Лекция 7

метод опорных векторов

Лектор – *Сенько Олег Валентинович*

Математические методы обучения

Содержание лекции

метод опорных векторов

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 раделены 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.

Метод опорных векторов

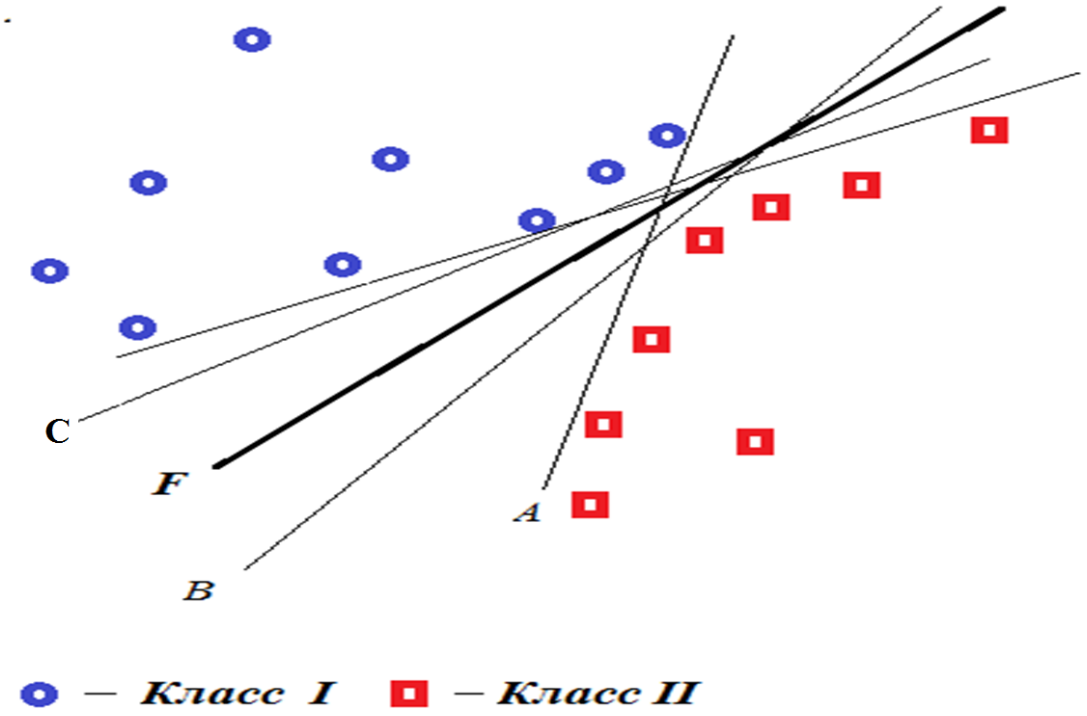

Рис.1

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

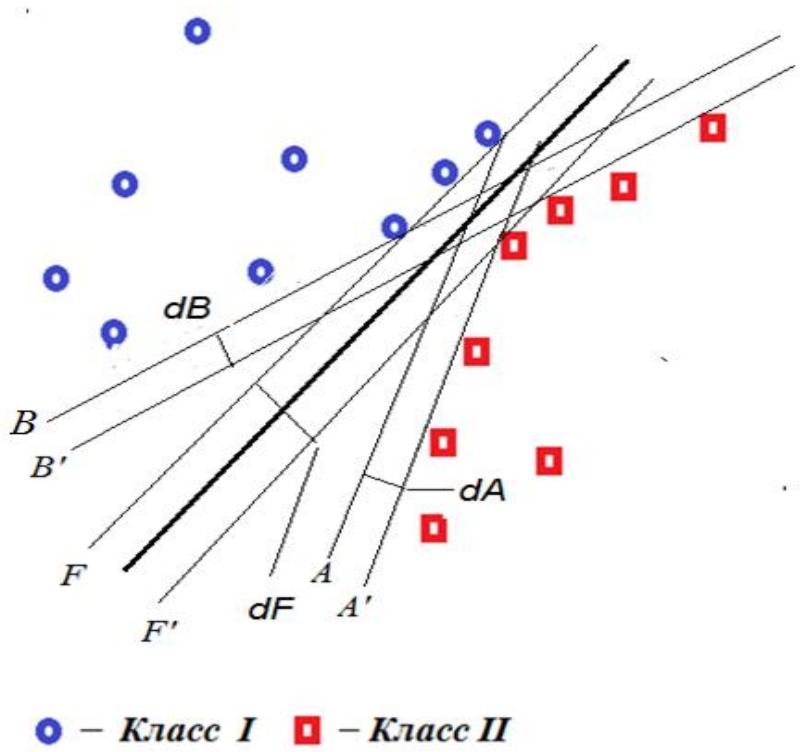

Рис.2

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

Метод опорных векторов

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

$$P_1 \to \boldsymbol{w}\boldsymbol{x}^t = b_1 \tag{1}$$

$$P_2 \to \boldsymbol{w}\boldsymbol{x}^t = b_2$$

От системы (1) нетрудно перейти к эквивалентной системе

$$\boldsymbol{z}\boldsymbol{x}^t = b + 1 \tag{2}$$

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

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

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

$$\delta = \frac{2}{|\,\boldsymbol{z}\,|} \to \max \tag{3}$$

$$\boldsymbol{z}\boldsymbol{x}_j^t \geq b + 1 \text{ при } s_j \in K_1 \bigcap \widetilde{S}_t,$$
$$\boldsymbol{z}\boldsymbol{x}_j^t \leq b - 1 \text{ при } s_j \in K_2 \bigcap \widetilde{S}_t.$$

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

$$\alpha_j = 1 \text{ при } s_j \in K_1 \text{ и}$$
$$\alpha_j = -1 \text{ при } s_j \in K_2$$

Тогда задача (3) оказывается эквивалентна задаче

$$\frac{1}{2}\sum_{i=1}^{n} z_i^2 \to \min \tag{4}$$

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

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

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

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

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

Из условия стационарности следует, что

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

В векторной форме система (5) принимает вид

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

,

Из условия стационарности следует выполнение равенства

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

Нетрудно показать, воспользовавшись уравнениями (5,6), что лагранжиан в точке может быть записан в виде . Отметим, что в силу соблюдения ограничений задачи (4) и неотрицательности множителей Лагранжа в точке выполняется неравенство

Из теории оптимизации следует, что оптимальные значения множителей множителей Лагранжа $(\lambda_1, \ldots, \lambda_m)$ могут быть найдены как решение двойственной задачи квадратичного программирования:

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

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

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

Метод опорных векторов

Пусть $(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$ - решение задачи (7) . Направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\sum_{j=1}^{m} \hat{\lambda}_j \alpha_j \boldsymbol{x}_j$ 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

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

Из условий дополняющей нежёсткости видно, при $j \in J_0$ обязательно должно выполняться равенство $\hat{\lambda}_j = 0$. Поэтому векторное описание $\boldsymbol{x}_j$ 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.

Метод опорных векторов

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

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

Объект s относится к классу K_1, если $g(\boldsymbol{x})$ 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 преодолён с помощью следующей модификации, основанной на использовании дополнительного вектора неотрицательных переменных $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$.

Метод опорных векторов. Случай отсутствия линейной разделимости.

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

$$\boldsymbol{z}\boldsymbol{x}_j^t \geq b + 1 - \xi_j \text{ при } s_j \in K_1 \bigcap \widetilde{S}_t,$$
$$\boldsymbol{z}\boldsymbol{x}_j^t \leq b - 1 + \xi_j \text{ при } s_j \in K_2 \bigcap \widetilde{S}_t.$$

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

$$\frac{1}{2}\sum_{i=1}^{n} z_i^2 + C\sum_{j=1}^{m} \xi_j \to \min \tag{9}$$

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

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

метод опорных векторов. Случай отсутствия линейной разделимости.

Из теоремы KKT следует, что для произвольной точки $(\boldsymbol{z}^*, b^*, \boldsymbol{\xi}^*)$, 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 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ и $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$ соблюдаются условия стационарности лагранжиана

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

Метод опорных векторов. Случай отсутствия линейной разделимости.

Данные условия записываются в виде

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

$$i = 1, \ldots, n$$

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

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

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

Метод опорных векторов. Случай отсутствия линейной разделимости.

Также выполняются условия дополняющей нежёсткости

$$\lambda_j[\alpha_j(\boldsymbol{z}\boldsymbol{x}_j^t - b) - 1 + \xi_j] = 0, \tag{11}$$

$$\eta_j\xi_j = 0,$$

$$j = 1, \ldots, m$$

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

$$\sum_{j=1}^{m}\lambda_j - \frac{1}{2}\sum_{j'=1}^{m}\sum_{j''=1}^{m}\lambda_{j'}\lambda_{j''}\alpha_{j'}\alpha_{j''}(\boldsymbol{x}_{j'}\boldsymbol{x}_{j''}^t) \to \max \tag{12}$$

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

$$C > \lambda_j \geq 0, j = 1, \ldots, m$$

Метод опорных векторов. Случай отсутствия линейной разделимости.

Как и в случае линейной разделимости направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\hat{\boldsymbol{z}} = \sum_{i=1}^{m} \hat{\lambda}_j \alpha_j \boldsymbol{x}_j$. Из условий $C - \lambda_j - \eta_j = 0$ 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 разделимости 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 $\alpha_j(\boldsymbol{z}\boldsymbol{x}_j^t - b) - 1 = 0$, эквивалентного равенству $b = \boldsymbol{z}\boldsymbol{x}_j^t - \alpha_j$. Распознавание нового объекта s производится по его описанию $\boldsymbol{x}$ также как и в случае линейно разделимых классов с помощью решающего правила (8) по величине распознающей функции $g(\boldsymbol{x})$.

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 $\{\boldsymbol{x}_j \mid j = 1, \ldots, m\}$ входят в задачи (7),(12) только через свои скалярные произведения $\boldsymbol{x}_{j'}\boldsymbol{x}_{j''}^t$ при $\lambda_{j'} > 0$ и $\lambda_{j''} > 0$. Аналогично при вычислении значения распознающей функции (8) по описанию распознаваемого объекта $\boldsymbol{x}$ на самом деле используются только скалярные произведения $\boldsymbol{x}\boldsymbol{x}_{j''}^t$. 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 $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$ -образы в пространстве H_y векторов описаний объектов обучающей выборки $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$.

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

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

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

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

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

Метод опорных векторо. Построение нелинейных разделяющих поверхностей

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

$$\mathfrak{K}(\boldsymbol{x}', \boldsymbol{x}'') = \boldsymbol{x}'(\boldsymbol{x}'')^t + \theta, \tag{14}$$

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 $\boldsymbol{x}'$ и $\boldsymbol{x}''$ из $\mathbb{R}^n$ выполняется равенство

$$\mathfrak{K}(\boldsymbol{x}', \boldsymbol{x}'') = \Phi(\boldsymbol{x}')\Phi^t(\boldsymbol{x}''). \tag{15}$$

Метод опорных векторов. Построение нелинейных разделяющих поверхностей

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

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

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

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

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 векторо. Построение нелинейных разделяющих поверхностей

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

Метод опорных векторов. Регрессия.

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 $X_1, \ldots, X_n$. Вместо требования максимизации величины «зазора» между распознаваемыми классами для задач распознавания в случае задач регрессии выдвигается требование минимизации вариации прогнозирующей функции на области $\widetilde{X}$, из которой принимают значения переменные $X_1, \ldots, X_n$ Уменьшение вариации прогнозирующей функции очевидно позволяет снизить вариационную составляющую обобщённой ошибки прогнозирования и уменьшить эффект переобучения. Предположим, что прогнозом величины Y являются значения функции $f(\boldsymbol{x})$.

Задача снижения вариации функции f, формализуется как задача максимизации параметра

$$\delta_\varepsilon = \inf_{(\boldsymbol{x}', \boldsymbol{x}'') \in \widetilde{X}_\varepsilon^C} (\mid \boldsymbol{x}' - \boldsymbol{x}'' \mid) \tag{16}$$

где ε - некоторый пороговый параметр;

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

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

$$\mid f(\boldsymbol{x}') - f(\boldsymbol{x}'') \mid = \mid \boldsymbol{\beta}(\boldsymbol{x}' - \boldsymbol{x}'')^t \mid$$

Очевидно, что минимум $| \boldsymbol{x}' - \boldsymbol{x}'' |$ достигается для пары векторов из $\widetilde{X}_\varepsilon^C$, для которой

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

В результате должно выполняться равенство

$$| \boldsymbol{\beta} | \delta_\varepsilon = 2\varepsilon,$$

эквивалентное равенству

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

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

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

$$\frac{1}{2}\sum_{i=1}^{n}\beta_i^2 \to \min \tag{17}$$

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

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

Решение задачи (17) может отсутствовать, если не будет найден вектор $\boldsymbol{\beta}$, при котором справедливы ограничения (17).

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

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

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

$$\beta_0 + \boldsymbol{\beta}\boldsymbol{x}_j^t - y_j \le \varepsilon + \xi_j^2, j = 1, \ldots, m$$

где $(\xi_j^1, \xi_j^2), j = 1, \ldots, m$ - 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.

Метод опорных векторо. Регрессия.

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.