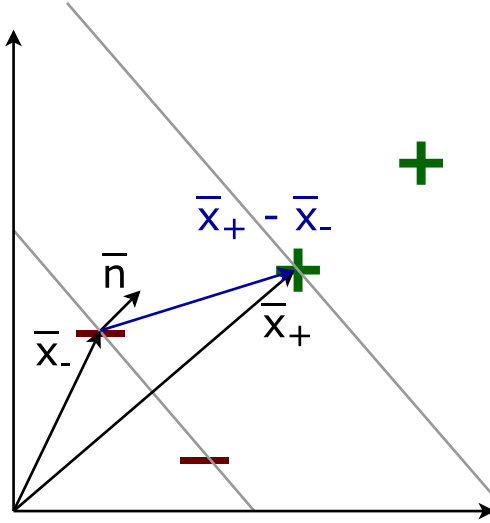


Рис. 3: Определение ширины зазора

Таким образом, мы получили простое выражение для ширины зазора:

$$Width = \frac{2}{\|\bar{w}\|} \quad (3)$$

Нам нужно максимизировать это значение, или максимизировать значение $\|\bar{w}\|$, но для удобства дальнейших вычислений мы перейдем к эквивалентной задаче поиска минимального значения функции $\frac{1}{2}\|\bar{w}\|^2$. Другими словами, необходимо найти решение следующей задачи:

$$\begin{cases} \frac{1}{2}\|\bar{w}\|^2 \rightarrow \min \\ y_i(\bar{w} \cdot \bar{x}_i - b) - 1 = 0, \bar{x}_i \in gutter \end{cases} \quad (4)$$

Для нахождения условного экстремума функции используется метод множителей Лагранжа, который позволяет получить новое выражение, для которого нужно будет найти экстремум, но уже без учёта ограничений. Применяв этот метод, задачу (4) можно свести к поиску экстремума следующей функции:

$$L = \frac{1}{2} \|\bar{w}\|^2 - \sum_i \alpha_i [y_i(\bar{w} \cdot \bar{x}_i - b) - 1] \quad (5)$$

Суммирование ограничений ведётся по всем примерам из обучающей выборки, α_i — некоторые константы. Для поиска экстремума функции возьмём частную производную по $\bar{w}$ и приравняем её к нулю:

$$\frac{\partial L}{\partial \bar{w}} = \bar{w} - \sum_i \alpha_i y_i \bar{x}_i = 0 \iff \bar{w} = \sum_i \alpha_i y_i \bar{x}_i \quad (6)$$

Мы получили, что вектор $\bar{w}$ является линейной комбинацией векторов из обучающей выборки. Некоторые из них в неё не войдут, поскольку для них коэффициент α_i будет равен нулю.

Далее подсчитаем частную производную по b :

$$\frac{\partial L}{\partial b} = - \sum_i \alpha_i y_i = 0 \iff \sum_i \alpha_i y_i = 0 \quad (7)$$

Подставим выражение, полученное для $\bar{w}$ в уравнении (6), в уравнение (5):

$$L = \frac{1}{2} \left(\sum_i \alpha_i y_i \bar{x}_i \right) \cdot \left(\sum_j \alpha_j y_j \bar{x}_j \right) - \sum_i \alpha_i y_i \bar{x}_i \cdot \left(\sum_j \alpha_j y_j \bar{x}_j \right) - \sum_i \alpha_i y_i b + \sum_i \alpha_i$$

С учётом уравнения (7) получим, что третья сумма равно нулю. Вторую сумму можно представить как произведение двух сумм, или как двойную сумму. Получим:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \bar{x}_i \cdot \bar{x}_j \quad (8)$$

Таким образом, нам удалось упростить исходное выражение (5). Далее решение экстремальной задачи (8) можно найти с помощью численных методов. Мы проделали эти упрощения для того, чтобы понять, как зависит поиск экстремума от векторов из обучающей выборки, и выяснили, что в результирующее выражение входит только скалярное произведение этих векторов.

Далее подставим выражение (6) для $\bar{w}$ в правило (1):

$$\sum_i \alpha_i y_i \bar{x}_i \cdot \bar{u} + b \geq 0 \quad (9)$$

Если выполняется условие (9), то неизвестный объект классифицируется как положительный. Заметим, что это условие также зависит только от скалярного произведения векторов.

Случай линейной неразделимости множеств

В случае, если множества из обучающей выборки линейно неразделимы, алгоритм будет работать с ошибкой даже на объектах этой выборки.

Эта проблема решается переходом в новое пространство, в котором множества можно будет линейно разделить. Пусть переход к новому пространству задаётся отображением $\phi(\bar{x})$.

Ранее было сказано, что функция, которую нужно максимизировать, зависит только от скалярного произведения векторов из обучающей выборки. То же самое верно и для правила, используемого для классификации. После трансформации пространства в выражения 8 и 9 будут входить множители $\phi(\bar{x}_i) \cdot \phi(\bar{x}_j)$ и $\phi(\bar{x}_i) \cdot \phi(\bar{u})$. Поэтому для решения задачи нам не нужно знать само отображение $\phi(x)$, а достаточно задать функцию $K(\bar{x}_i, \bar{x}_j) = \phi(\bar{x}_i) \cdot \phi(\bar{x}_j)$. Такая функция называется ядром. На практике чаще всего используются следующие ядра:

1. Полиномиальное ядро $K(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y} + 1)^n$;
2. Радиальная базисная функция $K(\bar{x}, \bar{y}) = e^{-\frac{\|\bar{x} - \bar{y}\|}{\sigma}}$.

Теперь у нас есть обобщённый метод, который работает на выпуклых множествах, поэтому всегда может найти глобальный экстремум. Конечно, этот метод не лишён недостатков. В частности, если в качестве ядра выбрать радиальную базисную функцию с достаточно малым значением параметра σ , то проявится проблема переобучения.

Историческая справка

Приведём краткую историю создания метода опорных векторов. Создаётся впечатление, что эта идея гораздо новее того, что мы уже изучили.

Владимир Вапник иммигрировал из СССР в США примерно в 1991 году. Никто до этого не слышал об идее опорных векторов, но он работал над этим методом ещё в рамках своей кандидатской диссертации в начале 1960-х в Московском Институте проблем управления.

Для реализации метода в то время не хватало вычислительной мощности, поэтому следующие 25 лет Вапник работал над прикладными проблемами онкологии в одном из институтов СССР. О нём узнали в Bell Labs и пригласили работать в США, куда он впоследствии иммигрировал.

В 1992 году Вапник предлагает к публикации три статьи в журнал NIPS (Neural Information Processing Systems), и все они были отклонены.

В 1993 году Bell Labs в заинтересовываются проблемой распознавания рукописного ввода и нейронными сетями. Вапник считал, что нейронные сети не очень хорошо покажут себя в этом, поэтому поспорил со своим коллегой, что метод опорных векторов сработает лучше. В итоге коллега, который как раз работал над распознаванием рукописного ввода, решает попробовать метод опорных векторов с полиномиальным ядром при $n = 2$, мало отличающийся от линейного, и он работает как по волшебству.

Использовали ли тогда ядро впервые? Эта идея появлялась ещё в диссертации Вапника, но он не придавал ей значения. Но как только эта идея сработала на задаче распознавания почерка, он оживил её и начал развивать, и в итоге она стала существенной частью всего метода опорных векторов.

Главная мысль заключается в том, что между концепцией идеи и моментом, когда о ней кто-то услышал, прошло около 30 лет. И так случается довольно часто: великие идеи появляются, затем о них надолго забывают, а потом случается переломный момент, когда исходная идея, лишь немного изменённая, приобретает огромную мощь. А Вапник,

о котором никто не слышал до начала 1990-х, стал знаменит благодаря идее, о которой сейчас знают все, кто интересуется машинным обучением.