МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор
Сенько Олег Валентинович

Лекция 7

Ядерные методы

Напомним, что байесовское решающее правило или оптимальное решающее правило в смысле леммы Неймана-Пирсона могут быть легко восстановлены, если для каждого из распознаваемых классов $K_1,\ldots,K_L$ известны соответствующие плотности вероятности $f_1(\mathbf{x}),\ldots, f_L(\mathbf{x})$. Ранее нами рассматривался метод восстановления плотностей $f_1(\mathbf{x}),\ldots, f_L(\mathbf{x})$, основанный на гипотезе о нормальности соответствующих распределений. Альтернативным подходом является использование ядерных методов восстановления плотности.

Ядерные методы

Ядерные методы восстановления плотности в многомерном пространстве основаны на использовании так называемых ядровых функций. Ядровая функция с центром в точке $\mathbf{x}_j \in \mathbf{R}^n$ обычно записывается в форме $\frac{1}{h^n}\mathcal{K}(\frac{\mathbf{x}-\mathbf{x}_j}{h})$. При этом выдвигается требование $\int_{\mathbf{R}^n}[\frac{1}{h^n}\mathcal{K}(\frac{\mathbf{x}-\mathbf{x}_j}{h})]d\mathbf{x}=1$, где h - параметр сглаживания. В качестве ядровых функции может быть использовано многомерное ядро Гаусса

$$\mathcal{K}_{\mathcal{N}}(\frac{\mathbf{x}-\mathbf{x}_j}{h})=\frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}}\exp\{-\frac{1}{2h^2}(\mathbf{x}-\mathbf{x}_j)\Sigma^{-1}(\mathbf{x}-\mathbf{x}_j)^t\}$$

Ядерные методы

Плотность вероятности для класса K_i может быть вычислена по объектам обучающей выборки $\tilde{S}_t = \{s_1 = (y_1, \mathbf{x}_1), \ldots, s_m = (y_m, \mathbf{x}_m)\}$ по формуле $f_i(\mathbf{x}) = \frac{1}{m_i h^n} \sum_{s_j \in K_i}^{m_i} \mathcal{K}(\frac{\mathbf{x} - \mathbf{x}_j}{h})$

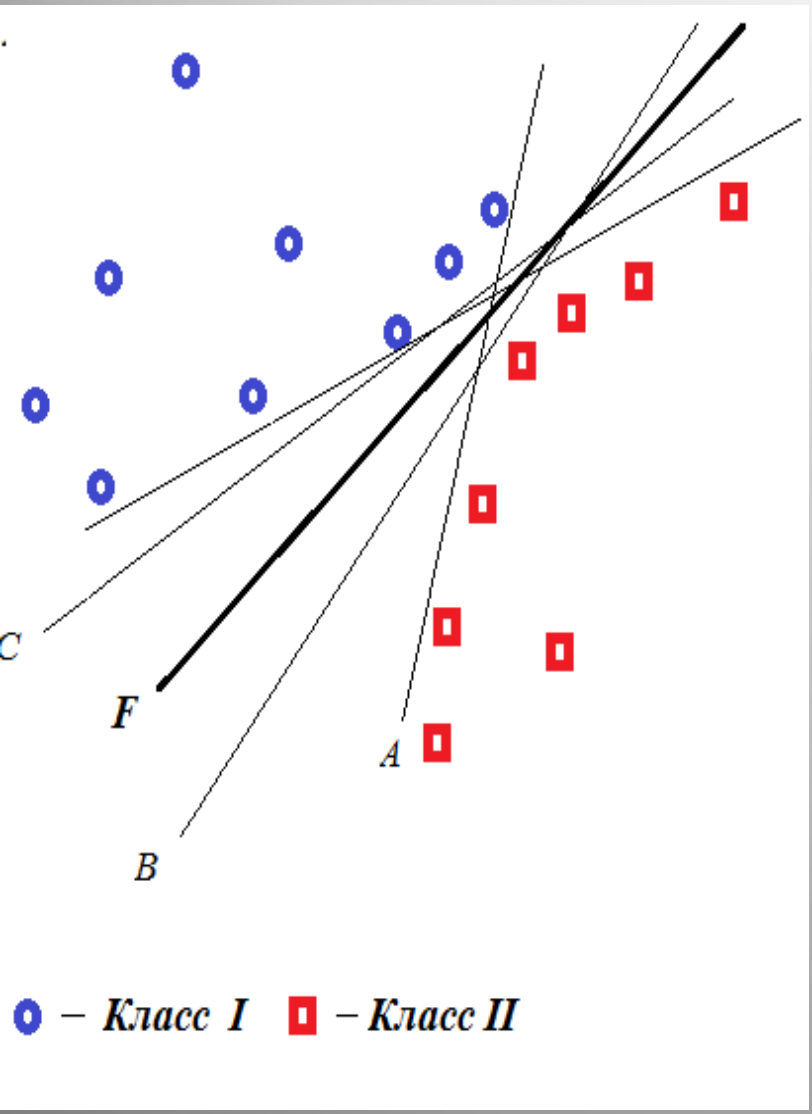
Согласно формуле Байеса распознаваемый объект относится в класс, для которого величина $m_i f_i(\mathbf{x}) h^n = \sum_{s_j \in K_i}^{m_i} \mathcal{K}(\frac{\mathbf{x} - \mathbf{x}_j}{h})$ максимальна .

Параметр сглаживания находится с помощью метода скользящий контроль.

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.

Метод опорных векторов

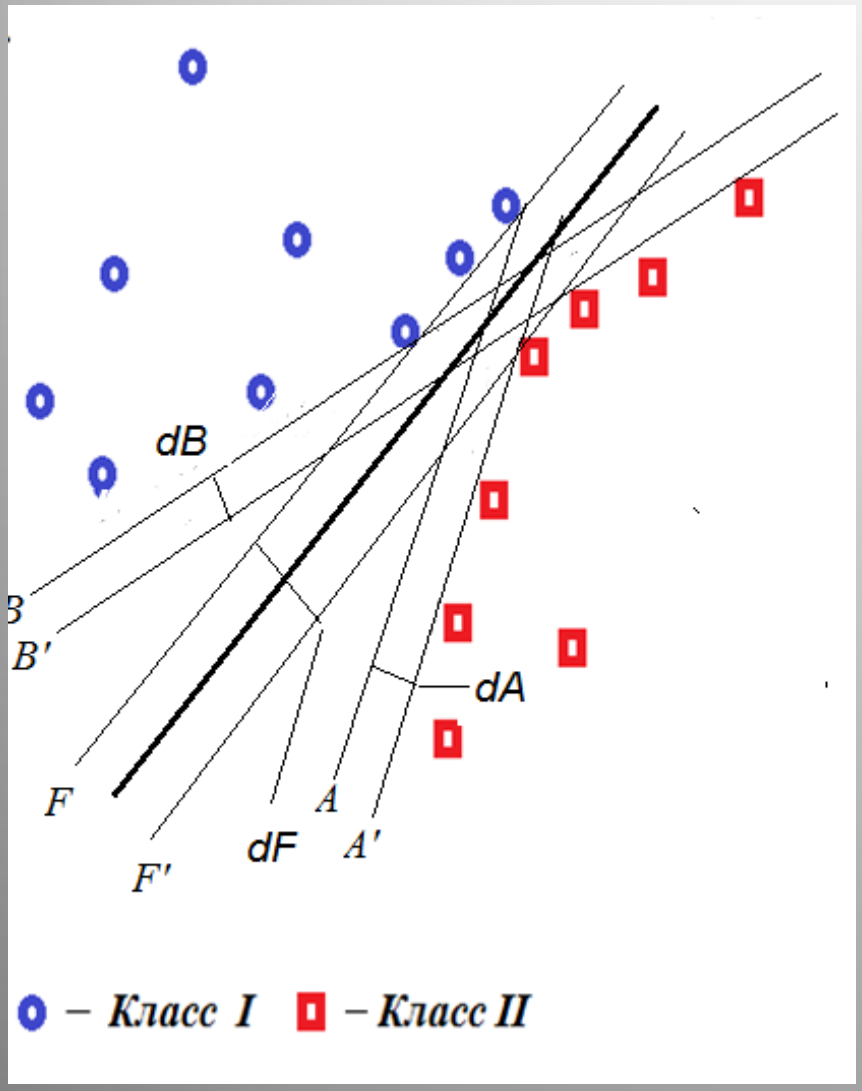


В случаях, когда объекты разных классов в обучающей выборке линейно разделимы, обычно существует целая совокупность линейных поверхностей, осуществляющих такое разделение. На рисунке представлены двумерные данные, где объекты двух классов могут быть раделены с помощью прямых ***A, B, C, D.***

Метод опорных векторов

- Однако наша интуиция, подсказывает что наилучшей обобщающей способностью должна обладать разделяющая прямая *F*, одинаково удалённая от групп объектов из разных классов. Интуитивные представления об оптимальной разделимости формализует проведение разделяющей гиперплоскости посередине между двумя параллельными гиперплоскостями, каждая из которых отделяет объекты одного из классов.

Метод опорных векторов



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'*.

Метод опорных векторов

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

$$\mathbf{w}\mathbf{x}^t = b_1 \qquad (P_1)$$
$$\mathbf{w}\mathbf{x}^t = b_2 \qquad (P_2) \qquad (1)$$

От системы (1) нетрудно перейти к эквивалентной системе

$$\mathbf{z}\mathbf{x}^t = b + 1 \qquad (P_1)$$
$$\mathbf{z}\mathbf{x}^t = b - 1 \qquad (P_2) \qquad (2),$$

описывающей те же самые гиперплоскости.

Метод опорных векторов

Расстояние (величина зазора) δ между гиперплоскостями P_1 и P_2 равна $\frac{2}{|\mathbf{z}|}$. Следовательно задача поиска двух параллельными гиперплоскостями, каждая из которых отделяет объекты одного из классов, может быть сведена к оптимизационной задаче с ограничениями

$$\delta = \frac{2}{|\mathbf{z}|} \to \max$$

$$\mathbf{z}\mathbf{x}_j^t \geq b+1 \quad при \quad s_j \in K_1 \qquad (3)$$

$$\mathbf{z}\mathbf{x}_j^t \leq b-1 \quad при \quad s_j \in K_2$$

Метод опорных векторов

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

Введём обозначение $\alpha_j = 1, \text{если } s_j \in K_1, \text{и } \alpha_j = -1, \text{если } s_j \notin K_1$

Тогда задача (3) оказывается эквивалентна задаче

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

$$\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) \geq 1 \quad j = 1, \ldots, m \qquad (4)$$

Метод опорных векторов

Из известной теоремы Каруша-Куна-Такера 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 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$

соблюдаются условия стационарности лагранжиана,

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

а также условие дополняющей нежёсткости

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

Метод опорных векторов

Из условия стационарности следует, что

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda})}{\partial z_i}\Big|_{z^*} = z_i^* - \sum_{j=1}^{m} \lambda_j \alpha_j x_i = 0 \qquad i = 1, \ldots, n$$

или $$\mathbf{z}^* = \sum_{j=1}^{m} \lambda_j \alpha_j \mathbf{x}_j \quad , \qquad (5)$$

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

Метод опорных векторов

Оптимальные значения множителей $(\lambda_1, \ldots, \lambda_m)$

Могут быть найдены как решение двойственной задачи квадратичного программирования

$$\sum_{i=1}^{m} \lambda_j - \frac{1}{2} \sum_{j'=1}^{m} \sum_{j''=1}^{m} \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} \mathbf{x}_{j'} \mathbf{x}_{j''}^{t} \to \max \qquad (6)$$

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

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

Метод опорных векторов

Пусть $(\hat{\lambda}_1,\ldots,\hat{\lambda}_m)$ - решение задачи (6) .

Направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\hat{\mathbf{z}} = \sum_{j=1}^{m} \hat{\lambda}_j \alpha_j \mathbf{x}_j$

То есть направляющий вектор разделяющей гиперплоскости является линейной комбинацией векторных описаний объектов обучающей выборки, для которых значения соответствующих оптимальных множителей Лагранжа отличны от 0 ..

Метод опорных векторов

Такие векторные описания принято называть опорными векторами. Из условий дополняющей нежёсткости видно, что

$$\hat{\lambda}_j = 0 \;\; \forall j \in J_0$$

$$\text{где } J_0 = \{ j \in \{1,\ldots,m\} \,|\, [\alpha_j (\hat{\mathbf{z}}\mathbf{x}_j^t - b^*) - 1] \neq 0 \}$$

Оценка параметра сдвига $\hat{b}$ находится из ограничения, соответствующего произвольному опорному вектору.

Иными словами

$$\hat{b} = \hat{\mathbf{z}}\mathbf{x}_{j'}^t - \alpha_{j'} \quad \forall j' \in J_s \text{ где } J_s = \{ j \in \{1,\ldots,m\} \,|\, \lambda_j > 0 \}$$

Метод опорных векторов

Таким образом классификация нового распознаваемого объекта $\mathbf{x}$ с описанием s вычисляется согласно знаку выражения

$$g(\mathbf{x}) = \sum_{j=1}^{m} \hat{\lambda}_j \alpha_j \mathbf{x}_j \mathbf{x}^t - \hat{b} \qquad (7)$$

s относится классу K_1 $при\ g(\mathbf{x}) > 0$ и классу K_2 в противном случае.

Метод опорных векторов

Случай отсутствия линейной разделимости

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 $\xi = (\xi_1, \ldots, \xi_m)$. Требования об отделимости классов (3) заменяются требованиями более мягкими требованиями

$$\mathbf{z}\mathbf{x}_j^t \geq b + 1 - \xi_j \quad при \quad s_j \in K_1$$

$$\mathbf{z}\mathbf{x}_j^t \leq b - 1 + \xi_j \quad при \quad s_j \in K_1 \qquad j = 1, \ldots, m$$

Метод опорных векторов

Случай отсутствия линейной разделимости

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

$$\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) \geq 1 - \xi_j,\ \xi_j \geq 0 \quad j = 1,\ldots,m \qquad (8)$$

где C - некоторая положительная константа, являющаяся открытым параметром алгоритма.

Метод опорных векторов

Случай отсутствия линейной разделимости

Из теоремы ККТ следует, что для произвольной точки $(\mathbf{z}^*, b^*, \boldsymbol{\xi}^*)$, 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 (8) , и некоторых векторов неотрицательных множителей Лагранжа $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$ соблюдаются условия стационарности лагранжиана,

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

Метод опорных векторов

Случай отсутствия линейной разделимости

Данные условия записываются в виде

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda}, \boldsymbol{\xi}, \eta)}{\partial z_i}\Big|_{z^*} = z_i^* - \sum_{j=1}^{m} \lambda_j \alpha_j x_i = 0 \qquad i = 1, \ldots, n$$

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda}, \boldsymbol{\xi}, \eta)}{\partial b} = \sum_{j=1}^{m} \lambda_j \alpha_j = 0 \qquad (9)$$

$$\frac{\partial L(\mathbf{z}, b, \boldsymbol{\lambda}, \boldsymbol{\xi}, \eta)}{\partial \xi_j} = C - \lambda_j - \eta_j = 0 \qquad j = 1, \ldots, m$$

Метод опорных векторов

Случай отсутствия линейной разделимости

Также выполняются условия дополняющей нежёсткости

$$\eta_j \xi_j = 0 \qquad j = 1, \ldots, m$$

$$\lambda_j [\alpha_j (\mathbf{z}\mathbf{x}_j^t - b) - 1 + \xi_j] = 0 \qquad j = 1, \ldots, m \qquad (10)$$

Метод опорных векторов

Случай отсутствия линейной разделимости

Оптимальные значения множителей $(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$

Могут быть найдены как решение двойственной задачи квадратичного программирования

$$\sum_{i=1}^{m} \lambda_j - \frac{1}{2} \sum_{j'=1}^{m} \sum_{j''=1}^{m} \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} \mathbf{x}_{j'} \mathbf{x}^t_{j''} \to \max$$

$$\sum_{j=1}^{m} \lambda_j \alpha_j = 0 \qquad (11)$$

$$0 \le \lambda_j \le C \qquad j = 1, \ldots, m$$

Метод опорных векторов

Случай отсутствия линейной разделимости

Как и в случае линейной разделимости направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\hat{\mathbf{z}} = \sum_{j=1}^{m} \hat{\lambda}_j \alpha_j \mathbf{x}_j$ Из условий (10) и (11) следует что $\hat{\eta}_j > 0$ и $\xi_j = 0$ при $0 < \hat{\lambda}_j < C$. Следовательно Оценка параметра сдвига находится из ограничения, соответствующего произвольному опорному вектору $\hat{\lambda}_j < C$

Иными словами

$$\hat{b} = \hat{\mathbf{z}}\mathbf{x}_{j'}^{t} - \alpha_{j'} \quad \forall j' \in J_s \text{ где } J_s = \{j \in \{1,\ldots,m\} \mid C > \lambda_j > 0\}$$

Метод опорных векторов

Случай отсутствия линейной разделимости

Распознавание нового объекта s производится по его описанию $\mathbf{x}$ также как и в случае линейно разделимых классов по с помощью решающего правила (7) по величине распознающей функции $g(\mathbf{x})$.

Метод опорных векторов

Следует отметить, что вектора описаний объектов обучающей выборки $\mathbf{x}_j \quad j = 1, \ldots, m$ входят в задачу (9) только через свои скалярные произведения $\mathbf{x}_{j'}\mathbf{x}_{j''}^t$. Аналогично при вычислении значения распознающей функции (10) по описанию распознаваемого объекта $\mathbf{x}$ на самом деле используются только скалярные произведения $\mathbf{x}\mathbf{x}_j^t \quad при\ \lambda_j > 0$

Предположим что в исходном признаковом пространстве эффективное линейное разделение отсутствует

Метод опорных векторов

Нелинейная разделяющая поверхность

Однако такое разделение может существовать в пространстве $\mathbf{R}_y^{n_y}$ которое может быть получено из исходного признакового пространства $\mathbf{R}_x^n$ с помощью преобразования Φ, ставящего произвольному вектору из $\mathbf{x}' \in \mathbf{R}_x^n$ вектор из $\mathbf{y}' \in \mathbf{R}_y^{n_y}$.

Линейная разделимость означает существование решения.

Задачи квадратичного программирования

$$\sum_{i=1}^{m} \lambda_j - \tfrac{1}{2} \sum_{j'=1}^{m} \sum_{j''=1}^{m} \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} \mathbf{y}_{j'} \mathbf{y}_{j''}^t \to \max$$

$$\sum_{j=1}^{m} \lambda_j \alpha_j = 0 \qquad (12)$$

$$0 \le \lambda_j, \quad \mathbf{y}_j = \Phi(\mathbf{x}_j),\ j = 1, \ldots, m$$

Метод опорных векторов

Нелинейная разделяющая поверхность

Отметим, что необходимость полного восстановления преобразования для поиска всех коэффициентов задачи квадратичного программирования (12) отсутствует. Достаточно восстановить взаимосвязь между скалярными произведениями $\mathbf{x}'(\mathbf{x}'')^t$ и $\Phi(\mathbf{x}')\Phi^t(\mathbf{x}'')$. Одним из способов задания взаимосвязи является выбор такой ядровой функции $\mathcal{K}(\mathbf{x}',\mathbf{x}'')$: $\mathbf{R}_x^n \to \mathbf{R}$, что

(13)

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

.

Метод опорных векторов

Нелинейная разделяющая поверхность

Существование преобразования Φ, для которого выполняется равенство (13) было показано для ядровых функций

$$\mathcal{K}(\mathbf{x}', \mathbf{x}'') = \mathbf{x}'(\mathbf{x}'')^t + \theta, \theta \geq 0$$

$$\mathcal{K}(\mathbf{x}', \mathbf{x}'') = (\mathbf{x}'(\mathbf{x}'')^t + \theta)^d, \theta \geq 0$$

$$\mathcal{K}(\mathbf{x}', \mathbf{x}'') = \exp\left(\frac{|\mathbf{x}' - \mathbf{x}''|^2}{2\sigma^2}\right), \sigma \geq 0$$

Где θ, σ - вещественные неотрицательные параметры, а d - целочисленный параметр.

Метод опорных векторов

Нелинейная разделяющая поверхность

Следовательно поиск коэффициентов задачи квадратичного программирования (12), соответствующих оптимальному преобразованию Φ может быть сведён к подбору таких параметров перечисленных выше ядровых функций, при которых достигается линейная разделимость.

Поскольку в общем случае преобразование Φ является нелинейным, то прообразом в пространстве $\mathbf{R}_x^n$ линейной разделяющей гиперплоскости, существующей в пространстве $\mathbf{R}_y^{n_y}$, может оказаться нелинейная поверхность.

Метод опорных векторов

Нелинейная разделяющая поверхность

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 (8). 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 задания переменных $X_1,\ldots,X_n$ - $\tilde{X}$

Метод опорных векторов

Регрессия

Уменьшение вариации прогнозирующей функции очевидно позволяет снизить вариационную составляющую обобщённой ошибки прогнозирования и уменьшить эффект переобучения.

Задача снижения вариации прогнозируемой функции f формализуется как задача максимизации параметра

$$\delta_\varepsilon = \inf_{(\mathbf{x}',\mathbf{x}'')\in\tilde{X}_\varepsilon^c} (|\mathbf{x}' - \mathbf{x}''|), \text{ где}$$

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

где ε - пороговый параметр.

Метод опорных векторов

Регрессия

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

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

Очевидно, что минимум $|\mathbf{x}' - \mathbf{x}''|$ достигается для пары векторов из $\tilde{X}_{\varepsilon}^{c}$, для которой

а) $|\boldsymbol{\beta}(\mathbf{x}' - \mathbf{x}'')^t| = 2\varepsilon$

Б)вектор $(\mathbf{x}' - \mathbf{x}'')$ совпадает по направлению с вектором $\boldsymbol{\beta}$.

В результате мы получаем $|\boldsymbol{\beta}|\,\delta_\varepsilon = 2\varepsilon \;\; и \;\; \delta_\varepsilon = \dfrac{2\varepsilon}{|\boldsymbol{\beta}|}$

Метод опорных векторов

Регрессия

Наряду с требованиями максимизации параметра δ_ε выдвигается также требование точности аппроксимации на обучающей выборке: отклонение прогнозирующей функции f от значений прогнозируемой величины Y не должно превышать порогового параметра ε. Отметим, что задача максимизации $\frac{2\varepsilon}{|\boldsymbol{\beta}|}$ полностью эквивалентна задаче минимизации $\frac{1}{2\varepsilon}\sum_{i=1}^{n}\beta_i^2$.

Метод опорных векторов

Регрессия

В результате мы переходим к задаче квадратичного программирования

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

$$y_j - \boldsymbol{\beta}\mathbf{x}_j^t - \beta_0 \le \varepsilon \qquad (14)$$

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

Для решения задачи квадратичного программирования (14) используются методы, аналогичные тем, которые используются для решения задачи квадратичного программирования (3), лежащей в основе процедуры обучения алгоритмов распознавания.

Метод опорных векторов

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.