

Лекция 7

Ядерные методы, метод опорных векторов

Лектор – Сенько Олег Валентинович

Курс «Математические основы теории прогнозирования»
4-й курс, III поток

- 1 Ядерные методы
- 2 Метод опорных векторов
- 2 Метод опорных векторов. Регрессия.

Напомним, что байесовское решающее правило или оптимальное решающее правило в смысле леммы Неймана-Пирсона могут быть легко восстановлены, если для каждого из распознаваемых классов K_1, \dots, K_L известны соответствующие плотности вероятности $p_1(\mathbf{x}), \dots, p_L(\mathbf{x})$. Ранее нами рассматривались методы восстановления плотностей $p_1(\mathbf{x}), \dots, p_L(\mathbf{x})$, основанные на гипотезе о независимости переменных X_1, \dots, X_n , а также на гипотезе о нормальности соответствующих распределений. Альтернативным подходом является использование ядерных методов восстановления плотности. Плотность вероятностей для каждого из классов K_1, \dots, K_L в точках пространства \mathbb{R}^n оценивается по объектам стандартной обучающей выборки $\tilde{S}_t = \{s_1 = (\alpha_1, \mathbf{x}_1), \dots, s_m = (\alpha_m, \mathbf{x}_m)\}$ с использованием ядерных функций.

Плотность вероятности для класса K_i в точке $\mathbf{x} \in \mathbb{R}^n$ вычисляется по формуле

$$p_i(\mathbf{x}) = \frac{1}{m_i} \sum_{s_j \in K_i} \mathfrak{K}(\mathbf{x}, \mathbf{x}_j),$$

где m_i - число объектов в обучающей выборке из класса K_i , $\mathfrak{K}(\mathbf{x}', \mathbf{x}'')$ является неотрицательной вещественнозначной функцией заданной, заданной на декартовом произведении $\mathbb{R}^n \times \mathbb{R}^n$. При этом

- \mathfrak{K} обладает свойством симметричности, то есть

$$\mathfrak{K}(\mathbf{x}', \mathbf{x}'') = \mathfrak{K}(\mathbf{x}'', \mathbf{x}');$$

- \mathfrak{K} достигает максимума при $\mathbf{x}' = \mathbf{x}''$;
- \mathfrak{K} убывает или по крайней мере не возрастает по мере увеличения расстояния между точками \mathbf{x}' и \mathbf{x}'' .

В одномерном случае, когда оценка плотности вероятности производится в точках интервала допустимых значений одного признака X по обучающей выборке $\{(\alpha_1, x_1), \dots, (\alpha_m, x_m)\}$, в качестве ядерных используются функции от x вида $\mathbf{K}(\frac{x-x_j}{h})$, где h - параметр сглаживания, характеризующий подробность аппроксимации. Для одномерных ядерных функций предполагается выполнение требований:

- максимальное значение $\mathbf{K}(u)$ должно достигаться в точке 0;
- должно выполняться свойство симметричности $\mathbf{K}(u) = \mathbf{K}(-u)$;
- $\mathbf{K}(u)$ монотонно убывающая или не возрастающая по мере удаления u от 0 функция;
- требуется выполнение равенства $\int_{-\infty}^{\infty} \mathbf{K}(u) du = 1$.

Очевидно, что такими свойствами обладают:

- прямоугольное ядро $\mathbf{K}(u) = \frac{1}{2}$ при $|u| \leq 1$ и $\mathbf{K}(u) = 0$ при $|u| > 1$;
- гауссиана $\mathbf{K}(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$.

Перечисленными свойствами обладает также ядро Епанечникова $\mathbf{K}(u) = \frac{3}{4}(1 - u^2)$ при $|u| \leq 1$ и $\mathbf{K}(u) = 0$ при $|u| > 1$. Параметр масштаба h характеризует подробность аппроксимации. При $h \rightarrow 0$ ядерный аппроксиматор подробно описывает детали эмпирического распределения объектов обучающей выборки. Однако устойчивость аппроксимации при этом теряется. При увеличении h возрастает устойчивость аппроксимации, но теряется способность улавливать детали.

Многомерные ядра размерности n могут задаваться в виде произведения одномерных ядер:

$$\mathfrak{K}(\mathbf{x}_*, \mathbf{x}_j) = \frac{1}{h^n} \prod_{i=1}^n \mathbf{K}\left(\frac{x_{*i} - x_{ji}}{h}\right).$$

Используются также ядровые функции, где связь между точками задаётся через расстояние между ними. Например, могут использоваться ядровые функции типа гауссианы

$$\mathcal{K}(\mathbf{x}_*, \mathbf{x}_j) = \sqrt{2\pi}\sigma^n e^{-\frac{(\mathbf{x}_* - \mathbf{x}_j)^2}{2\sigma^2}}.$$

Согласно формуле Байеса распознаваемый объект относится в класс, для которого величина $p_i(\mathbf{x})P(K_i)$ максимальна. Параметр сглаживания h может подбираться таким образом, чтобы точность распознавания была максимальной. Например, может максимизироваться оценка точности на контрольной выборке или в оценке скользящего контроля.

Метод опорных векторов является универсальным методом распознавания, позволяющим наряду с линейными реализовывать также нелинейные решающие правила. Исходный вариант метода был предложен для задач с двумя распознаваемыми классами K_1 и K_2 . В случаях, когда объекты разных классов в обучающей выборке линейно разделимы, обычно существует целая совокупность линейных поверхностей, осуществляющих такое разделение. На рисунке представлены двумерные данные, где объекты двух классов могут быть разделены с помощью прямых A, B, C, D. Однако наша интуиция, подсказывает что наилучшей обобщающей способностью должна обладать разделяющая прямая F, одинаково удалённая от групп объектов из разных классов.

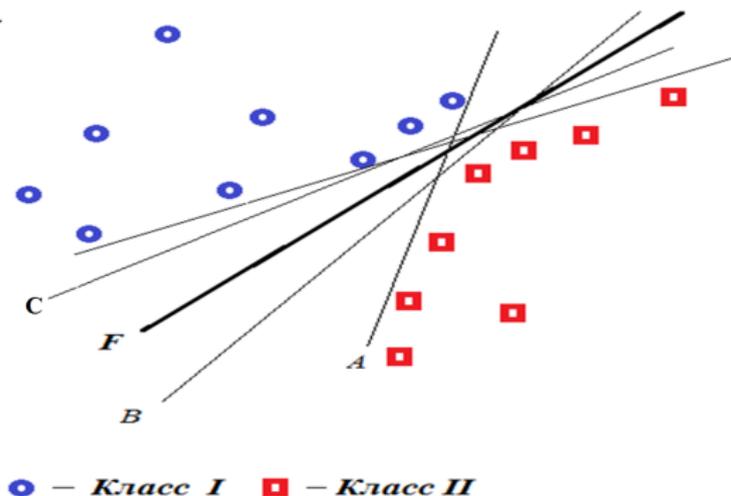


Рис.1

Интуитивные представления об оптимальной разделимости формализует проведение разделяющей гиперплоскости посередине между двумя параллельными гиперплоскостями, каждая из которых отделяет объекты одного из классов.

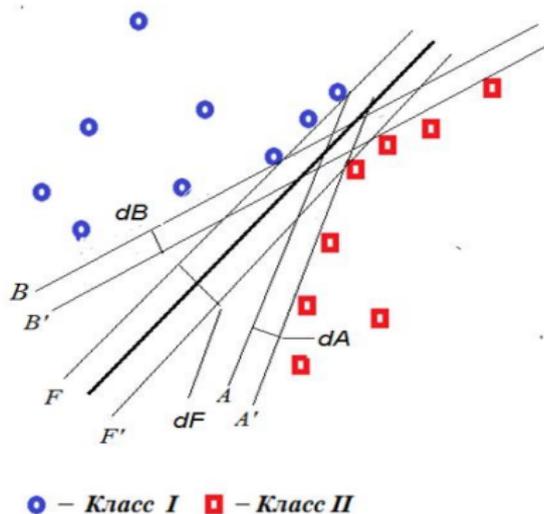


Рис.2

При этом две плоскости строятся таким образом, чтобы расстояние «зазор» между ними был бы максимальным. Из рисунка видно, что наибольшим является «зазор» между двумя параллельными прямыми F и F' .

Напомним, что пара параллельных гиперплоскостей P_1 и P_2 в многомерном пространстве \mathbb{R}^n описывается с помощью уравнений

$$P_1 \rightarrow \mathbf{w}\mathbf{x}^t = b_1 \quad (1)$$

$$P_2 \rightarrow \mathbf{w}\mathbf{x}^t = b_2$$

От системы (1) нетрудно перейти к эквивалентной системе

$$z\mathbf{x}^t = b + 1 \quad (2)$$

$$z\mathbf{x}^t = b - 1,$$

описывающей те же самые гиперплоскости. Расстояние (величина зазора) δ между гиперплоскостями P_1 и P_2 равно $\frac{2}{|z|}$.

Следовательно задача поиска двух максимально удалённых друг от друга параллельных гиперплоскостей, каждая из которых отделяет объекты одного из классов, может быть сведена к оптимизационной задаче с ограничениями

$$\delta = \frac{2}{|z|} \rightarrow \max \quad (3)$$

$$\begin{aligned} z x_j^t &\geq b + 1 \text{ при } s_j \in K_1 \cap \tilde{S}_t, \\ z x_j^t &\leq b - 1 \text{ при } s_j \in K_2 \cap \tilde{S}_t. \end{aligned}$$

При этом оптимизация производится по компонентам направляющего вектора $z = (z_1, \dots, z_n)$ и параметру сдвига b . Введём обозначение

$$\begin{aligned} \alpha_j &= 1 \text{ при } s_j \in K_1 \text{ и} \\ \alpha_j &= -1 \text{ при } s_j \in K_2 \end{aligned}$$

Тогда задача (3) оказывается эквивалентна задаче

$$\frac{1}{2} \sum_{i=1}^n z_i^2 \rightarrow \min \quad (4)$$

$$\alpha_j (\mathbf{z} \mathbf{x}_j^t - b) \geq 1, j = 1, \dots, m$$

Из известной теоремы Каруша-Куна-Такера (ККТ) следует, что для произвольной точки (\mathbf{z}^*, b^*) , в которой $\sum_{i=1}^n z_i^2$ достигает своего минимума при ограничениях задачи (4), и некоторого вектора неотрицательных множителей Лагранжа $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ соблюдаются условия стационарности лагранжиана

$$L(\mathbf{z}, b, \boldsymbol{\lambda}) = \frac{1}{2} \sum_{i=1}^n z_i^2 - \sum_{j=1}^m \lambda_j [\alpha_j (\mathbf{z} \mathbf{x}_j^t - b) - 1]$$

Также из теоремы ККТ следует необходимость выполнения m равенств, которые носят название условий дополняющей нежёсткости

$$\lambda_j [\alpha_j (\mathbf{z}^* \mathbf{x}_j^t - b^*) - 1] = 0, j = 1, \dots, m$$

Из условия стационарности следует, что

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda})}{\partial z_i} \Big|_{(\mathbf{z}^*, b^*)} = z_i^* - \sum_{j=1}^m \lambda_j \alpha_j x_{ji} = 0 \quad (5)$$

В векторной форме система (5) принимает вид

$$\mathbf{z}^* = \sum_{j=1}^m \lambda_j \alpha_j \mathbf{x}_j$$

Из условия стационарности следует выполнение равенства

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda})}{\partial b} = \sum_{j=1}^m \lambda_j \alpha_j = 0 \quad (6)$$

Таким образом для получения решения задачи (4) достаточно найти требуемый вектор значений неотрицательных множителей Лагранжа $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*)$.

Поиск оптимальных значений множителей Лагранжа.

Нетрудно показать, что лагранжиан в произвольной точке (z, b, λ) , для которой соблюдаются условия (5,6), может быть записан в виде

$$L(z, b, \lambda) = g(\lambda) = \sum_{j=1}^m \lambda_j - \frac{1}{2} \sum_{j'=1}^m \sum_{j''=1}^m \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} (\mathbf{x}_{j'} \mathbf{x}_{j''}^t).$$

Отметим, что в силу соблюдения ограничений задачи (4) и неотрицательности множителей Лагранжа выполняется неравенство

$$g(\lambda) \leq \sum_{i=1}^n z_i^2.$$

Из теории оптимизации с ограничениями следует, что при соблюдении ряда условий, которые справедливы для задачи (4), максимум $g(\boldsymbol{\lambda})$ по вектору множителей Лагранжа при условии их неотрицательности и при соблюдении равенства $\sum_{j=1}^m \lambda_j \alpha_j = 0$ равен $\frac{1}{2} \sum_{i=1}^n (z_i^*)^2$. Однако в силу соблюдения условий стационарности и условий дополняющей нежёсткости справедливо равенство

$$L(\mathbf{z}^*, \mathbf{b}^*, \boldsymbol{\lambda}) = g(\boldsymbol{\lambda}^*) = \frac{1}{2} \sum_{i=1}^n (z_i^*)^2.$$

То есть точка $\boldsymbol{\lambda}^*$ является точкой максимума $g(\boldsymbol{\lambda})$ при условии неотрицательности $(\lambda_1, \dots, \lambda_m)$ и при соблюдении равенства $\sum_{j=1}^m \lambda_j \alpha_j = 0$.

Таким образом оказывается, что оптимальные значения множителей множителей Лагранжа $(\lambda_1, \dots, \lambda_m)$ могут быть найдены как решение двойственной задачи квадратичного программирования:

$$\sum_{j=1}^m \lambda_j - \frac{1}{2} \sum_{j'=1}^m \sum_{j''=1}^m \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} (\mathbf{x}_{j'} \mathbf{x}_{j''}^t) \rightarrow \max \quad (7)$$

$$\sum_{j=1}^m \lambda_j \alpha_j = 0$$

$$\lambda_j \geq 0, j = 1, \dots, m$$

Пусть $(\hat{\lambda}_1, \dots, \hat{\lambda}_m)$ - решение задачи (7). Направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\sum_{j=1}^m \hat{\lambda}_j \alpha_j \mathbf{x}_j$. То есть направляющий вектор разделяющей гиперплоскости является линейной комбинацией векторных описаний объектов обучающей выборки, для которых значения соответствующих оптимальных множителей Лагранжа отличны от 0. Такие векторные описания принято называть опорными векторами. Пусть

$$J_0 = \{j = 1, \dots, m \mid [\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) - 1] \neq 0\}$$

Из условий дополняющей нежёсткости видно, при $j \in J_0$ обязательно должно выполняться равенство $\hat{\lambda}_j = 0$. Поэтому векторное описание \mathbf{x}_j соответствующего объекта обучающей выборки является опорным вектором, если j не принадлежит J_0 . Оценка параметра сдвига \hat{b} находится из ограничения, соответствующего произвольному опорному вектору.

Действительно, из условий дополняющей нежёсткости следует, что при произвольном j , когда $\lambda_j > 0$, обязательно должно выполняться равенство $\alpha_j(z\mathbf{x}_j^t - b) - 1 = 0$, эквивалентное равенству $b = z\mathbf{x}_j^t - \alpha_j$. Классификация нового распознаваемого объекта s с описанием \mathbf{x} вычисляется согласно знаку выражения

$$g(\mathbf{x}) = \sum_{j=1}^m \hat{\lambda}_j \alpha_j (\mathbf{x}_j \mathbf{x}^t) - \hat{b}. \quad (8)$$

Объект s относится к классу K_1 , если $g(\mathbf{x}) > 0$ и объект s относится к классу K_2 в противном случае.

Существенным недостатком рассмотренного варианта метода опорных векторов является требование линейной разделимости классов.

Однако данный недостаток может быть легко преодолен с помощью следующей модификации, основанной на использовании дополнительного вектора неотрицательных переменных

$$\boldsymbol{\xi} = (\xi_1, \dots, \xi_m).$$

Метод опорных векторов. Случай отсутствия линейной разделимости.

Требования об отделимости классов (3) заменяются более мягкими требованиями:

$$\begin{aligned} \mathbf{z} \mathbf{x}_j^t &\geq b + 1 - \xi_j \text{ при } s_j \in K_1 \cap \tilde{S}_t, \\ \mathbf{z} \mathbf{x}_j^t &\leq b - 1 + \xi_j \text{ при } s_j \in K_2 \cap \tilde{S}_t. \end{aligned}$$

Выдвигается также требование минимальности суммы $\sum_{j=1}^m \xi_j$. Поиск оптимальных параметров разделяющей гиперплоскости при отсутствии линейной разделимости таким образом сводится к решению задачи квадратично программирования

$$\frac{1}{2} \sum_{i=1}^n z_i^2 + C \sum_{j=1}^m \xi_j \rightarrow \min \quad (9)$$

$$\alpha_j (\mathbf{z} \mathbf{x}_j^t - b) \geq 1 - \xi_j, \xi_j \geq 0, j = 1, \dots, m$$

где α_j - некоторая положительная константа, являющаяся открытым параметром алгоритма. Иными словами оптимальное значение C подбирается пользователем.

метод опорных векторов. Случай отсутствия линейной делимости.

Из теоремы ККТ следует, что для произвольной точки (z^*, b^*, ξ^*) , в которой достигается минимум функционала

$$\frac{1}{2} \sum_{i=1}^n z_i^2 + C \sum_{j=1}^m \xi_j$$

при справедливости ограничений (9), и некоторых векторов неотрицательных множителей Лагранжа $\lambda = (\lambda_1, \dots, \lambda_n)$ и $\eta = (\eta_1, \dots, \eta_m)$ соблюдаются условия стационарности лагранжиана

$$L(z, b, \lambda, \xi, \eta) = \frac{1}{2} \sum_{i=1}^n z_i^2 - \sum_{j=1}^m \lambda_j [\alpha_j (z x_j^t - b) - 1 + \xi_j] - \sum_{j=1}^m \eta_j \xi_j$$

Метод опорных векторов. Случай отсутствия линейной делимости.

Данные условия записываются в виде

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\eta})}{\partial z_i} \Big|_{z_i^* = z_i^*} - \sum_{j=1}^m \lambda_j \alpha_j x_{ji} = 0 \quad (10)$$

$$i = 1, \dots, n$$

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\eta})}{\partial b} = \sum_{j=1}^m \lambda_j \alpha_j = 0,$$

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\eta})}{\partial \xi_j} = C - \lambda_j - \eta_j = 0,$$

$$j = 1, \dots, m.$$

Метод опорных векторов. Случай отсутствия линейной делимости.

Также выполняются условия дополняющей нежёсткости

$$\begin{aligned}\lambda_j[\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) - 1 + \xi_j] &= 0, \\ \eta_j\xi_j &= 0, \\ j &= 1, \dots, m\end{aligned}\tag{11}$$

Оптимальные значения множителей $(\lambda_1, \dots, \lambda_m)$ могут быть найдены как решение двойственной задачи квадратичного программирования

$$\begin{aligned}\sum_{j=1}^m \lambda_j - \frac{1}{2} \sum_{j'=1}^m \sum_{j''=1}^m \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} (\mathbf{x}_{j'} \mathbf{x}_{j''}^t) &\rightarrow \max \\ \sum_{j=1}^m \lambda_j \alpha_j &= 0\end{aligned}\tag{12}$$

$$C \geq \lambda_j \geq 0, j = 1, \dots, m$$

Метод опорных векторов. Случай отсутствия линейной делимости.

Как и в случае линейной делимости направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\hat{\mathbf{z}} = \sum_{i=1}^m \hat{\lambda}_i \alpha_i \mathbf{x}_i$. Из условий $C - \lambda_j - \eta_j$ и $\eta_j \xi_j = 0$ следует что $\eta_j > 0$ и $\xi_j = 0$ при $0 < \lambda_j < C$. Также как и в случае существования линейной делимости параметра сдвига \hat{b} находится из ограничения, соответствующего произвольному опорному вектору. Действительно, из условий дополняющей нежёсткости и и следующего из них равенства $\xi_j = 0$ следует выполнение равенства $\alpha_j (\mathbf{z} \mathbf{x}_j^t - b) - 1 = 0$, эквивалентного равенству $b = \mathbf{z} \mathbf{x}_j^t - \alpha_j$. Распознавание нового объекта s производится по его описанию \mathbf{x} также как и в случае линейно делимых классов с помощью решающего правила (8) по величине распознающей функции $g(\mathbf{x})$.

Метод опорных векторов. Построение нелинейных разделяющих поверхностей

Таким образом построение оптимального решающего правила сводится к решению двойственных задач квадратичного программирования (7) или (12). Следует отметить, что вектора описаний объектов обучающей выборки $\{\mathbf{x}_j \mid j = 1, \dots, m\}$ входят в задачи (7),(12) только через свои скалярные произведения $\mathbf{x}_{j'} \mathbf{x}_{j''}^t$ при $\lambda_{j'} > 0$ и $\lambda_{j''} > 0$. Аналогично при вычислении значения распознающей функции (8) по описанию распознаваемого объекта \mathbf{x} на самом деле используются только скалярные произведения $\mathbf{x} \mathbf{x}_{j''}^t$. Предположим что в исходном признаковом пространстве эффективное линейное разделение отсутствует. Однако может существовать такое евклидово пространство H_y и такое отображение Φ из области пространства \mathbb{R}^n , содержащей описания распознаваемых объектов, в пространство H_y , что образы объектов обучающей выборки из классов K_1 и K_2 оказываются разделимыми с помощью некоторой гиперплоскости P_y . Пусть $\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ - образы в пространстве H_y векторов описаний объектов обучающей выборки $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$.

Метод опорных векторов. Построение нелинейных разделяющих поверхностей

Линейная разделимость означает существование решения аналога задачи квадратичного программирования (4) для пространства H_y , которое сводится к решению двойственной задачи

$$\sum_{j=1}^m \lambda_j - \frac{1}{2} \sum_{j'=1}^m \sum_{j''=1}^m \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} (\mathbf{y}_{j'} \mathbf{y}_{j''}^t) \rightarrow \max \quad (13)$$

$$\sum_{j=1}^m \lambda_j \alpha_j = 0$$

$$\lambda_j \geq 0, j = 1, \dots, m$$

Отметим, что необходимость полного восстановления преобразования $\Phi(\mathbf{x})$ для поиска всех коэффициентов задачи квадратичного программирования (13) отсутствует. Достаточно найти функцию, связывающую скалярное произведение $\mathbf{y}_{j'} \mathbf{y}_{j''}^t$ с векторами $\mathbf{x}_{j'}$ и $\mathbf{x}_{j''}$, где $\mathbf{y}_{j'} = \Phi(\mathbf{x}_{j'})$ и $\mathbf{y}_{j''} = \Phi(\mathbf{x}_{j''})$.

Метод опорных векторо. Построение нелинейных разделяющих поверхностей

Такую функцию мы далее будем называть потенциальной и обозначать $\mathcal{K}(\mathbf{x}', \mathbf{x}'')$. Можно подобрать потенциальную функцию таким образом, чтобы решение (13) было оптимальным. При этом поиск оптимальной потенциальной функции может производиться внутри некоторого заранее заданного семейства. Например, потенциальную функцию можно задать с помощью простого сдвига

$$\mathcal{K}(\mathbf{x}', \mathbf{x}'') = \mathbf{x}'(\mathbf{x}'')^t + \theta, \quad (14)$$

Решение, полученное путём замены скалярных произведений на потенциальные функции, может рассматриваться как построении линейной разделяющей поверхности в трансформированном пространстве, если удаётся доказать существование отображения Φ , для которого при произвольных \mathbf{x}' и \mathbf{x}'' из \mathbb{R}^n выполняется равенство

$$\mathcal{K}(\mathbf{x}', \mathbf{x}'') = \Phi(\mathbf{x}')\Phi^t(\mathbf{x}''). \quad (15)$$

Метод опорных векторов. Построение нелинейных разделяющих поверхностей

Существование преобразования Φ , для которого выполняется равенство (15), было показано для неотрицательных симметричных потенциальных функций вида

$$\mathfrak{K}(\mathbf{x}', \mathbf{x}'') = (\mathbf{x}'(\mathbf{x}'')^t + \theta)^2,$$

, где $d \geq 1$ -целое число а также ядерных функции типа гауссианы

$$\mathfrak{K}(\mathbf{x}', \mathbf{x}'') = \sqrt{2\pi}\sigma^n e^{-\frac{(\mathbf{x}' - \mathbf{x}'')^2}{2\sigma^2}},$$

где σ - вещественная неотрицательная константа (размер ядра). Поскольку в общем случае преобразование является нелинейным, то прообразом в пространстве \mathbb{R}^n линейной разделяющей гиперплоскости, существующей в пространстве H_y , может оказаться нелинейная поверхность.

Метод опорных векторов. Построение нелинейных разделяющих поверхностей

Для большого числа прикладных задач линейная разделимость является недостижимой. Поэтому выбор ядерной функции может производиться из требования о минимальности числа ошибок в смысле задачи квадратичного программирования (9). На практике подбор ядерных функций и их параметров производится исходя из требования достижения максимальной обобщающей способности, которая оценивается с помощью скользящего контроля или оценок на контрольной выборке.

Методика улучшения обобщающей способности, лежащая в основе Метода опорных векторов (МОВ) может быть распространена также на задачи регрессии, то есть на задачи прогнозирования некоторой переменной Y , принимающей значения из интервала вещественной оси по значениями вещественных переменных X_1, \dots, X_n . Вместо требования максимизации величины «зазора» между распознаваемыми классами для задач распознавания в случае задач регрессии выдвигается требование минимизации вариации прогнозирующей функции на области \tilde{X} , из которой принимают значения переменные X_1, \dots, X_n . Уменьшение вариации прогнозирующей функции очевидно позволяет снизить вариационную составляющую обобщённой ошибки прогнозирования и уменьшить эффект переобучения. Предположим, что прогнозом величины Y являются значения функции $f(\mathbf{x})$.

Задача снижения вариации функции f , формализуется как задача максимизации параметра

$$\delta_\varepsilon = \inf_{(\mathbf{x}', \mathbf{x}'') \in \tilde{X}_\varepsilon^C} (|\mathbf{x}' - \mathbf{x}''|) \quad (16)$$

где ε - некоторый пороговый параметр;

$$\tilde{X}_\varepsilon^C = \{(\mathbf{x}', \mathbf{x}'') \in \tilde{X} \times \tilde{X} \mid |f(\mathbf{x}') - f(\mathbf{x}'')| > 2\varepsilon\}$$

Предположим, что регрессия является линейной, то есть $f(\mathbf{x}) = \beta \mathbf{x}^t + \beta_0$, где $\beta = (\beta_1, \dots, \beta_n)$ - вектор регрессионных коэффициентов, β_0 - параметр сдвига. Откуда следует, что

$$|f(\mathbf{x}') - f(\mathbf{x}'')| = |\beta(\mathbf{x}' - \mathbf{x}'')^t|$$

→

Очевидно, что минимум $\|x' - x''\|$ достигается для пары векторов из \tilde{X}_ε^C , для которой

- справедливо равенство $\|\beta(x' - x'')^t\| = 2\varepsilon$;
- вектор $(x' - x'')$ совпадает по направлению с вектором β .

В результате должно выполняться равенство

$$\|\beta\| \delta_\varepsilon = 2\varepsilon,$$

эквивалентное равенству

$$\delta_\varepsilon = \frac{2\varepsilon}{\|\beta\|}.$$

Наряду с требованиями максимизации параметра δ_ε выдвигается также требование точности аппроксимации на обучающей выборке.

Отклонение прогнозирующей функции $f(x)$ от значений прогнозируемой величины Y не должно превышать порогового параметра ε . Отметим, что задача максимизации $\delta_\varepsilon = \frac{2\varepsilon}{|\beta|}$ полностью эквивалентна задаче минимизации $\frac{1}{2} \sum_{i=1}^n \beta_i^2$.
В результате мы переходим к задаче квадратичного программирования

$$\frac{1}{2} \sum_{i=1}^n \beta_i^2 \rightarrow \min \quad (17)$$

$$y_j - \beta_0 - \beta \mathbf{x}_j^t \leq \varepsilon$$

$$\beta_0 + \beta \mathbf{x}_j^t - y_j \leq \varepsilon, 1, \dots, m$$

Решение задачи (17) может отсутствовать, если не будет найден вектор β , при котором справедливы ограничения (17).

Поэтому от задачи (17) переходим к задаче, допускающей существование решений в произвольном случае:

$$\frac{1}{2} \sum_{i=1}^n \beta_i^2 + C \sum_{j=1}^m (\xi_j^1 + \xi_j^2) \rightarrow \min \quad (18)$$

$$y_j - \beta_0 - \beta \mathbf{x}_j^t \leq \varepsilon + \xi_j^1$$

$$\beta_0 + \beta \mathbf{x}_j^t - y_j \leq \varepsilon + \xi_j^2, j = 1, \dots, m$$

где $(\xi_j^1, \xi_j^2), j = 1, \dots, m$ - неотрицательные коэффициенты, имеющие тот же самый смысл, что и аналогичные коэффициенты в задаче (9).

Параметр C - неотрицательный штрафной коэффициент.

Для решения задачи квадратичного программирования (18) используются методы, аналогичные тем, которые используются для решения задачи квадратичного программирования (9), лежащей в основе процедуры обучения алгоритмов распознавания.

Подобно тому как вариант МОВ для решения задач распознавания допускает расширение на случаи с линейно неотделимыми классами и принципиально позволяет строить нелинейные разделяющие поверхности, вариант МОВ для решения задач регрессионного анализа допускает расширение на задачи, в которых присутствуют выпадающие наблюдения, а также позволяет строить нелинейные прогнозирующие функции.