Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

КУРСОВАЯ РАБОТА

Тема: «Машина опорных векторов. Использование SVM в задачах прогнозирования. Сравнительный анализ методов прогнозирования.»

по дисциплине
«Методы оптимизаций»

Выполнили студенты
группы 3630102/80401

А. С. Мамаева
Описание задач SVM и LR, проведение эксперимента, анализ результатов
Д. А. Веденичев
Реализация метода SVR, построение графиков, работа с датасетом
Я. А. Тырыкин
Описание LR, перевод иностранной литературы

Руководитель
Доцент, к.ф.-м.н.

Е.А. Родионова

_____ 2021 г.

Санкт-Петербург
2021

СОДЕРЖАНИЕ

Введение

Метод опорных векторов (*Support vector machine*) – набор алгоритмов, использующихся для задач классификации и регрессионного анализа. Основная идея заключается в переводе исходных векторов в пространство более высокой размерности и поиске разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Помимо выполнения линейной классификации, SVM могут эффективно выполнять нелинейную, используя так называемый трюк с ядром, неявно отображая свои входные данные в пространственные объекты большой размерности.

Первые идеи были предложены ещё в 1950-ые годы. Данный классификатор был создан на основе статистической теории обучения Вапником и Червоненкисом в 1992 году. В настоящее время метод успешно используется во многих областях.

1 Постановка задачи SVM

Рассмотрим задачу бинарной классификации, в которой объектам из конечномерного евклидова пространства $X = \mathbb{R}^n$ соответсвует один из двух классов $Y = \{-1; 1\}$. Пусть задана обучающая выборка $(\vec{x}_i, y_i)_{i=1}^l$. Необходимо построить алгоритм классификации $a(\vec{x}) : \ X \to Y$.

1.1 Разделяющая гиперплоскость

В пространстве $\mathbb{R}^n$ уравнение $(\vec{w},\ \vec{x}) - b = 0$ при заданных $\vec{w}$ и b определяет гиперплоскость – множество векторов $\vec{x} = (x_1, \ldots, x_n)$ принадлежащих пространству меньшей размерности $\mathbb{R}^{n-1}$. Параметр $\vec{w}$ определяет вектор нормали к гиперплоскости, а через $\frac{b}{\vec{w}}$ выражается расстояние от гиперплоскости до начала координат. Гиперплоскость делит $\mathbb{R}^n$ на два полупространства:

$$(\vec{w},\ \vec{x}) - b > 0$$

$$(\vec{w},\ \vec{x}) - b < 0$$

Говорят, что гиперплоскость разделяет два класса C_1 и C_2, если объекты этих классов лежат по разные стороны от гиперплоскости, то есть выполнено либо:

$$\begin{cases} (\vec{w},\ \vec{x}) - b > 0, & \forall x \in C_1 \\ (\vec{w},\ \vec{x}) - b < 0, & \forall x \in C_2 \end{cases} \tag{1}$$

либо

$$\begin{cases} (\vec{w},\ \vec{x}) - b < 0, & \forall x \in C_1 \\ (\vec{w},\ \vec{x}) - b > 0, & \forall x \in C_2 \end{cases} \tag{2}$$

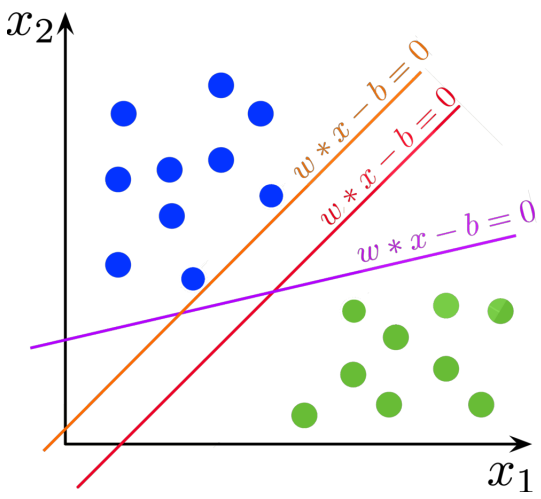


Рис. 1: Примеры разделяющих гиперплоскостей в $\mathbb{R}^2$

1.2 Линейно разделимая выборка

Пусть выборка линейно разделима, то есть существует некоторая гиперплоскость, разделяющая классы -1 и +1. Тогда в качестве алгоритма классификации можно использовать линейный пороговый классификатор:

$$a(\vec{x}) = sign\left((\vec{w},\ \vec{x}) - b\right) = sign\left(\sum_{i=1}^{l} w_i x_i - b\right) \tag{3}$$

где $\vec{x} = (x_1, \ldots, x_n)$ – вектор значений признаков объекта, а $\vec{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ и b – параметры гиперплоскости. Но для двух линейно разделимых классов возможны различные варианты построения разделяющих гиперплоскостей. Метод опорных векторов выбирает ту гиперплоскость, которая максимизирует отступ между классами:

Определение 1. *Отступ (англ. margin) – характеристика, оценивающая, насколько объект «погружён» в свой класс, насколько типичным представителем класса он является. Чем меньше значение отступа M_i, тем ближе объект $\vec{x}_i$ подходит к границе классов и тем выше становится вероятность ошибки. Отступ M_i отрицателен тогда и только тогда, когда алгоритм $a(x)$ допускает ошибку на объекте $\vec{x}_i$. Для линейного классификатора отступ определяется уравнением:* $M_i(\vec{w},\ b) = y_i((\vec{w},\ \vec{x}_i) - b)$

Если выборка линейно разделима, то существует такая гиперплоскость, отступ от которой до каждого объекта положителен:

$$\exists\ \vec{w},\ b : M_i(\vec{w},\ b) = y_i((\vec{w},\ \vec{x}_i) - b) > 0,\ i = 1, \ldots, l \tag{4}$$

Мы хотим построить такую разделяющую гиперплоскость, чтобы объекты выборки находились на наибольшем расстоянии от неё.

Заметим, что при умножении $\vec{w}$ и b на константу $c \neq 0$ уравнение $(c\vec{w},\ \vec{x}) - cb = 0$ определяет ту же самую гиперплоскость, что и $(\vec{w},\ \vec{x}) - b = 0$. Для удобства проведём нормировку: выберем константу c таким образом, чтобы $\min M_i(\vec{w},\ b) = 1$. При этом в каждом из двух классов найдётся хотя бы один «граничный» объект выборки, отступ от которого равен этому минимуму: иначе можно было бы сместить гиперплоскость в сторону класса с большим отступом, тем самым увеличив минимальное расстояние от гиперплоскости до объектов обучающей выборки.

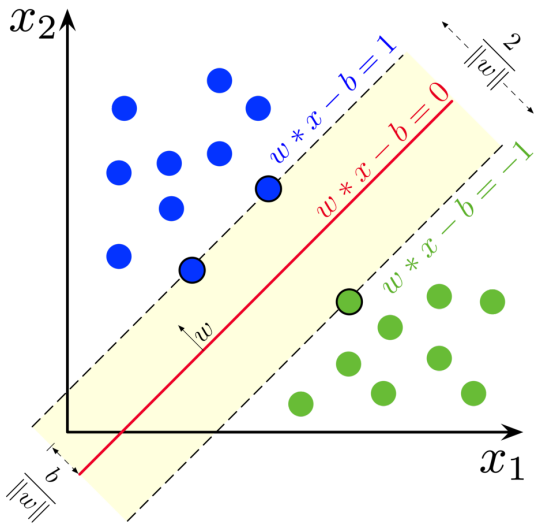


Рис. 2: Оптимальная разделяющая гиперплоскость в $\mathbb{R}^2$

Обозначим любой «граничный» объект из класса $+1$ как $\vec{x}_+$, из класса -1 как $\vec{x}_-$. Их отступ равен единице, то есть

$$\begin{cases} M_+(\vec{w},\ b) = (+1)((\vec{w},\ \vec{x}_+) - b) = 1 \\ M_+(\vec{w},\ b) = (-1)((\vec{w},\ \vec{x}_-) - b) = 1 \end{cases} \tag{5}$$

Нормировка позволяет ограничить разделяющую полосу между классами: $\{x : -1 < (\vec{w},\ \vec{x}_i) - b < 1\}$. Внутри неё не может лежать ни один объект обучающей выборки. Ширину разделяющей полосы можно выразить как проекцию вектора $\vec{x}_+ - \vec{x}_-$ на нормаль к гиперплоскости $\vec{w}$. Чтобы разделяющая гиперплоскость находилась на наибольшем расстоянии от точек выборки, ширина полосы должна быть максимальной:

$$\frac{(x_+ - x_-,\ \vec{w})}{||w||} = \frac{(\vec{x}_+,\ \vec{w}) - (\vec{x}_-,\ \vec{w}) - b + b}{||w||} = \frac{(+1)\big((\vec{x}_+,\ \vec{w}) - b\big) + (-1)\big((\vec{x}_-,\ \vec{w}) - b\big)}{||w||} =$$

$$= \frac{M_+(\vec{w},\ b) + M_-(\vec{w},\ b)}{||w||} = \frac{2}{||w||} \to \max \Rightarrow ||w|| \to \min$$

Это приводит нас к постановке задачи оптимизации в терминах квадратичного программирования:

$$\begin{cases} \frac{||w||}{2} = \frac{1}{2}(w_1^2 + w_2^2 + \cdots + w_l^2) \to \min \\ M_i(\vec{w},\ b) \geq 1,\ i = 1, \ldots, l \end{cases} \tag{6}$$

1.2.1 Применение теоремы Куна-Таккера

Теорема 1. *Пусть поставлена задача выпуклого программирования с ограничениями:*

$$\begin{cases} \varphi_0(x) \to \min\limits_{x \in S} \\ \varphi_i(x) \leq 0,\ i = \overline{1, n} \end{cases} \tag{7}$$

Если x – точка локального минимума при положительных ограничениях, то существуют такие множители $\lambda_i,\ i = \overline{1, m}$, что для функции Лагранжа $L(x; \lambda)$ выполняются условия:

$$\begin{cases} \frac{\partial L}{\partial x} = 0, \quad L(x; \lambda) = \varphi_0(x) + \sum_{i=1}^{n} \lambda_i \varphi_i(x) \\ \varphi_i(x) \leq 0 \quad \text{(исходные ограничения)} \\ \lambda_i \geq 0, \quad \text{(двойственные ограничения)} \\ \lambda_i \varphi_i(x) = 0, \ \text{(условия дополняющей нежёсткости)} \end{cases} \tag{8}$$

При этом искомая точка является седловой точкой функции Лагранжа: минимумом по x и максимумом по двойственным переменным λ.

По теореме Куна—Таккера, поставленная нами задача минимизации эквивалентна двойственной задаче поиска седловой точки функции Лагранжа:

$$L(\vec{x}, \vec{w}, b, \lambda) = \frac{1}{2} \sum_{i=1}^{m} w_i^2 - \sum_{i=1}^{n} \lambda_i (y_i(\vec{w}^T \vec{x} + b) - 1) \tag{9}$$

Продифференцируем функцию Лагранжа и приравняем к нулю производные. Получим следующие ограничения:

$$\begin{cases} \frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \lambda_i y_i = 0 \\ \frac{\partial L}{\partial w_k} = w_k - \sum_{i=1}^{n} \lambda_i y_i x_i^{(k)} = 0,\ k = \overline{1, m} \end{cases} \tag{10}$$

Подставляя условия седловой точки в функцию Лагранжа, получаем двойственную задачу

$$\max\Big(\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \vec{x}_i^T \vec{x}_j\Big) \tag{11}$$

при ограничениях

$$\lambda_i \geq 0,\ i = \overline{1,n},\quad \sum_{i=1}^{n} \lambda_i y_i = 0 \tag{12}$$

Разделяющая функция принимает вид:

$$a(\vec{x}) = \vec{w}^T \vec{x} + b = \sum_{i=1}^{n} \lambda_i \vec{x}_i^T \vec{x} + b \tag{13}$$

1.3 Линейно неразделимая выборка

На практике линейно разделимые выборки практически не встречаются: в данных возможны выбросы и нечёткие границы между классами. В таком случае поставленная выше задача не имеет решений, и необходимо ослабить ограничения, позволив некоторым объектам попадать на «территорию» другого класса. Для каждого объекта отнимем от отступа некоторую положительную величину ε_i, но потребуем чтобы эти введённые поправки были минимальны. Это приведёт к следующей постановке задачи, называемой также SVM с мягким отступом (англ. *soft-margin SVM*): [4], [5]

$$\begin{cases} \frac{1}{2}||\overrightarrow{\omega}||^2 + C\sum\limits_{i=1}^{l} \xi_i \to min \\ M_i(\overrightarrow{\omega}, b) \geq 1 - \xi_i, i = 1, ..., l \\ \xi_i \geq 0, i = 1, ..., l \end{cases} \tag{14}$$

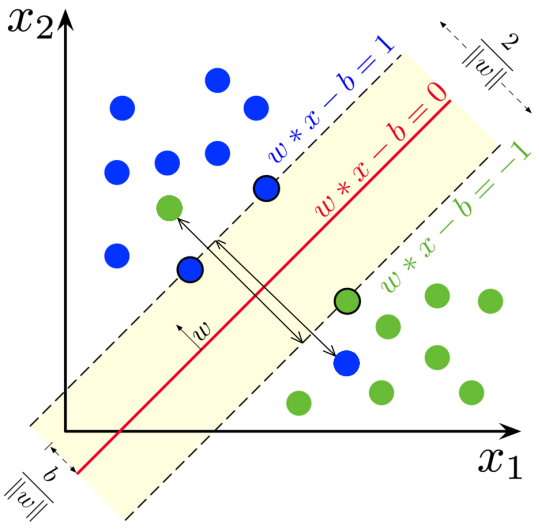


Рис. 3: Оптимальная разделяющая гиперплоскость в $\mathbb{R}^2$ для линейно неразделимых данных

Мы не знаем, какой из функционалов $\frac{1}{2}||\overrightarrow{\omega}||^2$ и $\sum\limits_{i=1}^{l} \xi_i$ важнее, поэтому вводим коэффициент C, который будем оптимизировать с помощью кросс-валидации. В итоге мы получили задачу, у ко-

торой всегда есть единственное решение.

Заметим, что мы можем упростить постановку задачи:

$$\begin{cases} \xi_i \geq 0 \\ \xi_i \geq 1 - M_i(\overrightarrow{\omega}, b) \\ \sum\limits_{i=1}^{l} \xi_i \to min \end{cases} \Rightarrow \begin{cases} \xi_i \geq max(1 - M_i(0, 1 - M_i(\overrightarrow{\omega}, b)) \\ \sum\limits_{i=1}^{l} \xi_i \to min \end{cases} \Rightarrow \xi_i = 1 - M_i(\overrightarrow{\omega}, b))_+$$

Получим эквивалентную задачу безусловной оптимизации:

$$\frac{1}{2}||\overrightarrow{\omega}||^2 + C\sum_{i=1}^{l}(1 - M_i(\overrightarrow{\omega}, b))_+ \to \min_{\omega, b} \tag{15}$$

По аналогии с предыдущим пунктом получаем двойственную задачу для линейно неразделимых данных

$$\max\Big(\sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j \vec{x}_i^T \vec{x}_j\Big) \tag{16}$$

при ограничениях

$$0 \leq \lambda_i \leq C,\ i = \overline{1, n},\ \ \sum_{i=1}^{n}\lambda_i y_i = 0 \tag{17}$$

Разделяющая функция принимает вид:

$$a(\vec{x}) = \vec{w}^T\vec{x} + b = \sum_{i=1}^{n}\lambda_i \vec{x}_i^T\vec{x} + b \tag{18}$$

2 Постановка задачи регрессионного анализа

Рассмотрим модель описания данных, называемую регрессионной и дающую пример статистической зависимости между переменными. Будем предполагать, что наблюдаемое значение отклика y представляет собой сумму двух слагаемых

$$y = f(x_1, \ldots, x_m) + \varepsilon \tag{19}$$

где $y = f(x_1, \ldots, x_m)$ является функцией наблюдаемых значений факторов, а ε – случайная величина, называемая ошибкой регрессионной модели. Случайная ошибка ε может отражать и влияние случайных факторов, и внутренне свойтсвенную отклику случайную изменчивость, и случайные погрешности результатов измерений.

Общая постановка задачи регрессионного анализа такова: по выборке $(x_{1i}, \ldots, x_{mi}, y_i),\ i = 1, 2, \ldots, n$

наблюдаемых (эксперементальных) данных о значениях факторов и отклика

$$y = f(x_{1i}, \ldots, x_{mi}, y_i) + \varepsilon_i,\ i = 1, \ldots, n \tag{20}$$

требуется построить оценку функциональной зависимости y от x: $\tilde{y} = \tilde{f}(x)$, где $x = (x_1, \ldots, x_m)$ вектор факторов. Сформулируем теперь дополнительные предположения о характере случайных ошибок ε_i в виде регрессионной зависимости $f(x)$. В регрессионной модели обычно принимают, что $\varepsilon_1, \ldots, \varepsilon_n$ – независимые, одинаково распределённые случайные величины с нулевыми математическим ожиданием $M\varepsilon_i = 0$ и дисперсией $D\varepsilon_i = \sigma^2,\ i = 1, \ldots, n$.

Относительно регрессионной функции f, как правило, предполагают, что её вид задан с точностью до неизвестных параметров $\beta_1, \ldots, \beta_i$, или, что то же, – относительно векторного параметра $\beta = (\beta_1, \ldots, \beta_i)$

$$y_i = f(x_i, \beta) + \varepsilon_i,\ i = 1, \ldots, n \tag{21}$$

В этом случае построение оценки регрессии $\tilde{f}$ сводится к построению оценок неизвестных параметров этой функции: $\tilde{y} = f(x, \tilde{\beta})$

Важный на практике класс регрессионных моделей возникает при рассмотрении линейных относительно неизвестных параметров β регрессионных зависимостей $f(x, \beta)$ – это задачи линейного регрессионного анализа. При этом зависимость отклика от факторов x может быть и нелинейной. Соответсвенно задача оценивания $f(x, \beta)$ при нелинейной зависимости от β носит название нелинейного регрессионного анализа.

Опредление 2. *Регрессией называется функциональная связь в среднем любых случайных величин.*

Опредление 3. *Регрессионным анализом называется раздел математической статистики, изучающий зависимость между случайными величинами с помощью уравнений регрессии.*

2.1 Простая линейная регрессия

2.1.1 Модель простой линейной регрессии

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1..n \tag{22}$$

где $x_1, ..., x_n$– заданные числа (значения фактора); $y_1, ...y_n$– наблюдаемые значения отклика; $\varepsilon_1, ..., \varepsilon_n$– независимые, нормально распределенные $N(0, \sigma)$ с нулевым математическим ожиданием и одина-

ковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1- неизвестные параметры, подлежащие оцениванию.

В модели (22) отклик y зависит зависит от одного фактора x, и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика y. Погрешности результатов измерений x в этой модели полагают существенно меньшими погрешностей результатов измерений y, так что ими можно пренебречь [1, с. 507].

2.1.2 Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распрстранённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0, \beta_1} \tag{23}$$

Задача минимизации квадратичного критерия $Q(\beta_0, \beta_1)$ носит название задачи метода наименьших квадратов (МНК), а оценки $\hat{\beta}_0, \hat{\beta}_1$ параметров β_0, β_1, реализующие минимум критерия $Q(\beta_0, \beta_1)$, называют МНК-оценками [1, с. 508].

2.1.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров $\hat{\beta}_0, \hat{\beta}_1$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум. Для нахождения МНК-оценок $\hat{\beta}_0, \hat{\beta}_1$ выпишем необходимые условия экстремума

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases} \tag{24}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из (24) системы получим:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases} \tag{25}$$

Разделим оба уравнения на n:

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 (\frac{1}{n} \sum x_i) = \frac{1}{n} \sum y_i \\ \hat{\beta}_0 (\frac{1}{n} \sum x_i) + \hat{\beta}_1 (\frac{1}{n} \sum x_i^2) = \frac{1}{n} \sum x_i y_i \end{cases} \tag{26}$$

и, используя известные статистические обозначения для выборочных первых и вторых начальных моментов

$$\bar{x} = \frac{1}{n}\sum x_i, \bar{y} = \frac{1}{n}\sum y_i, \bar{x^2} = \frac{1}{n}\sum x_i^2, \bar{xy} = \frac{1}{n}\sum x_i y_i, \tag{27}$$

получим

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \bar{x^2} = \bar{xy}, \end{cases} \tag{28}$$

откуда МНК-оценку $\hat{\beta}_1$ наклона прямой регрессии находим по формуле Крамера

$$\hat{\beta}_1 = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - (\bar{x})^2} \tag{29}$$

а МНК-оценку $\hat{\beta}_0$ определяем непосредственно из первого уравнения (28) :

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \tag{30}$$

Заметим, что определитель системы (24):

$$\bar{x^2} - (\bar{x})^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = s_x^2 > 0, \tag{31}$$

если среди значений $x_1, ..., x_n$ есть различные, что и будем предполагать.
Доказательство минимальности функции $Q(\beta_0, \beta_1)$ в стационарной точке проведём с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2 = 2n\bar{x^2}, \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2\sum x_i = 2n\bar{x} \tag{32}$$

$$\triangle = \frac{\partial^2 Q}{\partial \beta_0^2} \cdot \frac{\partial^2 Q}{\partial \beta_1^2} - (\frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0})^2 = 4n^2\bar{x^2} - 4n^2(\bar{x})^2 = 4n^2\left[\bar{x^2} - (\bar{x})^2\right] = 4n^2\left[\frac{1}{n}\sum (x_i - \bar{x})\right] = 4n^2 s_x^2 > 0. \tag{33}$$

Этот результат вместе с условием $\frac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0$ означает, что в стационарной точке функция Q имеет минимум [1, с. 508-511].

3 Применение SVM в задачах регрессии

Регрессия опорных векторов (SVR) представляет собой наиболее общую форму SVM. Исходные данные берутся в виде $\{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\} \subset \mathbb{R} \times \mathbb{R}^n$. В $\varepsilon-$ опорной векторной регрессии цель состоит в нахождении функции $f(x)$, которая отклоняется на величину, не превосходящую ε, от фактически полученных значений y_i для всех обучающих данных, в то же время достаточно

гладких. Случай с линейной функцией можно описать следующим образом:

$$f(x) = (\omega, x) + b, \omega \in \mathbb{R}^n, b \in \mathbb{R} \tag{34}$$

где $(\cdot, \cdot)$ обозначает скалярное покомпонентное произведение в $\mathbb{R}^n$
Гладкость в (34) означает малость ω . Следовательно, требуется минимизировать значение евклидовой нормы $||\omega||^2$. Формально это можно записать в виде задачи выпуклой оптимизации:

$$min||\omega||^2 \tag{35}$$

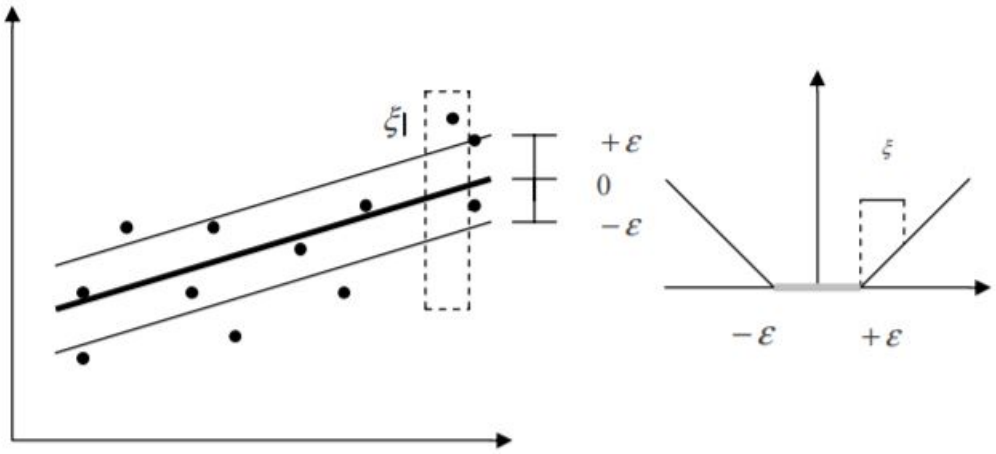
$$\begin{cases} y_i - (\omega, x_i) - b \leq \varepsilon \\ (\omega, x_i) + b - y_i \leq \varepsilon \end{cases}$$

Приведенная выше задача выпуклой оптимизации считается корректной в случае существования такой функции f и апроксимации ею всех пар (ω, x) с требуемой точностью ε. Если ввести вспомогательные переменные ξ_i, ξ_i^*, чтобы справиться с недопустимыми решениями задачи (35), перейдем к следующе формулировке все той же задачи выпуклой оптимизации:

$$min[||\omega||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)] \tag{36}$$

$$\begin{cases} y_i - (\omega, x_i) - b \leq \varepsilon + \xi_i \\ (\omega, x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Положительная константа C определяет согласованность между гладкостью функции и величиной, до которой допускаются отклонения, превышающие ε.



Функция интенсивных потерь $|\xi|_\varepsilon$ описывается следующим образом:

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{при } |\xi| < \varepsilon \\ |\xi| - \varepsilon & \text{в остальных случаях} \end{cases} \tag{37}$$

На вышеприведенном рисунке изображено графическое прдеставление приведенной функции. Двойственная формулировка дает возможность применять SVM к нелинейным функциям. Стандартный метод дуализации с использованием множителей Лагранжа:

$$L = \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)] - \sum_{i=1}^{l}[\alpha_i(\varepsilon + \xi_i - y_i + (\omega, x_i) + b)] - \tag{38}$$
$$- \sum_{i=1}^{l}[\alpha_i^*(\varepsilon + \xi_i^* - y_i + (\omega, x_i) + b)] - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

Двойственные переменные в выражении (38) должны быть положительными, $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$. Это следует из условия седловой точки, частные производные функции L относительно переменных $\omega, b, \xi_i, \xi_i^*$ должны удовлетворять условиям оптимальности:

$$\frac{\delta L}{\delta b} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \tag{39}$$

$$\frac{\delta L}{\delta \omega} = \omega - \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)x_i = 0 \tag{40}$$

$$\frac{\delta L}{\delta \xi_i^*} = C - \alpha_i^* - \eta_i^* \tag{41}$$

Подстановка выражений (39), (40), (41) в (38) дает двойственную задачу:

$$max\{-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i, x_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l}y_i(\alpha_i - \alpha_i^*)\} \tag{42}$$

$$\begin{cases} \sum\limits_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

Двойственные переменные η_i, η_i^* через условие (41) были устранены для получения задачи (42). Уравнение (40) можно переписать в следующем виде:

$$\omega = \sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)x_i \Rightarrow f(x) = \sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(x_i, x) + b \tag{43}$$

Это так называемое расширение опорных векторов, то есть ω можно задать линейной комбинацией тренировочтных. Для оценки функции $f(x)$ нет необходимости явно находить значение ω. Вычисление b выполняется с использованием условия Каруша - Куна - Таккера (ККТ), которое утверждает, что в оптимальном решении произведения между двойственными переменными и ограничениями должны исчезнуть. В случае опорных векторов это означает:

$$\begin{cases} \alpha_i(\varepsilon + \xi_i - y_i + (\omega, x_i) + b) = 0 \\ \alpha_i^*(\varepsilon + \xi_i^* + y_i - (\omega, x_i) + b) = 0 \end{cases} \tag{44}$$

И:

$$\begin{cases} (\mathrm{C} - \alpha_i)\xi_i = 0 \\ (\mathrm{C} - \alpha_i^*)\xi_i^* = 0 \end{cases} \tag{45}$$

Можно сделать следующие выводы: вне ε-цилиндра вокруг функции f лежат только пары (x_i, y_i) с соответствующими $\alpha_i^* = C$. Также $\alpha_i\alpha_i^* = 0$,то есть, не может быть одновременно двух ненулевых переменных α_i, α_i^*, поскольку для этого потребуется ненулевое ослабление в обоих направлениях. Наконец, для $\alpha_i^* \in (0, C), \xi_i^* = 0$ и, кроме того , второе равенство в выражении (44) должен исчезнуть. Поэтому b можно посчитать следующим образом:

$$\begin{cases} b = y_i - (\omega, x_i) - \varepsilon & \text{для } \alpha_i \in (0, C) \\ b = y_i - (\omega, x_i) + \varepsilon & \text{для } \alpha_i^* \in (0, C) \end{cases} \tag{46}$$

Из условий (44) следует, что только при $|f(x_i) - y_i| \geq \varepsilon$ множители Лагранжа принимают ненулевые значения, или иными словами, для любых последовательностей внутри цилиндра радиуса ε $\alpha_i, _i^*$ исчезнут: при $|f(x_i) - y_i| < \varepsilon$ второе уравнение в выражении (44) будет ненулевым, следовательно α_i, α_i^* должно быть равно нулю, чтобы выполнялись условия Каруша-Куна-Таккера. Поэтому, допустимое разложение ω существует в терминах x_i (то есть, для описания нужны не все x_i). Опорными векторами называются те ω, которые получаются при ненулевых коэффициентах α_i, α_i^* [2], [3].

4 Показатели эффективности построенных моделей

Качество работы модели линейной регрессии (LR) и регрессии опорных векторов (SVR) оценивается по следующим параметрам: RMSE, R^2, MAE и коэффициентом корреляции.

1. Средняя абсолютная ошибка (Mean Absolute Error – MAE) даёт общую вариацию прогнозируемых и оцененных значений параметра.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - f(x_i)| \tag{47}$$

где i – индекс выборки, y_i – прогнозируемые значения, $f(x_i)$ – фактические значения, n – номера выборки набора данных.

2. Среднеквараличная ошибка (Root Mean Square Error – RMSE) даёт квадратный корень из среднего отклонения результатов прогноза и ожидаемых результатов.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{n}} \tag{48}$$

где i – индекс выборки, y_i – прогнозируемые значения, $f(x_i)$ – фактические значения, n – номера выборки набора данных.

3. Значение коэффициента корреляции определяется как значения, представляющие независимый параметр x и зависимый параметр y. Он обозначается r и выражается как:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{49}$$

4. Коэффициент детерминации R^2 – отображает изменение зависимого параметра и независимого параметра модели в процентах. В частном случае линейной зависимости R^2 является квадратом так называемого множественного коэффициента корреляции между зависимой переменной и объясняющими переменными. В частности, для модели парной линейной регрессии коэффициент детерминации равен квадрату обычного коэффициента корреляции между y и x.

5 Вычислительный эксперимент

5.1 Актуальность

В конце 2019 года в китайском городе Ухань произошла вспышка новой коронавирусной инфекции, получившей название COVID-19. Всемирная организация здравоохранения объявила эту вспышку чрезвычайной ситуацией в области общественного здравоохранения, имеющей международное значение, а 11 марта — пандемией. Россию, и в частности Санкт-Петербург, болезнь не обошла стороной.

5.2 Цель работы

Необходимо собрать актуальную информацию по количеству заразившихся COVID-19 за текущий 2021 год по городу Санкт-Петербург. Визуализировать данные на графике. Применить модель линейной регрессии (LR) и машину опорной регрессии (SVR) для построения линии наилучшего

соответсвия тенденции заболеваемости. Сравнить данные методы между собой с помощью оценок, приведённых в параграфе (4).

5.3 Обработка данных

Для нашей работы воспользуемся данными, взятыми с официального сайта открытых датасетов России. Построим график, на оси абсцис которого будем откладывать дни, а на оси ординат количества заразившихся.

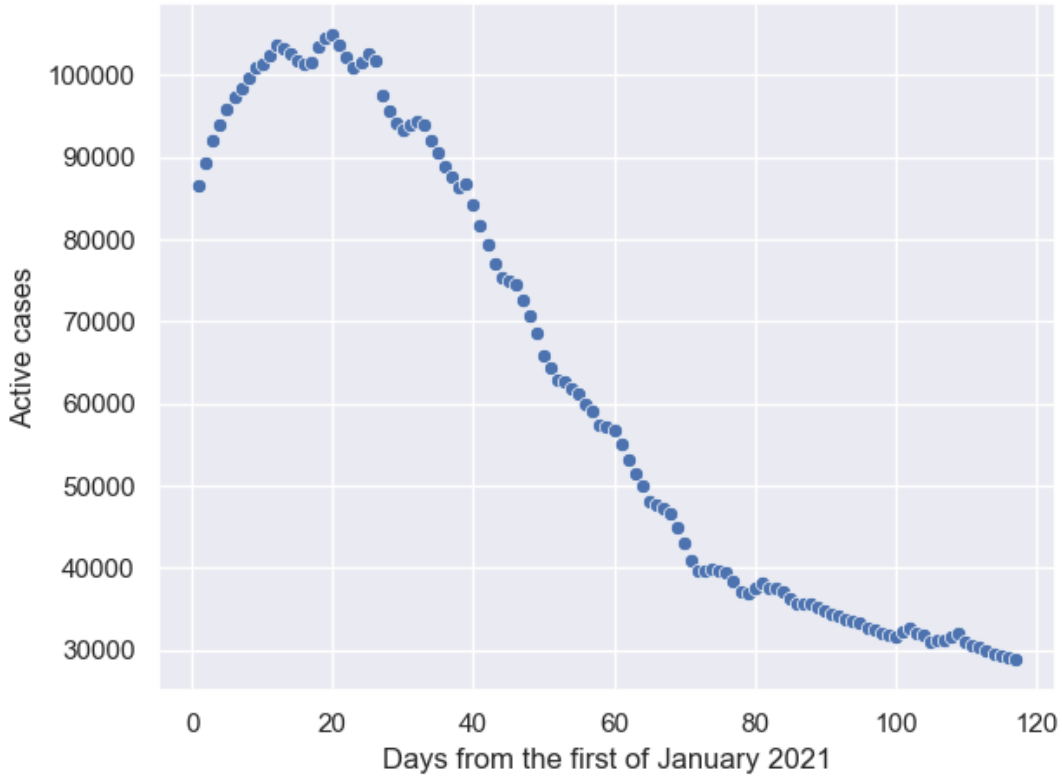


Рис. 4: Статистика числа заболевших COVID-19

5.4 Алгоритм нахождения оптимальных констант SVR

Для нахождения оптимальных значений C и ε метода SVR выполнялась следующая последовательность действий:

1. Используя поиск по сетке значений ε и C с большим разбросом исходного и итогового значений находилась область, где результаты получались наиболее подходящими для исходной задачи.

2. В месте на области, где получались минимальные отклонения результатов, снова выполнялся поиск по сетке. Таким образом удавалось локализовать оптимальную область.

3. Из оптимальной области выбиралось C, для которого, с различными значениями ε (границы значений получены из пункта 2), выполнялось обучение и определение оптимального значения ε.

4. С полученным ε выполнялось обучение с различными значениями C. Таким образом находя оптимальный параметр C.

Сказанное выше проиллюстрируем графиком, полученным при выполнении программного кода:

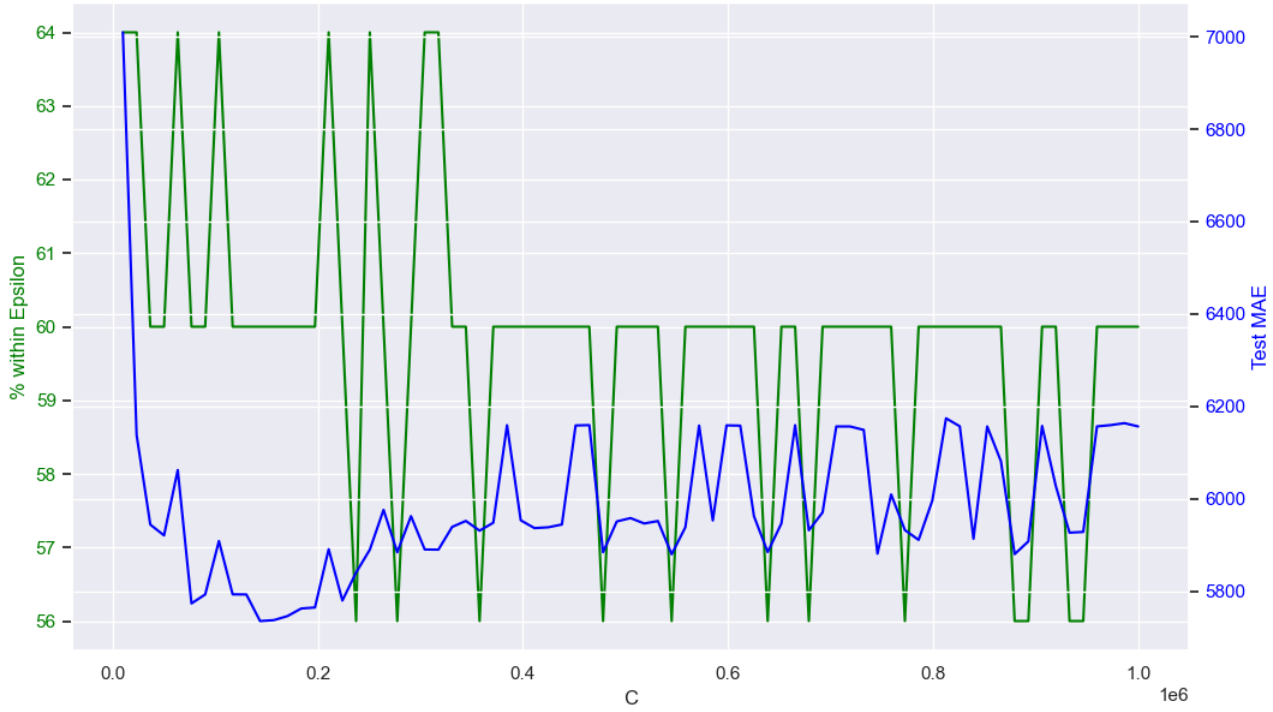


Рис. 5: Выбор оптимальных параметров

Таким образом, для нашей задачи было решено положить константу C равной 200000, константу ε равной 6900.

5.5 Поиск тенденции заболеваемости

Применяя алгоритмы, описанные в нашей работе для поиска регрессии, получаем следующие результаты.

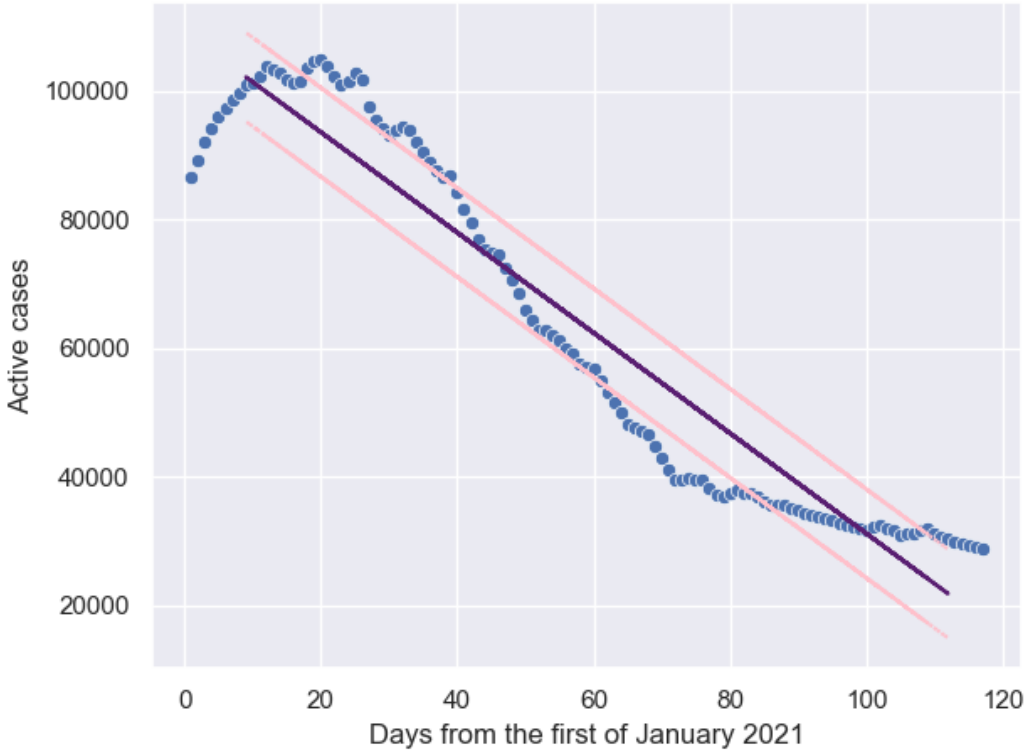


Рис. 6: Регрессия, полученная при помощи SVR

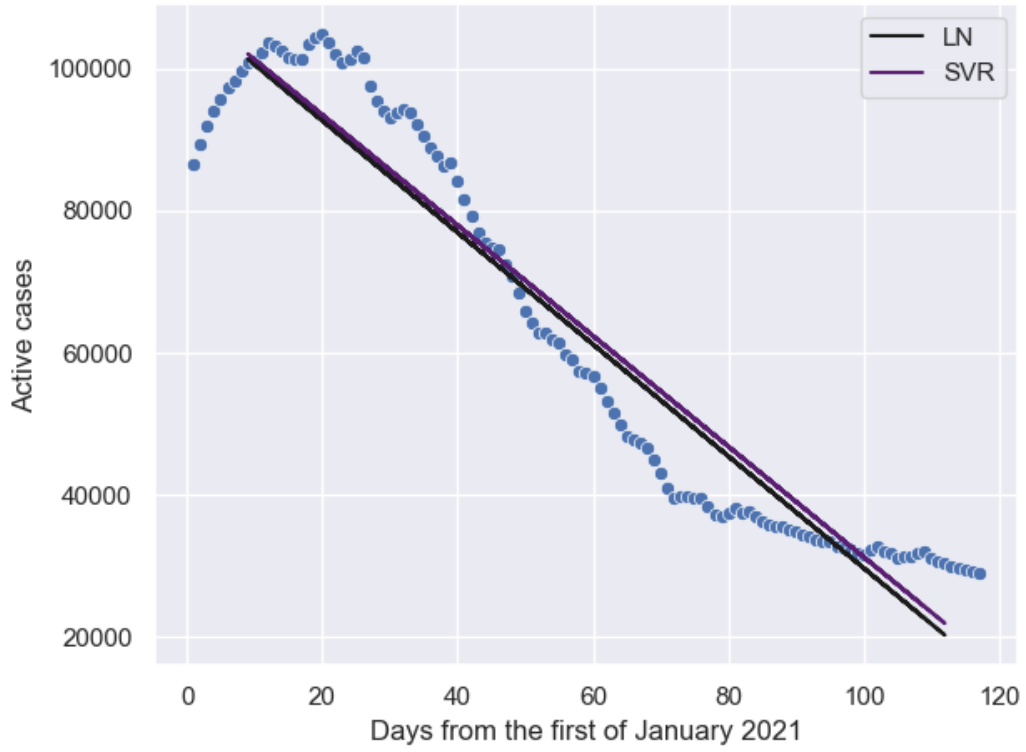


Рис. 7: Совмещенный график линий регрессии SVR и LR

6 Анализ результатов

Составим сводную таблицу, в которую запишем значения показателей эффективности.

	LR	SVR
MAE	6013.153	5762.637
RMSE	48261115.85	43848996.00
r	0.974	0.974
R^2	0.947	0.941

Таблица 1: Сравнительный анализ показателей эффективности методов

Анализируя параметры, видим, что при оптимально выбранных константах метода SVR, получаем результаты точнее относительно модели линейной регрессии. Это отчетливо прослеживается при сравнении среднеквадратичных ошибок. Таким образом можем заключить, что машина опорных векторов лучше справилась с задачей построения тенденции заболеваемости.

7 Программная реализация

В процессе реализации алгоритмов использовался язык программирования Python3.6 и среда разработки PyCharm. Для визуализации результатов и построения графиков, мы пользовались встроенными библиотеками:

1. matplotlib
2. seaborn

Для того чтобы работать с датасетом, пользовались библиотекой pandas, для работы с моделями – sklearn.
Исходный код находится в системе контроля версий GitHub
https://github.com/Brightest-Sunshine/regression-analysis-of-patients-with-COVID-19

Список литературы

[1] Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров»,2001. — 592 с., илл

[2] Platt, John. Fast Training of Support Vector Machines using Sequential Minimal Optimization, in Advances in Kernel Methods – Support Vector Learning, B. Scholkopf, C. Burges, A. Smola, eds., MIT Press (1998)

[3] Support Vector Regression/ Debasish Basak, Srimanta Pal and Dipak Chandra Patranabis/ Vol. 11, No. 10, October 2007

[4] Vapnik, V., Estimation of Dependences Based on Empirical Data, Springer-Verlag, (1982).

[5] Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, (1995).