УДК 004.85

АНАЛИЗ ЯДЕРНЫХ МЕТОДОВ В КОНТЕКСТЕ SVM

Березина С. В., Романов Д. А., студенты

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Лущакова И. Н. – канд. физ.-мат. наук, доцент

Аннотация. В данной работе рассматриваются теоретические и практические аспекты ядерных методов в машинном обучении с фокусом на метод опорных векторов (SVM). Анализируется математический аппарат гильбертовых пространств и "ядерного трюка", позволяющего эффективно решать нелинейные задачи. Подробно разбираются геометрические свойства разделяющих гиперплоскостей, формулируются прямая и двойственная задачи оптимизации SVM, а также исследуется роль опорных векторов в построении классификатора.

Ключевые слова. Машинное обучение, метод опорных векторов, ядерные методы, нелинейная классификация, линейная разделимость, радиальная базисная функция, полиномиальное ядро, сигмоидальное ядро, обучение модели.

Ядерные методы решают нелинейные задачи классификации данных, неявно перенося данные в пространство большей размерности, где классы становятся линейно разделимыми, с помощью функции ядра $K(x, y)$, которая вычисляет скалярное произведение образов объектов без явного построения отображения. Метод опорных векторов (SVM) использует эту идею для поиска разделяющей гиперплоскости с максимальным зазором между классами: линейный SVM применяется, когда данные линейно разделимы; в противном случае нелинейный SVM за счёт ядерного трюка строит более гибкие разделяющие поверхности.

1. Математические основы ядерных функций и метода опорных векторов

1.1 Линейный классификатор: постановка задачи

Фундаментальной задачей машинного обучения является задача классификации (разделения) данных на два класса. Рассмотрим формальную постановку: имеется обучающая выборка $(X, t) = \{x_n, t_n\}_{n=1}^{N}$, где каждый объект представлен признаковым описанием $x_n \in R^d$, и для него известна метка класса $t_n \in \{-1, +1\}$. Цель заключается в построении модели, способной предсказывать метку класса t для нового объекта x на основе выявленных закономерностей.

Одним из ключевых представителей линейных классификаторов является метод опорных векторов (Support Vector Machines, SVM). Линейные классификаторы принимают решение о принадлежности объекта классу на основе линейной комбинации признаков. Формально, это решение определяется знаком линейного решающего правила:

$$f(\boldsymbol{x}) = \sum_{j=1}^{d} \omega_j x(j) + b = \boldsymbol{\omega^T x} + b, \qquad t(\boldsymbol{x}) = \begin{cases} +1, \text{если } f(\boldsymbol{x}) \geq 0, \\ -1, \text{если } f(\boldsymbol{x}) < 0. \end{cases}$$

Здесь $\omega_j \in R$ представляют собой веса признаков, $b \in R$ — параметр сдвига (смещение). С геометрической точки зрения, линейный классификатор определяет в признаковом пространстве R^d разделяющую плоскость, задаваемую уравнением $f(\boldsymbol{x}) = 0$. Объекты, для которых значение $f(\boldsymbol{x})$ положительно, относятся к первому классу (+1), а объекты с отрицательным значением — ко второму (-1). Введем вспомогательную функцию $h(\boldsymbol{x}, t)$, формализующую правило принятия решения.

$$t(\boldsymbol{x}) = arg \max_{t \in \{-1,+1\}} h(\boldsymbol{x}, t) = arg \max_{t \in \{-1,+1\}} tf(\boldsymbol{x})$$

Объект *x* относится к такому классу *t*, для которого значение соответствующей линейной функции максимально.

1.2 Геометрия разделяющей гиперплоскости

Пусть заданы вещественные числа $\boldsymbol{\alpha_1, \ldots, \alpha_n}$, b, причем не все $\{\boldsymbol{\alpha_j}\}_{\boldsymbol{j=1}}^{\boldsymbol{n}}$ равны нулю. Уравнение

$$\boldsymbol{\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n = b}$$

относительно переменных $\boldsymbol{x_1, \ldots, x_n}$ задает в пространстве $\boldsymbol{R_n}$ множество, называемое гиперплоскостью. Скалярное произведение векторов $\boldsymbol{X} = (\boldsymbol{x_1, x_2, \ldots, x_n})$ и $\boldsymbol{W} = (\boldsymbol{\omega_1, \omega_2, \ldots, \omega_n})$ задается формулой $\langle \boldsymbol{X, W} \rangle = \boldsymbol{x_1 \omega_1 + x_2 \omega_2 + \ldots + x_n \omega_n}$. Тогда уравнение гиперплоскости принимает следующий вид:

$$\langle \boldsymbol{X, W} \rangle + b = 0$$

Вектор $\boldsymbol{W} = (w_1, w_2, \ldots, w_n)$ интерпретируется как нормальный вектор к гиперплоскости.

Гиперплоскость разделяет R_n на два открытых полупространства: множество точек, для которых $\langle \boldsymbol{X}, \boldsymbol{W} \rangle + b > 0$, и множество точек, для которых $\langle \boldsymbol{X}, \boldsymbol{W} \rangle + b < 0$. Если к открытому полупространству присоединить все точки самой гиперплоскости, то получится замкнутое полупространство.

Если множество $S_1 \subset R^n$ целиком содержится в замкнутом полупространстве, а множество S_2 полностью лежит в противоположном открытом полупространстве, то гиперплоскость отделяет S_1 от S_2 и называется разделяющей гиперплоскостью для этих множеств. Для классов S_1, S_2 разделяющая гиперплоскость удовлетворяет условию $S_1 \subset \{\langle \boldsymbol{X}, \boldsymbol{W} \rangle + b \geq 0\}, S_2 \subset \{\langle \boldsymbol{X}, \boldsymbol{W} \rangle + b < 0\}$ (или наоборот).

1.3 Оптимальная разделяющая гиперплоскость

Среди бесконечного множества гиперплоскостей, разделяющих два класса, SVM выбирает ту, которая максимизирует зазор. Оптимальная гиперплоскость находится максимально далеко от точек обоих классов. Рассмотрим опорные векторы — объекты, которые лежат ближе всего к разделяющей гиперплоскости. Обозначим через $\boldsymbol{X}^+$ опорный вектор (т.е. точку) из класса +1, а через $\boldsymbol{X}^-$ — из класса −1. Оптимальная разделяющая гиперплоскость проходит через середину отрезка, соединяющего точки $\boldsymbol{X}^+$ и $\boldsymbol{X}^-$, и перпендикулярна этому отрезку. В векторной форме это условие записывается как:

$$\langle \boldsymbol{X} - \boldsymbol{X_C}, \boldsymbol{X}^+ - \boldsymbol{X}^- \rangle = 0, \qquad \text{где} \quad \boldsymbol{X_C} = \frac{\boldsymbol{X}^+ + \boldsymbol{X}^-}{\boldsymbol{2}}$$

Рассмотрим случай, когда объекты двух классов могут быть разделены линейно без ошибок. В этом случае мы можем провести гиперплоскость таким образом, чтобы она не просто отделяла классы друг от друга, а делала это с максимально возможным «зазором».

Введем понятие зазора (margin). Для фиксированной разделяющей гиперплоскости зазором между ней и классом называется наименьшее расстояние от гиперплоскости до объектов данного класса. Обозначим зазоры для положительного и отрицательного классов как d_+ и d_- соответственно. Тогда оптимальной разделяющей гиперплоскостью будет та, которая максимизирует минимальный из этих двух зазоров:

$$\min(d_+, d_-) \rightarrow \max_{\boldsymbol{\omega}, b}.$$

Чтобы выразить этот критерий в терминах параметров модели $\boldsymbol{\omega}$ и b, необходимо преодолеть их неоднозначность. Дело в том, что гиперплоскость, задаваемая уравнением $\boldsymbol{\omega^T x} + b = 0$, не меняется при умножении $\boldsymbol{\omega}$ и b на произвольную ненулевую константу. Устраним это, зафиксировав шкалу специальным образом.

Рассмотрим ближайшие к разделяющей гиперплоскости объекты обучения. Условимся, что линии (поверхности) уровня, проходящие через эти объекты, будут описываться уравнениями $f(\boldsymbol{x}) = 1$ для объектов класса (+1) и $f(\boldsymbol{x}) = -1$ для объектов класса (-1). Такая нормировка возможна, поскольку оптимальная гиперплоскость, всегда проходит ровно посередине между этими двумя линиями (поверхности) уровня.

Теперь вычислим величину зазора. Возьмем произвольный вектор $\boldsymbol{u_1}$, лежащий на самой гиперплоскости $\boldsymbol{\omega^T u_1} + b = 0$, и вектор $\boldsymbol{u_2}$ на линии (поверхности) уровня для первого класса $\boldsymbol{\omega^T u_2} + b = 1$. Вычитая первое равенство из второго, получаем $\boldsymbol{\omega^T}(\boldsymbol{u_2} - \boldsymbol{u_1}) = 1$. Зазор d_+ — это не что иное, как длина проекции вектора $(\boldsymbol{u_2} - \boldsymbol{u_1})$ на направление нормального вектора $\boldsymbol{\omega}$. Вычислив эту проекцию, находим:

$$pr_{\boldsymbol{\omega}}(\boldsymbol{u_2} - \boldsymbol{u_1}) = \frac{\boldsymbol{\omega^T}(\boldsymbol{u_2} - \boldsymbol{u_1})}{\|\boldsymbol{\omega}\|} = \frac{1}{\|\boldsymbol{\omega}\|}$$

Таким образом, искомый зазор обратно пропорционален норме вектора весов: $d_+ = d_- = 1/\|\boldsymbol{\omega}\|$. Следовательно, максимизация зазора эквивалентна минимизации величины $\|\boldsymbol{\omega}\|$.

Условия корректной классификации всех объектов с учетом введенной нормировки принимают вид системы неравенств:

$$\begin{cases} \boldsymbol{\omega^T x_n} + b \geq 1, & \text{если } t_n = +1 \\ \boldsymbol{\omega^T x_n} + b < 1, & \text{если } t_n = -1 \end{cases}$$

Эти два условия можно объединить в одно:

$$t_n(\boldsymbol{\omega^T x_n} + b) \geq 1, \quad \forall n.$$

Объединив критерий максимизации зазора с ограничениями на корректную классификацию и заменив максимизацию $1/\|\omega\|$ на более удобную для оптимизации минимизацию $\frac{1}{2}\|\omega\|^2$, приходим к стандартной постановке задачи обучения SVM для линейно разделимого случая:

$$\begin{cases} \frac{1}{2}\|\boldsymbol{\omega}\|^2 \to \min_{\boldsymbol{\omega},b} \\ t_n(\boldsymbol{\omega}^T\boldsymbol{x_n} + b) \geq 1, \quad n = 1,\dots,N \end{cases}$$

1.4 Прямая и двойственная задачи SVM.

Метод опорных векторов базируется на двух эквивалентных формулировках оптимизационной задачи. Прямая задача заключается в непосредственном поиске параметров разделяющей гиперплоскости — вектора весов $\boldsymbol{\omega}$ и сдвига b. Её цель — минимизировать норму $\boldsymbol{\omega}$ (что эквивалентно максимизации зазора) и суммарную ошибку на обучающих данных (если введены ослабляющие переменные ξ_n):

$$\frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{n=1}^{N}\xi_n \to \min_{\boldsymbol{\omega},b,\xi}, \qquad t_n(\boldsymbol{\omega}^T\boldsymbol{x_n} + b) \geq 1 - \xi_n, \quad \xi_n \geq 0 \ \ \forall n.$$

Однако для дальнейшего развития метода гораздо удобнее оказывается двойственная задача, которая получается из прямой с помощью метода множителей Лагранжа и теоремы Куна–Таккера. Вместо $\boldsymbol{\omega}$ и b здесь переменными становятся множители λ_n (по одному на каждый объект обучения), а целевая функция и ограничения принимают вид:

$$-\frac{1}{2}\sum_{n,m=1}^{N}\lambda_n\lambda_m t_n t_m\langle x_n, x_m\rangle + \sum_{n=1}^{N}\lambda_n \to \max_{\lambda}, \qquad 0 \leq \lambda_n \leq C, \qquad \sum_{n=1}^{N}\lambda_n t_n = 0.$$

Решающее правило для нового объекта x представляется следующим образом:

$$f(x) = \sum_{n=1}^{N}\lambda_n t_n(x_n, x) + b^*$$

где b^* находится из условий дополняющей нежёсткости для любого опорного вектора.

Двойственная формулировка удобна тем, что объекты входят в неё только через скалярные произведения, что позволяет заменить их на ядерную функцию и неявно перейти в пространство большей размерности. Кроме того, решение получается разреженным: большинство множителей λ_n равны нулю, а ненулевые соответствуют опорным векторам, которые и определяют положение разделяющей границы.

1.5 Нелинейный SVM, применение "ядерного трюка"

До сих пор мы рассматривали метод опорных векторов как способ построения оптимальной разделяющей гиперплоскости в исходном признаковом пространстве R^d. Однако реальные задачи классификации редко бывают линейно разделимыми. Граница между классами часто имеет сложную нелинейную форму. Основная идея преодоления нелинейности: если данные не разделяются линейно в текущем пространстве, нужно преобразовать их в другое пространство, где они станут линейно разделимыми.

В более общем виде, мы рассматриваем некоторое отображение $\Phi : R^d \to H$, которое переводит исходный объект *x* в новое пространство признаков *H*, обычно гораздо более высокой (иногда бесконечной) размерности. Задача SVM теперь заключается в построении оптимальной разделяющей гиперплоскости не для исходных объектов x_n, а для их образов $\Phi(x_n)$ в этом новом пространстве.

В данном случае не получится решить стандартную задачу SVM в простанстве *H.* Обратимся к *двойственной задаче оптимизации SVM*. Если мы переходим в пространство признаков *H*, все скалярные произведения заменяются на $\langle\Phi(x_n), \Phi(x_m)\rangle$. Ключевая идея ядерного метода заключается в том, чтобы заменить это скалярное произведение в пространстве признаков специальной функцией ядра, которая работает непосредственно с исходными объектами:

$$K(x_n, x_m) = \langle\Phi(x_n), \Phi(x_m)\rangle_H$$

Это и есть "ядерный трюк". Нам не нужно знать, как выглядит преобразование *Ф* или какова размерность пространства *H*. Достаточно определить функцию *K*, которая, будучи примененной к двум объектам исходного пространства, возвращает результат, эквивалентный их скалярному произведению в новом, более сложном пространстве. Все вычисления остаются в исходном пространстве, но мы при этом решаем нелинейную задачу.

Двойственная задача и решающее правило приобретают вид:

$$-\frac{1}{2}\sum_{n,m=1}^{N}\lambda_n\lambda_m\, t_n t_m K(x_n,x_m)+\sum_{n=1}^{N}\lambda_n \to \max_{\lambda}, \qquad 0\le\lambda_n\le C, \qquad \sum_{n=1}^{N}\lambda_n t_n = 0,$$

$$f(x)=\sum_{n=1}^{N}\lambda_n t_n K(x_n,x)+b^*.$$

Очевидно, что не любая функция двух аргументов может служить ядром. Чтобы функция $K(x,y)$ действительно соответствовала скалярному произведению в некотором пространстве *H*, она должна удовлетворять условию Мерсера. Не вдаваясь в тонкости, это условие требует, чтобы для любой интегрируемой со своим квадратом функции g выполнялось неравенство:

$$\iint_{X\,x\,X} K(x,y)g(x)g(y)dxdy \ge 0,$$

где X x X – область определения признаковых описаний. На практике это означает, что матрица попарных сходств (матрица Грама) для любого набора данных должна быть положительно полуопределенной.

Существует несколько классических ядер, удовлетворяющих этому условию и широко используемых на практике:

1. Линейное ядро: $K(x,y)=x^Ty$. Фактически возвращает нас к линейному SVM.
2. Полиномиальное ядро: $K(x,y)=(x^Ty+\theta)^d$, где $\theta\ge 0$, а $d\in N$ — степень полинома. Позволяет строить полиномиальные разделяющие поверхности.
3. Радиально-базисное ядро (RBF) или гауссово ядро: $K(x,y)=exp(-\frac{\|x-y\|^2}{2\sigma^2})^d$, где $\sigma>0$ — ширина окна. Это самое популярное ядро, способное аппроксимировать любую гладкую границу. Оно обладает свойством локальности: близкие объекты влияют друг на друга сильнее, чем далекие.
4. Сигмоидное ядро: $K(x,y)=tanh(\alpha x^Ty+\beta)$. При определенных параметрах оно ведет себя как нейронная сеть.

2. Визуализация метода на основе простейших примеров.

2.1. Визуализация метода опорных векторов.

Пусть нам дана выборка линейно разделимых признаков, которые можно однозначно разместить на плоскости OXY. Выборка представлена на рисунке 1.

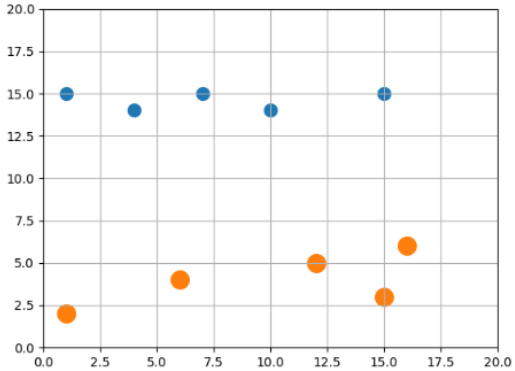

Рисунок 1 - Линейно разделимая выборка на плоскости.

Для решения задачи классификации осуществляется построение прямой, разделяющей пространство признаков. Существует множество возможных вариантов такого разделения, часть из которых представлена на рисунке 2. С использованием метода опорных векторов определяется оптимальная прямая, изображённая на рисунке 3. На данном рисунке центральная линия,

обозначенная как «separating hyperplane», представляет собой оптимальную разделяющую гиперплоскость, в рассматриваемом случае — прямую, максимизирующую ширину зазора.

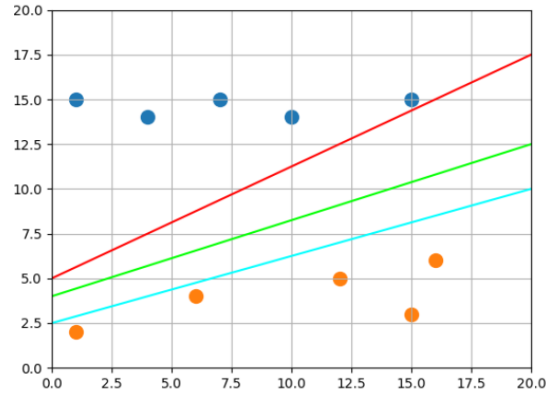

Рисунок 2 - Различные прямые, которые можно построить.

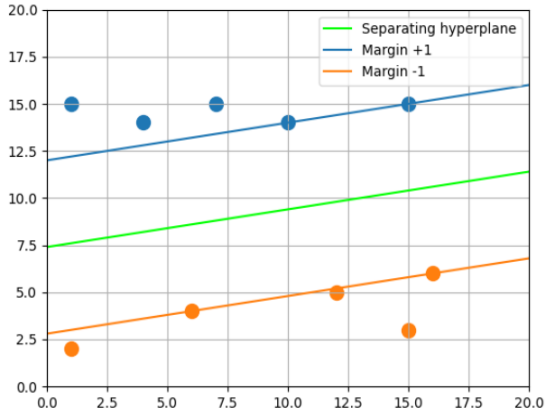

Рисунок 3 - Оптимальная прямая.

После определения оптимальной разделяющей гиперплоскости (в данном случае — прямой), можно перейти к проверке её практической эффективности. Для этого выполним предсказание принадлежности новых точек к одному из двух классов. На представленных рисунках объекты класса 1 отображены синими маркерами, тогда как объекты класса -1 обозначены оранжевыми маркерами. Такой шаг позволяет убедиться, что построенная модель корректно разделяет пространство признаков и способна обобщать информацию, выходящую за пределы обучающей выборки.

Пусть дана выборка из 3 точек с координатами (5, 19), (18, 1), (10, 7) (см. рисунок 4). Выполним предсказание для этих точек. Если модель предсказала что точка принадлежит классу 1, перекрасим метку в синий цвет, для класса -1 - в оранжевый. Результаты представлены на рисунке 5.

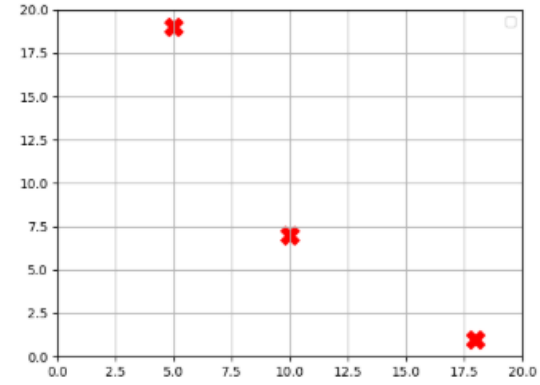

Рисунок 4 - Отображение на плоскости выборки для тестирования модели.

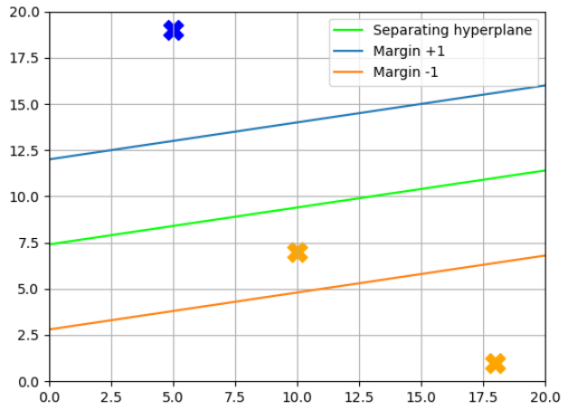

Рисунок 5 - Отображение предсказаний модели.

Проведённая проверка подтверждает, что обученная модель SVM обеспечивает устойчивое разделение классов и корректно классифицирует новые объекты, не участвовавшие в обучении.

2.2 Визуализация метода опорных векторов с применением ядерных методов.

Пусть дана выборка, представляющая собой классическую задачу "два кольца" — две концентрические окружности с добавлением небольшого шума. Такой набор данных широко используется для демонстрации ограничений линейных методов классификации: точки внутреннего круга относятся к первому классу, точки внешнего кольца — ко второму. В исходном двумерном пространстве невозможно провести прямую линию, которая разделила бы эти классы (см. рисунок 6). Возникает проблема линейной неразделимости, которую классический линейный SVM не способен преодолеть.

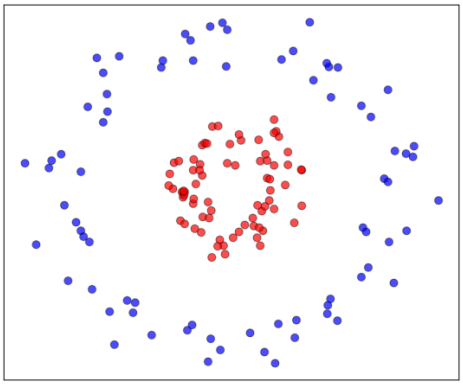

Рисунок 6 - Исходные данные (два кольца).

Для демонстрации эффективности ядерного подхода были обучены две модели SVM: с линейным ядром и с радиальным базисным ядром (RBF). Параметры моделей выбраны стандартными, обеспечивающими устойчивую работу алгоритма.

Как видно из представленной визуализации (см. рисунок 7), линейное ядро не справляется с поставленной задачей — разделяющая граница представляет собой прямую линию, которая не может корректно отделить один класс от другого. Точность классификации в данном случае оказывается на уровне случайного угадывания (~50%).

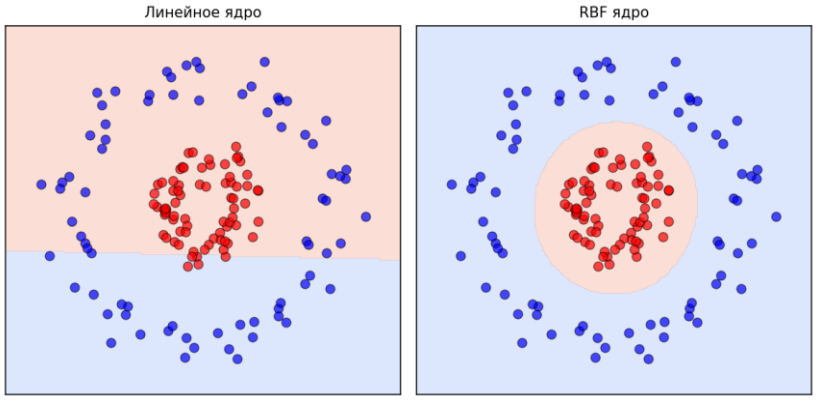

Рисунок 7 - Сравнение линейного и RBF ядер.

В отличие от линейного подхода, RBF ядро демонстрирует высокую эффективность. Разделяющая граница принимает сложную нелинейную форму, повторяющую геометрию исходных данных. Это наглядно иллюстрирует основное преимущество ядерных методов — возможность построения нелинейных разделяющих поверхностей без явного перехода в пространство признаков большей размерности.

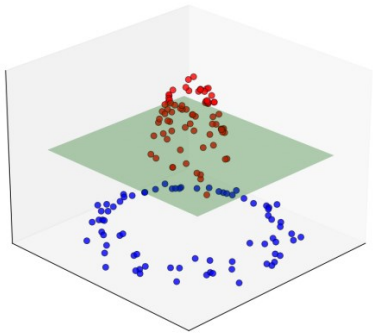

Рисунок 8 - Преобразование данных в трехмерное пространство.

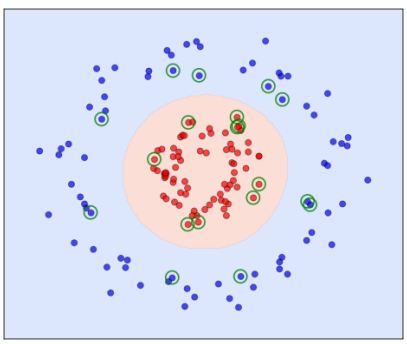

Рисунок 9 - Опорные векторы в модели с RBF ядром.

Для понимания механизма работы ядерных методов рассматривается геометрическая интерпретация RBF-ядра (см. рисунок 8). Идея заключается в неявном отображении исходных признаков в пространство большей размерности, где новые признаки зависят от расстояний между точками. Для каждой точки формируются характеристики, значения которых убывают с увеличением расстояния до других точек. Чем ближе точки расположены друг к другу, тем больше соответствующее значение признака; чем дальше — тем оно меньше. Благодаря этому классы становятся разделимыми простой плоскостью, проведенной на некоторой промежуточной высоте (зеленая плоскость на графике). То, что было невозможно в двумерном пространстве с помощью прямой, достигается в трехмерном пространстве с помощью плоскости.

Важной особенностью SVM является то, что решение определяется только опорными векторами — точками, которые лежат на границе разделяющей полосы или внутри нее. В случае нелинейной границы опорные векторы располагаются вдоль кривой, разделяющей классы. На представленном графике (рисунок 9) опорные векторы обведены окружностями. Можно заметить, что они располагаются преимущественно вблизи границы раздела классов, формируя "скелет" разделяющей поверхности. Именно эти точки вносят основной вклад в определение формы границы решения.

Проведенные эксперименты демонстрируют принципиальное преимущество ядерных методов перед линейным подходом в случае нелинейно разделимых данных. Геометрическая интерпретация ядерного преобразования показывает, что добавление даже одного признака, нелинейно зависящего от исходных, может перевести задачу в разряд линейно разделимых в пространстве большей размерности. Роль опорных векторов в формировании нелинейной границы решения оказывается определяющей — именно они задают форму разделяющей поверхности через ядерную функцию.

3. Применение метода опорных векторов на практике.

В рамках практического применения метода опорных векторов была рассмотрена задача бинарной классификации уровня дохода на основе социально‑демографических характеристик. Исходная выборка включала параметры, описывающие возраст, сферу занятости, уровень образования и другие социально‑экономические признаки. Для упрощения постановки задачи непрерывная переменная дохода была преобразована в бинарный признак: «доход выше 50 тысяч» и «доход ниже или равен 50 тысячам». Такое преобразование позволяет корректно использовать метод опорных векторов, ориентированный на решение задач бинарной классификации, и оценить его эффективность в условиях реальных данных.

Для начала была проведена предварительная обработка исходных данных, включающая анализ структуры выборки и устранение наиболее очевидных источников шума. Фрагмент необработанных данных представлен в таблице 1. С целью повышения линейной разделимости признакового пространства и соответственно улучшения качества классификации методом опорных векторов в набор данных была добавлена новая категориальная переменная *age_type*, отражающая возрастные группы (young, adult, old). Дополнительно было выявлено, что в столбцах *workclass*, *occupation* и *native_country* часто встречается символ «?», что, вероятно, связано с нежеланием респондентов раскрывать соответствующую информацию. Вместо удаления таких записей, что могло бы привести к потере значимых наблюдений, отсутствующие значения были заменены на *None*, что позволяет корректно обработать их в процессе обучения модели. Также для целевой переменной *income* была сформирована бинарная версия *income_num*, принимающая значение 1 при доходе выше или равном 50 тысячам и 0 в противном случае, что делает задачу полностью совместимой с бинарной природой метода опорных векторов. Фрагмент очищенной выборки приведён в таблице 2.

После выполнения предварительной очистки данных был проведён анализ ключевых признаков с целью выявления потенциальных недостатков выборки. Анализ распределения целевой переменной выявил существенный дисбаланс классов: доля наблюдений с доходом ниже 50 тысяч составила 76%, тогда как на долю наблюдений с доходом выше данного порога пришлось лишь 24% от общего объема выборки. Подобная диспропорция способна привести к переобучению модели и снижению её обобщающей способности, поэтому на последующих этапах была выполнена нормализация и балансировка данных. Дополнительный анализ признаков выявил выраженные выбросы в столбцах *capital_gain* и *capital_loss*, что видно из таблицы 1. Наличие подобных аномальных значений может негативно повлиять на устойчивость и точность модели, поэтому наблюдения с экстремальными значениями по данным признакам были исключены из выборки.

Таблица 1 - Необработанные данные.

	age	workclass	education	education_num	marital_status	occupation	relationship	sex	capital_gain	capital_loss	hours_per_week	native_country	income	age_types	income_num
20227	40	?	Masters	14	Divorced	?	Own-child	Female	0	0	55	United-States	<=50K	adult	0
1888	22	Self-emp-not-inc	HS-grad	9	Never-married	Other-service	Own-child	Male	0	0	40	Greece	<=50K	young	0
10564	38	Federal-gov	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	Male	0	2051	40	United-States	<=50K	adult	0
9117	25	Private	Some-college	10	Never-married	Craft-repair	Own-child	Female	0	0	50	United-States	<=50K	adult	0
21754	49	Self-emp-inc	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Male	15024	0	50	United-States	>50K	adult	1
4922	30	Local-gov	Some-college	10	Never-married	Adm-clerical	Not-in-Family	Male	0	0	40	United-States	<=50K	adult	0
9339	46	Private	Some-college	10	Divorced	Sales	Unmarried	Female	0	0	40	Cuba	<=50K	adult	0
4680	39	Private	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Male	15024	0	50	United-States	>50K	adult	1
17258	23	Private	HS-grad	9	Never-married	Other-service	Own-child	Male	0	0	12	United-States	<=50K	young	0
16294	34	?	HS-grad	9	Married-civ-spouse	?	Husband	Male	2885	0	80	United-States	<=50K	adult	0

Таблица 2 - Обработанные данные.

	age	workclass	education	education_num	marital_status	occupation	relationship	sex	capital_gain	capital_loss	hours_per_week	native_country	income	age_types	income_num
21086	43	Private	Some-college	10	Married-civ-spouse	Handlers-cleaners	Husband	Male	0	0	40	United-States	>50K	adult	1
22657	51	Private	HS-grad	9	Married-civ-spouse	Tech-support	Husband	Male	0	0	40	United-States	<=50K	adult	0
6037	26	Self-emp-not-inc	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Male	0	0	75	United-States	<=50K	adult	0
25145	52	Private	7th-8th	4	Married-civ-spouse	Other-service	Husband	Male	0	0	48	Cuba	<=50K	adult	0
15638	39	Private	Prof-school	15	Divorced	Exec-managerial	Not-in-family	Male	0	0	50	United-States	<=50K	adult	0
20894	19	Private	HS-grad	9	Never-married	Other-service	Own-child	Male	0	0	15	United-States	<=50K	young	0
15313	57	Private	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	Male	0	0	40	United-States	<=50K	adult	0
25813	56	NaN	HS-grad	9	Divorced	NaN	Not-in-family	Male	0	0	40	United-States	<=50K	adult	0
4719	39	Private	Bachelors	13	Divorced	Exec-managerial	Not-in-family	Female	0	0	25	United-States	<=50K	adult	0
6492	50	Self-emp-inc	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	Male	0	0	75	United-States	>50K	adult	1

После выполнения предварительной очистки данных был сформирован набор признаков, пригодный для последующего обучения модели. Далее была проведена категоризация всех строковых признаков посредством процедуры кодирования, что позволило преобразовать их в числовое представление, совместимое с алгоритмами машинного обучения. На заключительном этапе была выполнена нормализация данных. Результат всех этапов подготовки данных представлен в таблице 3.

Таблица 3. Результат подготовки данных.

	age	workclass	education	education_num	marital_status	occupation	relationship	sex	capital_gain	capital_loss	hours_per_week	native_country
0	0.301370	0.750	0.600000	0.800000	0.666667	0.000000	0.2	1.0	0.052626	0.0	0.397959	0.926829
1	0.452055	0.625	0.600000	0.800000	0.333333	0.214286	0.0	1.0	0.000000	0.0	0.122449	0.926829
2	0.287671	0.375	0.733333	0.533333	0.000000	0.357143	0.2	1.0	0.000000	0.0	0.397959	0.926829
3	0.493151	0.375	0.066667	0.400000	0.333333	0.357143	0.0	1.0	0.000000	0.0	0.397959	0.926829
4	0.150685	0.375	0.600000	0.800000	0.333333	0.642857	1.0	0.0	0.000000	0.0	0.397959	0.097561

После завершения этапов очистки и нормализации данных выборка была разделена на обучающую и тестовую подвыборки. Такой подход позволяет объективно оценить способность модели обобщать информацию и корректно работать с данными, которые модель еще не видела.

Проведённые эксперименты показали, что метод опорных векторов демонстрирует устойчивые результаты при решении задачи бинарной классификации уровня дохода. Обучение модели с линейным ядром дало среднюю точность 0.834 на обучающей выборке (по результатам перекрёстной проверки) и 0.827 на тестовой выборке, что свидетельствует о хорошем балансе между способностью модели выявлять закономерности и её обобщающей способностью. Применение радиально‑базисного ядра (RBF) не привело к улучшению качества классификации: показатели точности полностью совпали с результатами линейной модели. Это может указывать на то, что после предварительной обработки, нормализации и кодирования признаков структура данных стала близкой к линейно разделимой, и использование более сложного ядра не дало дополнительного выигрыша. В совокупности результаты позволяют заключить, что для данной задачи линейная модель SVM является не только достаточной, но и оптимальной с точки зрения соотношения сложности и качества классификации.

Список использованных источников:

1. Метод опорных векторов [Электронный ресурс]. – URL: https://vmath.ru/vf5/users/au/svm. – Дата доступа: 03.03.2026.

2. Метод опорных векторов в стандартной задаче классификации [Электронный ресурс]. – URL: http://www.machinelearning.ru/wiki/images/2/25/smais11_svm.pdf. – Дата доступа: 03.03.2026.

3. Понимание пространства ядер в машинном обучении [Электронный ресурс] / Хабр. – URL: https://habr.com/ru/articles/814343/. – Дата доступа: 03.03.2026.

4. Income classification dataset [Электронный ресурс] – URL: https://www.kaggle.com/datasets/lodetomasi1995/income-classification - Дата доступа: 06.03.2026.

UDC 004.85

ANALYSIS OF KERNEL METHODS IN THE CONTEXT OF SVM

Biarezina S. V., Romanov D. A., students

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Lushchakova I. N. – PhD in Physics and Mathematics

Annotation. This article examines the theoretical and practical aspects of kernel methods in machine learning, focusing on the support vector machine (SVM). It analyzes the mathematical formalism of Hilbert spaces and a "kernel trick" that allows for the efficient solution of nonlinear problems. The geometric properties of separating hyperplanes are examined in detail, the primal and dual SVM optimization problems are formulated, and the role of support vectors in classifier construction is explored.

Keywords. Machine learning, support vector machine, SVM, kernel methods, kernel trick, classification, feature space, nonlinear classification, linear separability, radial basis function, polynomial kernel, sigmoid kernel, model training, test set.