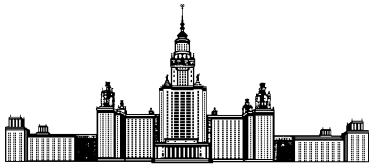
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

Байесовский подход к обучению распознаванию образов с учетом критерия гладкости решающего правила на основе метода опорных векторов

Выполнил:

студент 3 курса 317 группы

Квасов Андрей Федорович

Научный руководитель:

к.ф-м.н.

Красоткина Ольга Вячеславовна

Москва, 2015

Содержание

Аннотация

Обращаясь к конкретным прикладным задачам распознавания образов, часто объектами, данными, являются упорядоченные последовательности величин, которые характеризуют явления, свойства объектов и их изменение по какой-либо оси аргумента, в частности, по временной. Такого роды задачи, задачи распознавания сигналов отражены в более прикладных областях, к примеру, в распознавании рукописных символов. Из-за своей специфики возникают различные проблемы связанные с признаковым описанием объектов распознавания.

В данной работе рассматривается метод опорных векторов, с использованием мягких отступов и ядер, применяемый в конкретной задаче распознавания рукописных символов, а также рассматривается вероятностная интерпретация наложенных на модель ограничений, связанных со спецификой задач распознавания символов.

1 Введение

1.1 Основные понятия и определения

Задача распознавания сигналов, как часть задачи распознавания образов, оперирует измерениями каких-либо реальных объектов или явлений, расположенных упорядоченно по аргументу $t \in T = 1, 1, \ldots, N, \ldots$, и представляющие из себя наборы векторов $x_t = x_t^1, \ldots, x_t^n$, каждая координата которой может принадлежать множеству $\mathbb{X}^n$, не совпадающему для разных координат.

Помимо различия в значениях данных наблюдается одна из важных проблем распознавания сигналов: к примеру, рассмотрев написание букв от руки автором (*online handwriting*), и с какой-либо частотой регистрировать такие признаки как координата (x, y) в ограничивающем символ окне, угол под которым находиться электронное перо или ручка, сила надавливания на экран регистрирующего устройства, скорость движения точек и т.п., для каждой буквы будет создано не фиксированное по длине признаковое описание. Таким образом, будет затруднительно сформировать пространство, удовлетворяющее гипотезе компактности, как основной гипотезе построения моделей и алгоритмов машинного обучения. Также нельзя использовать в качестве признаков отсчеты сигнала, взятые с некоторым шагом вдоль оси аргумента, поскольку сигналы, полученные от разных написаний даже одного и того же символа неизбежно будут иметь разную длину, и, следовательно. Значит не будет существовать единого линейного пространства, в котором могли бы быть представлены написания распознаваемых символов.

Фиксированность размеров признакового пространства в задаче машинного обучения является необходимым условием для не метрических алгоритмов таких как *метод опорных векторов (Support Vector Machine)*, рассматриваемый в данной работе.

Вторая проблема, которая и была рассмотрена в данной работе, обозначает ситуацию, в которой исходная обучающая выборка мала и может лишь сравниться с размером признакового пространства. Для задачи распознавания рукописных символов свойственно, что количество объектов в обучающей совокупности почти равно среднему размеру признакового пространства, или много меньше (около 100 отсче-

тов на символ: букву, цифру и т.п.). Поэтому для получения большей точности на контрольной выборке, имея малоинформативную обучающую выборку, необходимо наложить априорные ограничения на вид признакового пространства. Одним из таких ограничений является *критерий гладкости* для компонент в описании сигнала.

Как будет показано ниже, критерий гладкости представляет собой вид перехода пространства признаков из исходного пространства, в котором классификация имеет недостаточно хорошие результаты, в так называемое *спрямляющее пространство*, определяющее новое, линейно разделимое множество точек выборки с новым признаковым описанием.

Данная концепция используется повсеместно, в особенности, в задачах распознаваня сигналов.

1.2 Обзор литературы

Для решения указанной ранее проблемы различного количества измерений объектов, когда признаковое описание реальных объектов, даже принадлежащих одному и тому же классу, является переменной величиной. В работе К. Бальмана [1] представлен *Gaussian Dynamic Time Warping (GDTW)* ядровой переход. При котором, степень близости между объектами, необходимая в RBF-ядре, меняется с евклидового расстояния на метрику $D(\tau_i, \tau_j)$ другого вида. В свою очередь, данная метрика для объектов с разным количеством признаков выполняет процесс выравнивания, т.е. удлинения меньшего вектора признаков объекта, и далее рассматривает евклидовое расстояние между парами признаков. Этот процесс также называемый как Алгоритм динамической трансформации временной шкалы (DTW) [2], применяется для перевода пары объектов в новое метрическое пространство. Похожая концепция применима и для критерия гладкости [6]. Концепция DTW может быть применена в любых метрических классификаторах, таких как метод ближайших соседей NN или многослойный перцептрон Розенблатта MLP [4]. Она не только дает возможность работы с сигналами с разной длинной признакового описания, но и повышает устойчивость алгоритма к шумам, позволяет использовать меру близости не только как метрику, но и новое признаковое описание объекта.

Во многих работах наряду с SVM в различных комбинациях применяются алгоритмы предобработки сигналов, генерации признаков, на основе которых и проводиться дальнейшая классификация. Состоящий из двух стадий алгоритм HMM-SVM [5] применяет скрытую марковскую модель, а точнее совокупность скрытых марковских моделей для каждого отдельного класса, чтобы извлечь признаки из исходных "глобальных"признаков этого класса. Исходный вектор признаков может иметь различную длину для различных сигналов. Если так, то конечный вектор признаков определяется лишь тем, какие признаки были сгенерированны наборами HMM и тогда, помимо извлечения новых более информативных признаков, дающих вероятностную характеристику каждого класса и принадлежности данного объекта к классу, можно добиться эффекта уменьшения размерности признакового пространства. Иначе, мы добавляем извлеченные признаки к вектору глобальных и подаем на вход SVM с RBF ядром.

В данной работе будет рассмотрен вероятностный подход к распознаванию образов с помощью метода *Soft-margin SVM с критерием гладкости* решающего правила.

2 Вероятностная постановка задачи

2.1 Байесовское правило обучения

Рассмотрим вероятностное пространство вида $X \times Y \times W$, которое соответствует декартовому произведению множества объектов на множество классов принадлежности объектов умноженное на множество параметров вероятностной модели распределения объектов и их классов.

Обозначим $X^l = (\mathbf{x}_i; y_i)_{i=1}^N$, $\mathbf{x}_i \in X = \mathbb{R}^d$, $y_i \in Y = \{-1, 1\}$, где N - количество объектов, d - количество признаков объекта. X^l - есть генеральная совокупность реализаций объектов x_i и соответствующих ответов y_i. Пусть $\mathbf{w} = \begin{bmatrix}\mathbf{a}, & b\end{bmatrix}^\intercal$ - элемент множества $W = \mathbb{R}^{d+1}$, где $\mathbf{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$.

Исходя из предложенного метода вероятностной интерпретации генеральной совокупности при решении задачи SVM [7], рассмотрим отнесение объекта выборки к классу, как разделение объектов в пространстве $\mathbb{R}^d$ гиперплоскостью $\mathbf{a}^\intercal\mathbf{x} + b = 0$, соответствующей границе двух параметрических семейств - положительного класса

$y_i = 1$ и отрицательного класса $y_i = -1$. Далее, будем учитывая гипотезу компактности распределения классов выборки, определим параметрическую плотность несобственного совместного распределения объектов генеральной совокупности $\mathbf{x}_i$ и его класса y_i равной:

$$\varphi(\mathbf{x}_i, y_i|\mathbf{w}) = \varphi(\mathbf{x}_i, y_i|\mathbf{a}, b) = \\ = \varphi_{y_i}(\mathbf{x}_i|\mathbf{a}, b) = \begin{cases} 1, y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) \geq 1; \\ \exp[-\lambda(1 - y_i(\mathbf{a}^\intercal \mathbf{x}_i + b))], y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) < 1. \end{cases} \quad (1)$$

Несобственные распределения имеют особенность - интеграл от плотности распределения по всему пространству, на котором задано распределение, не равен единице,а бесконечен. Такая размытость распределения показывает свойство неопределенности распределения объектов в классе, находящихся за разделяющей полосой внутри своего класса, т.е. могут свободно находиться в любом месте в пределах области своего класса $\{x_i : y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) \geq 1\}$. В то же время, вероятность найти объект класса внутри разделяющей полосы ($\{x_i : |\mathbf{a}^\intercal \mathbf{x}_i + b| \leq 1\}$) или в полосе чужого класса ($\{x_i : y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) < -1\}$) экспоненциально снижается к нулю, при удалении от границы своего класса разделяющей полосы.

Под совместной плотностью распределения конечного множества векторов признаков объектов известных классов $(\mathbf{x}_i, y_i)$ в составе обучающей совокупности X^l, полученных при наблюдении, будем понимать плотность распределения выборки независимых реализаций этих двух распределений:

$$\Phi(X^l|\mathbf{a}, b) = \Phi(\mathbf{x}_i, y_i, i = 1, \ldots, N|\mathbf{a}, b) = \quad (2)$$

$$= \prod_{i=1}^{N} \varphi_{y_i}(\mathbf{x}_i|\mathbf{a}, b) = \left(\prod_{i:\ y_i=1} \varphi_1(\mathbf{x}_i|\mathbf{a}, b) \right) \left(\prod_{i:\ y_i=-1} \varphi_{-1}(\mathbf{x}_i|\mathbf{a}, b) \right). \quad (3)$$

Вектор параметров $\mathbf{w}$ задается априорным распределением вероятности на W для параметров распределений $\varphi_y(\mathbf{x}|\mathbf{a}, b)$, $y \in \{-1, 1\}$. Часть вектора параметров $\mathbf{a}$ задается многомерным нормальным распределением с нулевым вектором средних значений и матрицей ковариации вида $\sigma^2 \cdot B^{-1}$, т.е. положительно определенной и симметричной матрицей. Соответствующее b распределение является несобственным равномерным распределением равным единице по всей прямой $\mathbb{R}$. Совместное несобственное априорное распределение в пространстве параметров модели генеральной

совокупности имеет вид:

$$\psi(\mathbf{w}) = \psi(\mathbf{a}, b) = \psi(\mathbf{a}, b|\sigma^2, B) \propto \exp(-\frac{1}{2\sigma^2} \cdot \mathbf{a}^\intercal B\mathbf{a}). \tag{4}$$

Видно, что совместное распределение $\psi(\mathbf{a}, b)$ является также несобственным, так как не имеет конечного интеграла на множестве всех $b \in \mathbb{R}$.

Если по генеральной совокупности X^l проводить восстановление вектора параметров $\mathbf{w}$, исходя из разделения объектов по классам, то мы приходим к задаче нахождения максимума апостериорной плотности распределения параметров $\mathbf{a}$ и b относительно обучающей совокупности, что соответствует задачи максимума правдоподобия (ML). Для ее определения воспользуемся формулой Байеса:

$$p(\mathbf{a}, b|X^l) = p(\mathbf{a}, b|x_i, y_i, i = 1, \ldots, N) = \frac{\psi(\mathbf{a}, b)\Phi(X^l|\mathbf{a}, b)}{\int\limits_{\mathbb{R}^{d+1}} \cdots \int \psi(\mathbf{a}', b')\Phi(X^l|\mathbf{a}', b')\,\mathrm{d}\mathbf{a}'\mathrm{d}b'}. \tag{5}$$

Знаменатель справа не зависит от вектора параметров, таким образом:

$$p(\mathbf{a}, b|X^l) \propto \psi(\mathbf{a}, b)\Phi(X^l|\mathbf{a}, b) = \psi(\mathbf{a}, b) \left(\prod_{i:\ y_i=1} \varphi_1(\mathbf{x}_i|\mathbf{a}, b) \right) \left(\prod_{i:\ y_i=-1} \varphi_{-1}(\mathbf{x}_i|\mathbf{a}, b) \right). \tag{6}$$

Принцип максимизации плотности апостериорного распределения в пространстве параметров модели генеральной совокупности приводит к байесовскому правилу обучения, взяв логарифм плотности (6):

$$\begin{aligned} \ln p(\mathbf{a}, b|X^l) &= \ln \psi(\mathbf{a}, b) + \sum_{i:y_i=1} \ln \varphi_1(\mathbf{x}_i|\mathbf{a}, b) + \sum_{i:y_i=-1} \ln \varphi_{-1}(\mathbf{x}_i|\mathbf{a}, b) = \\ &= \ln \psi(\mathbf{a}, b) + \sum_{i=1}^{N} \ln \varphi_{y_i}(\mathbf{x}_i|\mathbf{a}, b) \to \max_{\mathbf{a},b} \end{aligned} \tag{7}$$

Для несобственных плотностей распределения векторов признаков объектов двух классов (1) и априорного несобственного распределения параметров разделяющей гиперплоскости (4) байесовский критерий обучения (7) примет вид:

$$\begin{aligned} Q = -\ln p(\mathbf{a}, b|X^l) &= \frac{1}{2\sigma^2} \cdot \mathbf{a}^\intercal B\mathbf{a} + \sum_{i=1}^{N} \lambda(1 - y_i(\mathbf{a}^\intercal \mathbf{x}_i + b))\,[y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) < 1] = \\ &= \frac{1}{2\sigma^2 \cdot \lambda} \cdot \mathbf{a}^\intercal B\mathbf{a} + \sum_{i=1}^{N} max(0, 1 - y_i(\mathbf{a}^\intercal \mathbf{x}_i + b)) \to \min_{\mathbf{a},b} \end{aligned} \tag{8}$$

2.2 Soft-margin SVM и вероятностный подход

В формуле (8) обозначим $\xi_i = max(0, 1-y_i(\mathbf{a}^\intercal \mathbf{x}_i+b))$. Получим SVM в виде задачи оптимизации, с ограничениями типа неравенства называемой **Soft-margin SVM**:

$$\begin{cases} \dfrac{1}{2C}\mathbf{a}^\intercal B\mathbf{a} + \sum\limits_{i=1}^{N} \xi_i \to \min\limits_{\mathbf{a},\xi} \\ y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i,\ i = 1, \ldots, N \\ \xi_i \geq 0,\ i = 1, \ldots, N \end{cases} \tag{9}$$

Такой подход к задаче классификации отражен в работе В. Вапника [3]. Он предполагает рассматривать пространство признаков объекта обучающей выборки, в котором они не могут быть разделены гиперплоскостью на два подпространства, и в каждом объекты только одного класса. То есть ищется гиперплоскость, которая бы минимизировала ошибку разделения классов на ней, но не обязательно приравнивала ее к нулю. Ошибками, в данных нами обозначениях, являются неотрицательные величины $\xi_i = max(0, 1 - y_i(\mathbf{a}^\intercal \mathbf{x}_i + b))$. Ошибка ξ_i будет нулевой, если будет верно неравенство $1 - y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) \leq 0 \Leftrightarrow y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) \geq 1$, т.е. отступ любого объекта соответствующего класса будет минимальным и выходящим за разделяющую полосу. соответственно отступом (margin) называют $M(x_i, y_i) = y_i(\mathbf{a}^\intercal \mathbf{x}_i + b)$, который равен проекции вектора признаков на нормаль к разделяющей гиперплоскости, т.е. расстояние от точки x до нее. Рассматривая задающуюся нормалью $\mathbf{a}$ и сдвигом b искомую разделяющую гиперплоскость $\mathbf{a}^\intercal \mathbf{x}_i + b = 0$, и ответом классификатора для нового объекта будет являться $a(\mathbf{x}) = \text{sign}\,(\mathbf{a}^\intercal \mathbf{x} + b)$.

Минимизируемый функционал качества в классической задаче Soft-margin SVM не включает матрицу B (см. пункт 2.4). Помимо минимизации ошибки разделения объектов гиперплоскостью идея SVM состоит в том, чтобы максимизировать расстояние от ближайших объектов выборки до разделяющей гиперплоскости. В работе Вапника показано, что ширина разделяющей полосы, определяемая расстоянием между объектами двух классов с отступами равными 1 и -1, равна $\dfrac{2}{\|\mathbf{w}\|}$. Для удобства решения задачи оптимизации ищется максимум $\gamma(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2}$. Таким образом функционал полностью состоит из суммы двух слагаемых, первый из которых - это функционал потери $L(X^l, \mathbf{w}) = \sum_{i=1}^{N} \xi_i$, а второй - регуляризатор $\gamma(\mathbf{w})$, т.е. ограничения накладываемые на значения искомого вектора параметров модели.

Проводя аналогию с вероятностным подходом, видно, что формула распределения объектов обучающей совокупности (1) соответствует функции потерь $L(X^l, \mathbf{w})$ на этой совокупности, а априорное распределение параметров модели $\psi(\mathbf{w})$ (4) - регуляризатору $\gamma(\mathbf{w})$.

В отличие от классической постановки, при учете критерия гладкости регуляризатор имеет другой вид $\gamma(\mathbf{w}) = \frac{1}{2C}\mathbf{a}^\intercal B\mathbf{a}$. Рассмотрим этот вид задачи позднее (см. пункт 2.4). Как уже было сказано, этап обучения состоит из минимизации функционала, состоящего из двух слагаемых:

$$\begin{aligned} &Q(X^l, \mathbf{w}) = \gamma(\mathbf{w}) + C \cdot L(X^l, \mathbf{w}) \to \min_{\mathbf{w}} \\ &\gamma(\mathbf{w}) = \frac{1}{2C}\mathbf{a}^\intercal B\mathbf{a} = -\ln \psi(\mathbf{w}); \\ &L(X^l, \mathbf{w}) = \sum_{i=1}^{N} \xi_i = -\sum_{i=1}^{N} \ln \varphi_{y_i}(\mathbf{x}_i|\mathbf{a}, b). \end{aligned} \tag{10}$$

Значимым на этапе обучения является также и положительный гиперпараметр $C = \sigma^2\lambda > 0$. Этот гиперпараметр, также как и матрица B^{-1} связывает параметры эвристического подхода к классификации SVM и байесовского критерия обучения, но в отличие от нее является глобальным параметром дисперсии в пространстве векторов параметров и признаков. В отличие от ковариационной матрицы B^{-1}, которая показывает степень зависимости распределений координат вектора параметров, глобальная дисперсия C с уменьшением дает меньшую область возможных значений вектора параметров модели, т.е. исходя из уменьшения σ в распределении (4) увеличивается и норма вектора $\mathbf{a}$. То же можно сказать и о гиперпараметре λ, который, исходя из (1), при своем уменьшении увеличивает вероятность появления объектов вне своего класса и внутри разделяющей полосы. Это однозначно соответствует ситуации, когда при уменьшении параметра C в эвристическом критерии обучения Soft-margin SVM (9), процесс минимизации всего функционала качества Q приводит к сильному уменьшению регуляризатора $\gamma(\mathbf{w})$ и увеличению функционала потерь.

Итак, для соответствия задачи, при которой проводиться поиск наилучшей обобщающей способности при малом количестве векторов, функционал потерь, т.е. сумма ошибок разделения объектов гиперплоскостью, может быть достаточно большой, но вместо этого, регуляризация помогает уменьшить возможные отклонения от предполагаемой формы множества параметров или объектов. В пункте 2.4, рассмотрим

подробнее критерий гладкости необходимый для наложения правила на вектор параметр и учет этого правила в регуляризаторе. Этим способом снижается качество распознавания на обучающей выборке и сужается область допустимых значений решающего правила, позволяя улучшить качество распознавания на генеральной совокупности.

2.3 Двойственная задача

Рассмотрим подробнее вид оптимизационной задачи Soft-margin SVM (9). Данная задача является классической задачей квадратичного программирования:

$$\begin{cases} (1/2)\mathbf{x}^\intercal P\mathbf{x} + \mathbf{q}^\intercal\mathbf{x} \to \min\limits_x \\ G\mathbf{x} \leq h \\ A\mathbf{x} = \mathbf{b} \end{cases} \tag{11}$$

Минимизируемый функционал состоит из квадратичной части по первым d переменным и линейной - по следующим N переменным, функции участвующие в ограничениях типа неравенств линейные, значит это также и задача выпуклого программирования, при существовании локального минимума он также будет являться глобальным минимумом. Для поиска глобального минимума необходимо выполнение условий Каруша — Куна — Таккера и условий регулярности задачи. Условия регулярности выполняются из свойств линейности ограничений и сильной двойственности (доказывается через выполнения условия Слейтора). Далее рассмотрим условия Каруша — Куна — Таккера:

Функция Лагранжа для данной задачи будет иметь вид:

$$L(\mathbf{a}, b, \xi, \lambda, \mu) = \frac{1}{2}\mathbf{a}^\intercal B\mathbf{a} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\lambda_i\left[y_i(\mathbf{a}^\intercal\mathbf{x}_i + b) - 1 + \xi_i\right] - \sum_{i=1}^{N}\mu_i\xi_i \tag{12}$$

, где $\lambda_i \geq 0, \mu_i \geq 0, i = 1, \ldots, N$- двойственные переменные. По условиям ККТ градиенты по переменным прямой задачи:

$$\begin{cases} \nabla_{\mathbf{a}} L = \frac{1}{2}(B + B^\intercal)\mathbf{a} - \sum\limits_{i=1}^{N} \lambda_i y_i \mathbf{x}_i = 0 \\ \nabla_{\xi} L = \lambda + \mu = \mathbf{0} \\ \nabla_b L = \sum\limits_{i=1}^{N} \lambda_i y_i = 0 \end{cases} \Leftrightarrow \begin{cases} \mathbf{a} = \sum\limits_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \\ 0 \leq \lambda_i \leq C,\, i = 1, \ldots, N \\ 0 \leq \mu_i \leq C,\, i = 1, \ldots, N \\ \sum\limits_{i=1}^{N} \lambda_i y_i = 0 \end{cases} \tag{13}$$

Существование обратной матрицы и ее симметричность доказывается в пункте 2.4. Двойственная функция будет равняться:

$$\begin{aligned} W(\lambda, \mu) &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i y_j (B^{-1}\mathbf{x}_i)^\intercal B (B^{-1}\mathbf{x}_j)) \lambda_i \lambda_j + C \sum_{i=1}^{N} \xi_i - \\ &- \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i y_j (B^{-1}\mathbf{x}_i)^\intercal \mathbf{x}_j) \lambda_i \lambda_j - \sum_{i=1}^{N} \lambda_i y_i b + \sum_{i=1}^{N} \lambda_i = (13) = \\ &= \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i y_j \mathbf{x}_i^\intercal B^{-1} \mathbf{x}_j) \lambda_i \lambda_j \end{aligned} \tag{14}$$

И двойственную задачу к прямой, зависящую только от λ, мы можем записать в виде:

$$\begin{cases} W(\lambda) = \frac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} (y_i y_j \mathbf{x}_i^\intercal B^{-1} \mathbf{x}_j) \lambda_i \lambda_j - \sum\limits_{i=1}^{N} \lambda_i \to \min\limits_{\lambda} \\ \sum\limits_{i=1}^{N} \lambda_i y_i = 0; \\ 0 \leq \lambda_i \leq C,\, i = 1, \ldots, N. \end{cases} \tag{15}$$

Соответствующий ему вектор параметров $\mathbf{a}$ и классификатор зависят от двойственной переменной λ:

$$\mathbf{a} = \sum_{i=1}^{N} \lambda_i y_i (B^{-1}\mathbf{x}_i); \tag{16}$$

$$a(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i^\intercal B^{-1} \mathbf{x}_j + b \right). \tag{17}$$

Из условий дополняющей нежесткости ККТ:

$$\begin{cases} \lambda_i \left(y_i (\mathbf{a}^\intercal \mathbf{x}_i + b) - 1 + \xi_i \right) = 0,\, i = 1, \ldots, N; \\ \mu_i \xi_i = 0,\, i = 1, \ldots, N. \end{cases}$$

Можно определить опорные вектора - объекты $\mathbf{x}_i$, у которых соответствующие им двойственные переменные положительные. При этом, если $\lambda_i \neq C$, то соответствующие $\mu_i > 0 \Rightarrow \xi_i = 0 \Rightarrow y_i(\mathbf{a}^\intercal \mathbf{x}_i + b) - 1 = 0$. Тогда можно взять все вектора x_i с $0 < \lambda_i < C$, домножим на последнее равенство и получим:

$$\sum_{i:0<\lambda_i<C} \lambda_i(\mathbf{a}^\intercal \mathbf{x}_i + b) = \sum_{i:0<\lambda_i<C} \lambda_i y_i; \qquad \Rightarrow$$

$$b = \frac{-\left(\sum\limits_{i:0<\lambda_i<C} \lambda_i \mathbf{a}^\intercal \mathbf{x}_i + \sum\limits_{i:0<\lambda_i<C} \lambda_i y_i\right)}{\sum\limits_{i:0<\lambda_i<C} \lambda_i}; \qquad \Rightarrow$$

$$b = \frac{-\sum\limits_{i:\lambda_i\neq C} \lambda_i \mathbf{a}^\intercal \mathbf{x}_i + C \sum\limits_{i:\lambda_i=C} y_i}{\sum\limits_{i:\lambda_i\neq C} \lambda_i}$$

Двойственную задачу удобнее решать чем прямую по нескольким причинам. Во-первых, как уже было сказано, концепция опорных векторов была изобретена именно в этом виде и предполагала нахождение нескольких объектов из обучающей совокупности, определяющих вид и форму гиперплоскости. Во-вторых, что более существенно, критерий гладкости в данной задачи представляет собой переход в новое спрямляющее пространство. Рассмотрим свойства матрицы B стоящей в прямой задаче при векторе параметров $\mathbf{a}$, а также ее обратная матрица при векторе признаков объекта $\mathbf{x}_i$.

2.4 Критерий гладкости и Kernel-trick (The Method of Convolution)

Метод SVM имеет свойство использовать лишь малое количество объектов выборки называемых опорными (см. пункт 2.3). Этот метод хорошо работает только с дополнительными ограничений на вид разделяющей гиперплоскости, потому как опорные вектора не всегда могут показать форму границ областей различных классов на малонаполненных выборках. Рассматривая объекты выборки X^l, представляющие собой результат упорядоченного измерения некоторого свойства объекта вдоль координаты той или иной природы, есть основания полагать, что соседние признаки несут почти идентичную информацию о принадлежности объекта к определенному классу. Данное свойство переносится на вид априорного распределения в пространстве векторов параметра $\mathbf{a} = (a_1, a_2, \ldots, a_d)$, которое накладывает ограничение на

плавное изменение коэффициентов a_i при увеличении индекса i в признаковом описании объекта $\mathbf{x} = (x_1, x_2, \ldots, x_d)$. Сравнивая соседние по индексу коэффициенты, можно записать их взаимное различие в виде дополнительного регуляризатора ($\frac{1}{2}$ используется для удобства):

$$J(\mathbf{a}) = \frac{1}{2}\sum_{i=1}^{d-1}(a_i - a_{i+1})^2 = \frac{1}{2}\left(a_1^2 - 2\sum_{i=1}^{d-1} a_i a_{i+1} + 2\sum_{i=1}^{d-1} a_i^2 + a_n^2\right) = \\ = \frac{1}{2}\sum_{i=1}^{N}\widetilde{B}_{ij}a_i a_j = \frac{1}{2}\mathbf{a}^\intercal\widetilde{B}\mathbf{a} \geq 0. \tag{18}$$

В матричном виде матрица в критерии гладкости $\widetilde{B}$ принимает вид:

$$\widetilde{B} = \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & & & \vdots \\ 0 & -1 & 2 & -1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & -1 & 2 & -1 & 0 \\ \vdots & & & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{d\times d} \tag{19}$$

Очевидно, что такая постановка может привести к бесконечному множеству вырожденных решений: когда $a_i = a$, $i = 1, \ldots, N$, т.е. не различающиеся коэффициенты, $J(\mathbf{a}) = 0$.

Поэтому критерий гладкости необходимо использовать вместе с квадратичным функционалом классической постановки SVM (10), тогда результирующий регуляризатор в функционале качества будет иметь вид:

$$\gamma(\mathbf{w}) = \frac{1}{2}\mathbf{a}^\intercal B\mathbf{a} = \frac{1}{2}\mathbf{a}^\intercal\left(I + \alpha\widetilde{B}\right) \tag{20}$$

, где $\alpha \geq 0$ будем считать параметром регуляризации, который при увеличении одновременно уменьшает критерий гладкости как слагаемое функционала. Данный вид регуляризатора $\gamma(\mathbf{w})$ используется в прямом виде постановки Soft-margin SVM (9).

Таким образом окончательная матрица гладкости, используемая в прямой задаче (9):

$$B = I + \alpha\widetilde{B} > 0 \tag{21}$$

, являющейся положительно определенной матрице: для любого ненулевого вектора $\mathbf{v} \neq 0, \mathbf{v}^\intercal(I + \alpha\widetilde{B})\mathbf{v} = \mathbf{v}^\intercal\mathbf{v}\cdots > 0$. Также, данная матрица является симметрической. Обращаясь к интерпретации матрицы B в пункте 2.2 будучи ковариационной матрицей многомерного случайного вектора она должна быть симметричной и неотрицательно определенной.

Из того, что матрица B положительно определенная и симметричная, можно найти ортонормированный базис из ее собственных векторов, соответствующих положительным собственным значениям: $B = Q^\intercal\Lambda Q$, где Q - матрица состоящая из ортонормированных собственных векторов, а Λ - диагональная матрица с положительными собственными значениями на диагонали. При этом обратная к данной матрице будет являться матрица вида $B^{-1} = Q^\intercal\Lambda^{-1}Q$, причем она будет также положительно определенной и симметричной: $det\,(B^{-1}) = (detB)^{-1} > 0; (B^{-1})^\intercal = (B^\intercal)^{-1} = B^{-1}$. Помимо обратной матрицы рассмотрим ее квадратный корень. Из единственности разложения по ортонормированному базису $B^{-1/2} = (B^{-1})1/2 = Q^\intercal\Lambda^{-1/2}Q$, отсюда видно, что матрица $B^{-1/2}$ также будет симметричной.

Приведем пример вида матрицы ковариации B^{-1}. Меняя параметр гладкости α, можно получить разную степень зависимости между элементами.

При $\alpha = 1$ (количество измерений $d = 10$):

$$\begin{bmatrix}
0.618 & 0.236 & 0.0902 & 0.0344 & 0.0132 & 0.00503 & 0.00192 & 0.000739 & 0.000296 & 0.000148 \\
0.236 & 0.472 & 0.18 & 0.0689 & 0.0263 & 0.0101 & 0.00384 & 0.00148 & 0.000591 & 0.000296 \\
0.0902 & 0.18 & 0.451 & 0.172 & 0.0658 & 0.0251 & 0.00961 & 0.0037 & 0.00148 & 0.000739 \\
0.0344 & 0.0689 & 0.172 & 0.448 & 0.171 & 0.0653 & 0.025 & 0.00961 & 0.00384 & 0.00192 \\
0.0132 & 0.0263 & 0.0658 & 0.171 & 0.447 & 0.171 & 0.0653 & 0.0251 & 0.0101 & 0.00503 \\
0.00503 & 0.0101 & 0.0251 & 0.0653 & 0.171 & 0.447 & 0.171 & 0.0658 & 0.0263 & 0.0132 \\
0.00192 & 0.00384 & 0.00961 & 0.025 & 0.0653 & 0.171 & 0.448 & 0.172 & 0.0689 & 0.0344 \\
0.000739 & 0.00148 & 0.0037 & 0.00961 & 0.0251 & 0.0658 & 0.172 & 0.451 & 0.18 & 0.0902 \\
0.000296 & 0.000591 & 0.00148 & 0.00384 & 0.0101 & 0.0263 & 0.0689 & 0.18 & 0.472 & 0.236 \\
0.000148 & 0.000296 & 0.000739 & 0.00192 & 0.00503 & 0.0132 & 0.0344 & 0.0902 & 0.236 & 0.618
\end{bmatrix}$$

Здесь наблюдается довольно большая дисперсия у каждой компоненты вектора параметров модели $\mathbf{a}$, хотя ее значения и меньше единицы, что обусловлено функционалом, уменьшающим норму $\mathbf{a}$ в решающем правиле. В тоже время происходит быстрое уменьшение ковариации $B_{ij}^{-1} = cov(a_i, a_j)$ между двумя компонентами вектора. Это обусловлено тем фактом, что при данном выборе гиперпараметра α мы не сильно сглаживаем вектор.

При $\alpha = 100$:

$$\begin{bmatrix}
0.126 & 0.118 & 0.11 & 0.104 & 0.0983 & 0.0939 & 0.0904 & 0.0878 & 0.0861 & 0.0853 \\
0.118 & 0.119 & 0.111 & 0.105 & 0.0993 & 0.0948 & 0.0913 & 0.0887 & 0.087 & 0.0861 \\
0.11 & 0.111 & 0.114 & 0.107 & 0.101 & 0.0967 & 0.0931 & 0.0905 & 0.0887 & 0.0878 \\
0.104 & 0.105 & 0.107 & 0.11 & 0.104 & 0.0996 & 0.0959 & 0.0931 & 0.0913 & 0.0904 \\
0.0983 & 0.0993 & 0.101 & 0.104 & 0.108 & 0.103 & 0.0996 & 0.0967 & 0.0948 & 0.0939 \\
0.0939 & 0.0948 & 0.0967 & 0.0996 & 0.103 & 0.108 & 0.104 & 0.101 & 0.0993 & 0.0983 \\
0.0904 & 0.0913 & 0.0931 & 0.0959 & 0.0996 & 0.104 & 0.11 & 0.107 & 0.105 & 0.104 \\
0.0878 & 0.0887 & 0.0905 & 0.0931 & 0.0967 & 0.101 & 0.107 & 0.114 & 0.111 & 0.11 \\
0.0861 & 0.087 & 0.0887 & 0.0913 & 0.0948 & 0.0993 & 0.105 & 0.111 & 0.119 & 0.118 \\
0.0853 & 0.0861 & 0.0878 & 0.0904 & 0.0939 & 0.0983 & 0.104 & 0.11 & 0.118 & 0.126
\end{bmatrix}$$

Как видно, дисперсия каждой компоненты значительно уменьшилась, для того, чтобы она и соседние с ней компоненты не сильно отличались как от нулевого век-

тора, так и друг от друга. Увеличение степени гладкости наблюдается при рассмотрении ковариации "дальних"компонент вектора - для a_i компоненты ее дисперсия и ковариация с дургими компонентами незначительно отличаются при $\alpha = 100$, чего не скажешь при меньшем значении параметра гладкости α. То есть, близость позиций координат вектора становится менее значимой, чем общая сглаженность их значений.

Интересно показать, как меняется форма множества векторов $\mathbf{a}$ в пространстве (к примеру при $d = 10$). Сгенерируем с помощью многомерного нормального распределения с нулевым вектором матожидания и матрице ковариации B^{-1} и возьмем проекцию на две из d координат. Все описанные зависимости отражены в пространстве координат вектора $\mathbf{a}$ на рис. 1 и 2.

В работе Вапника В. [3] описывается подход называемый The Method of Convolution [3], или по-другому - ядровой переход. Рассматривается переход из исходного пространства в спрямляющее, в каждом определена функция скалярного произведения. Спрямляющее пространство используется для удобства классификации, так как двойственная задача (15) зависит от скалярных произведений признаков объектов и таким образом линейная классификация может проводиться в спрямляющем пространстве, в то время как в исходном она имеет нелинейный вид. Таким образом найдем вид $\phi : \mathbb{R}^d \to \mathbb{R}^d$, такой что использованная нами матрица гладкости B задействована в скалярном произведении:

$$\begin{aligned} &\mathbf{x}_i^\intercal B^{-1}\mathbf{x}_j = (B^{-1}\mathbf{x}_j \cdot \mathbf{x}_i) = ((B^{-1/2}\mathbf{x}_j \cdot (B^{-1/2}\mathbf{x}_i)) = \\ &= (\phi(\mathbf{x}_j) \cdot \phi(\mathbf{x}_i)) \Rightarrow \phi(\mathbf{x}_i) = (B^{-1/2}\mathbf{x}_i). \end{aligned} \tag{22}$$

Из единственности разложения по ортонормированному базису верно:

$$B^{-1/2} = Q^\intercal \Lambda^{-1/2} Q$$

Для нее будут верны все свойства, что и для матрицы B^{-1} - симметричность, положительная определенность.

На рисунке 3 показаны модельные данные, сгенерированные в пункте 3, в своем исходном пространстве гладких синусоид размерности $d = 50$. Описание каждого класса синусоид проводиться по различным частотам ω и сдвигам ϕ. Из-за небольшого различия этих гиперпараметров, классы объектов тоже не сильно отличаются.

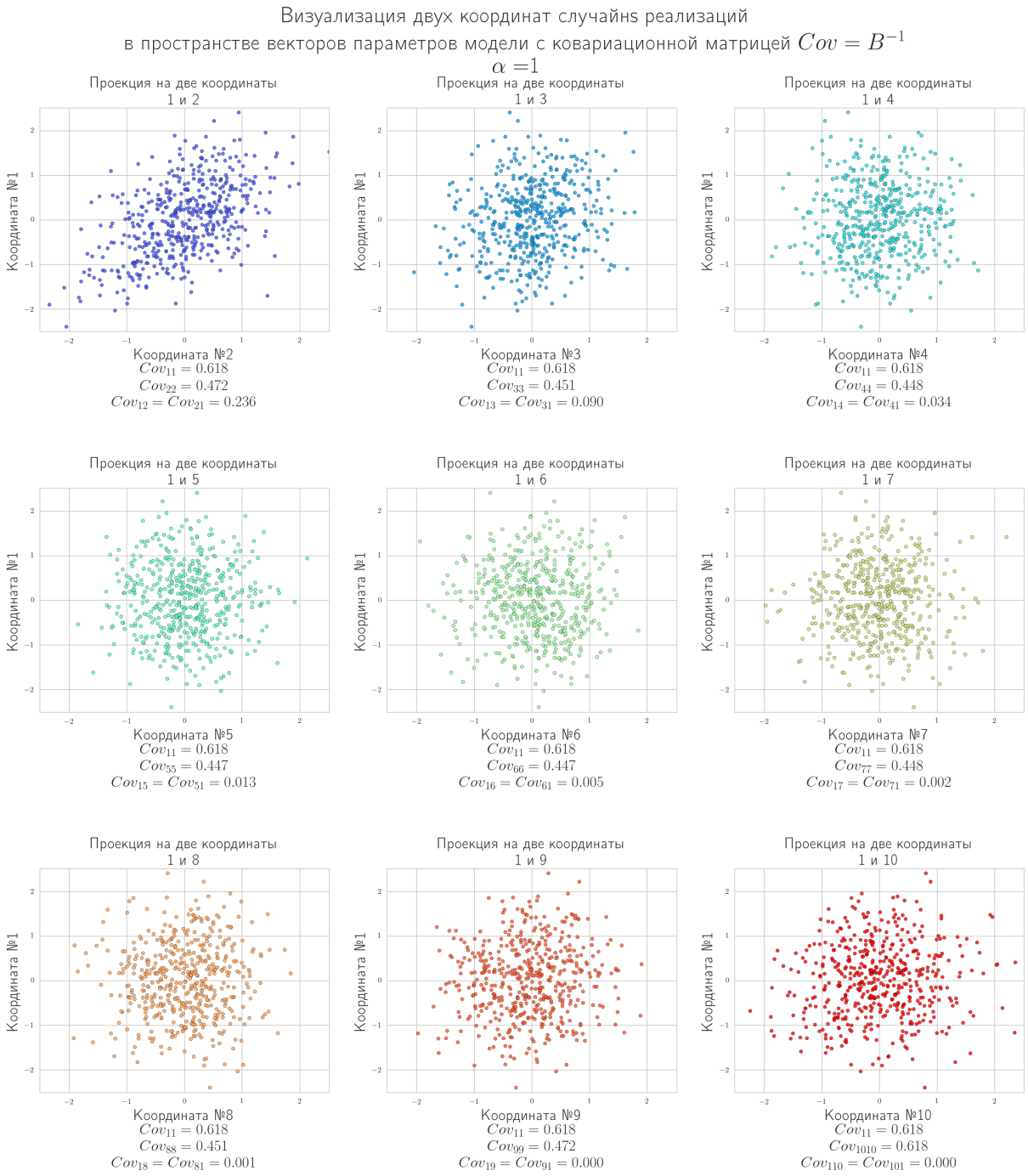


Рис. 1: Проекции пространства векторов **a** на две из 10 координат, при $\alpha = 1$

Добавив Гауссов шум для каждого из d измерений мы еще сильнее сглаживаем различия между классами.

Тем не менее в спрямляющем пространстве, определяемом матрицей гладкости B (на рис. 4) значения зашумленных синусоид сильно сглаживается, одновременно приближаясь к средним значениям в новом-спрямляющем пространстве. Стоит отме-

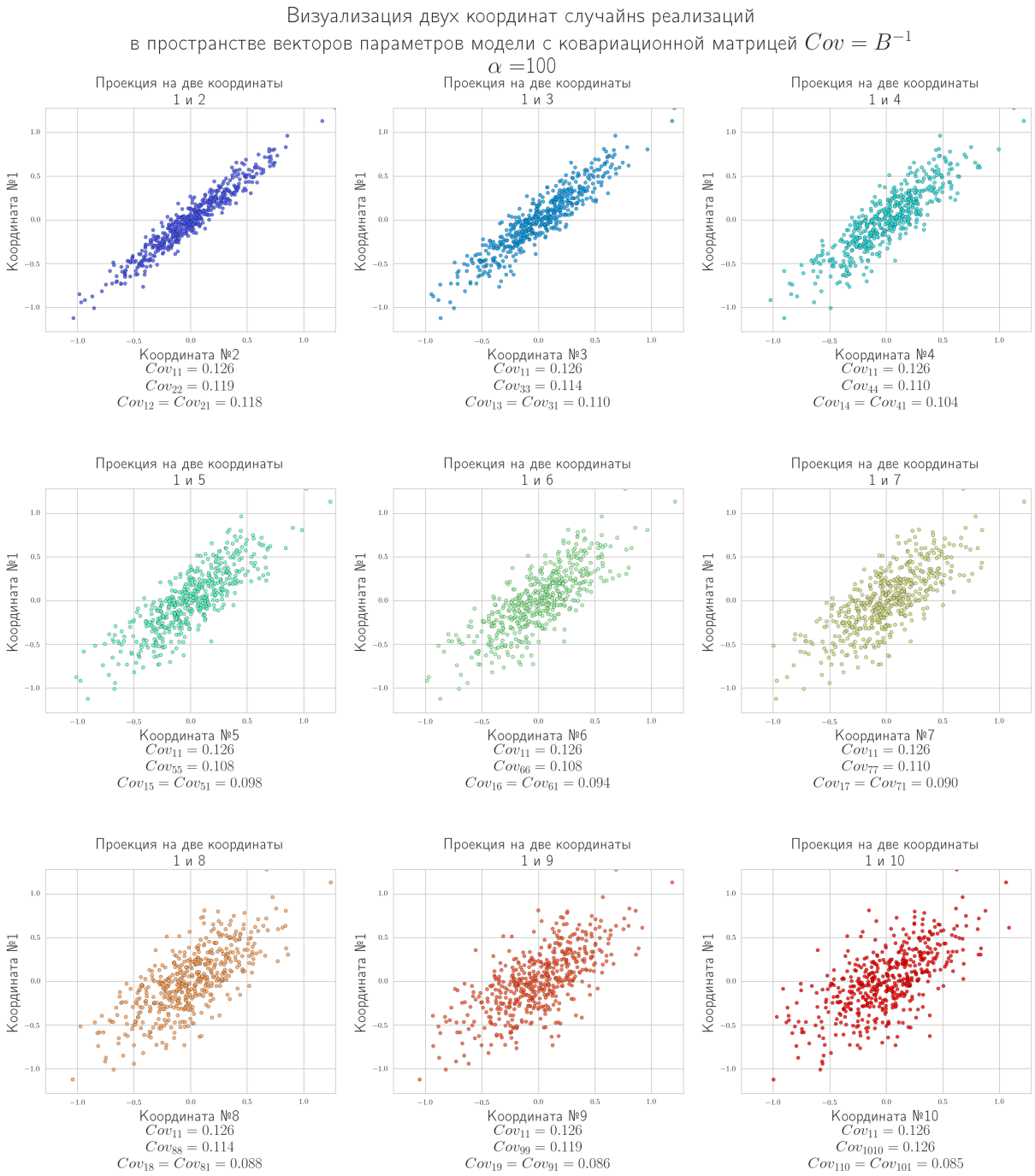


Рис. 2: Проекции пространства векторов **a** на две из 10 координат, при $\alpha = 100$

тить, что и средние значения синусоид были сильно изменены в новом-спрямляющем пространстве, но из-за сближения средних значений и зашумленных легче определить принадлежность нового объекта к тому или иному классу.

Суммируя сказанное выше, оказывается, что наша матрица гладкости, участвующая в априорной плотности распределения (4), определяется матрице B^{-1}, кото-

рая положительно определенная, симметричная и при увеличении параметра. Но она учитывает зависимости всех компонент признака параметра **a**: если сравнить диагональные элементы с вне диагональными, удаляясь от диагонали все дальше, наблюдается заметное уменьшение значения ковариации между элементами, а значит и уменьшается зависимость между рассматриваемой компонентой вектора **a** и остальными при удалении от исходной. Данная зависимость отражает эффект предполагаемой гладкости компонент.

К тому же, решения вопроса выбора ядра для реализации SMSVM сводиться к использовании ковариационной матрицы или ее квадратному корню, которые наилучшим образом показывают эффект гладкости в новом спрямляющем пространстве. Также, как было сказано в введении 1, сначала необходимо привести все объекты, являющиеся буквами алфавита, к единому признаковому пространству, что тоже подразумевает ядровой переход. Так как в данной работе не рассматривались реальные данные, здесь не приводится вид ядра и спрямляющего пространства при выравнивании DTW, в котором буквы будут иметь одинаковое количество измерений, описанное во введении 1.

3 Эксперименты на модельных данных

3.1 Исходные данные и условия эксперимента

Для генерирования модельных данных были выбраны синусоиды с различными частотами ω и сдвигами ϕ так, чтобы формы кривых синусоид отличались незначительно. Также для каждого из $d = 50$ значений синусоид, накладывался шум с параметром среднего значения в исходных точках синусоид, чтобы возможные различия между объектами классов еще сильнее уменьшались. Несмотря на это предполагалось, что по средним значениям координат различных объектов, можно оценить принадлежность к тому или иному классу, но для этого было необходимо применить критерий гладкости, усредняющий и приближающий зашумленные вектора к своим средним значениям. На рис. 3 и 4 наглядно показан вид этих данных как в исходном, так и в спрямляющем пространстве ядрового перехода с матрицей $B^{-1/2}$.

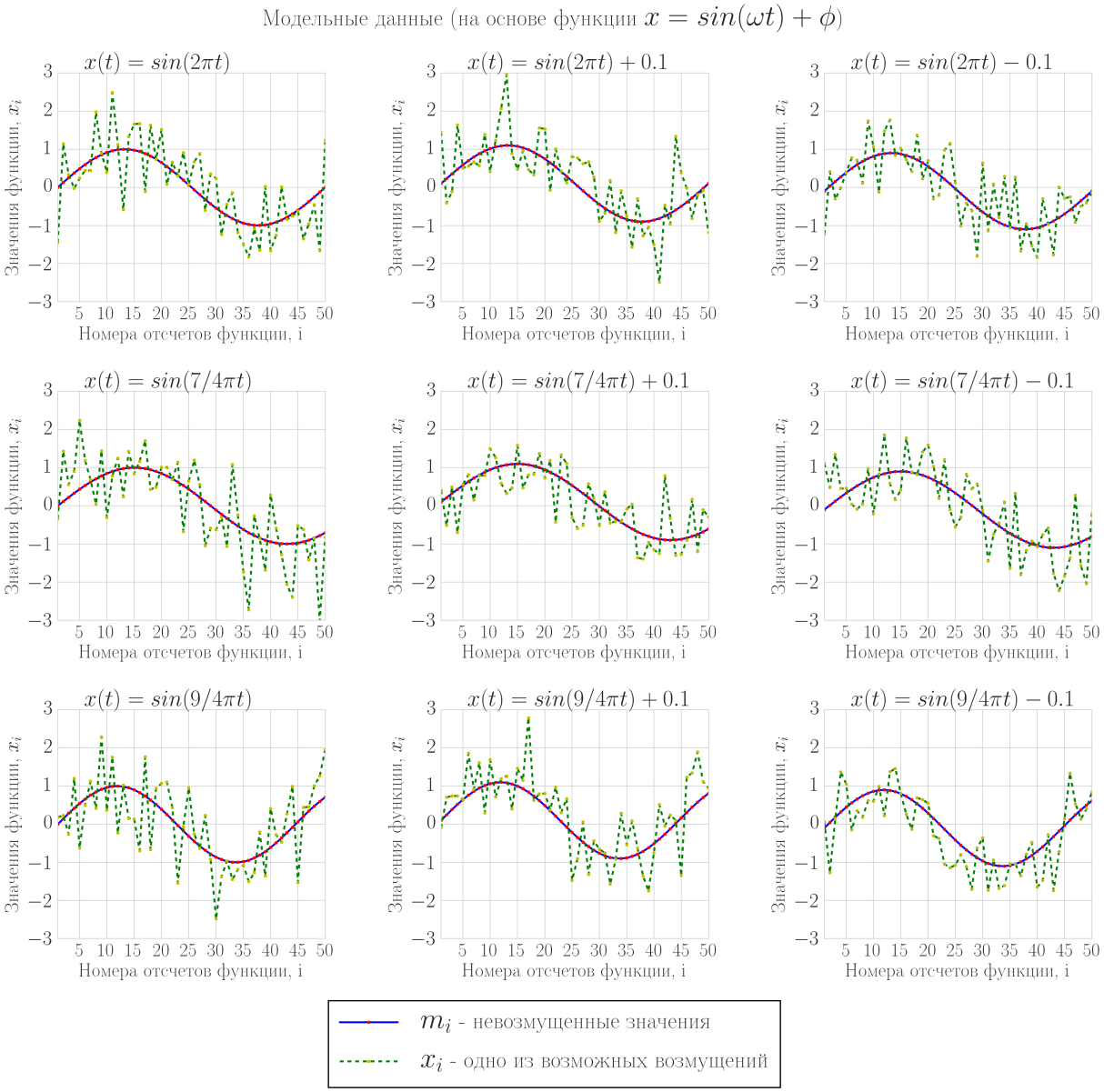


Рис. 3: Модельные данные в исходном пространстве

Далее проводилась попарная классификация всех классов. Для этого настраивался SMSVM с ядром на подвыборках объектов различных классов. Из предположения о малой наполненности исходной выборки, и сравнимого с размерность признакового пространства данных d количества объектов N в классифицируемой выборке, выбирались случайные подвыборки объектов двух классов (равномощные для обоих классов) размерами $N = 25, 50, 100, 150, 300$ для контроля и столько же для обучения.

Гиперпараметр $C = 0.1$ брался для того, чтобы оставалась возможность достаточно больших значений ошибок классификации на обучении $\sum_{i=1}^{N} \xi_i$, но критерий глад-

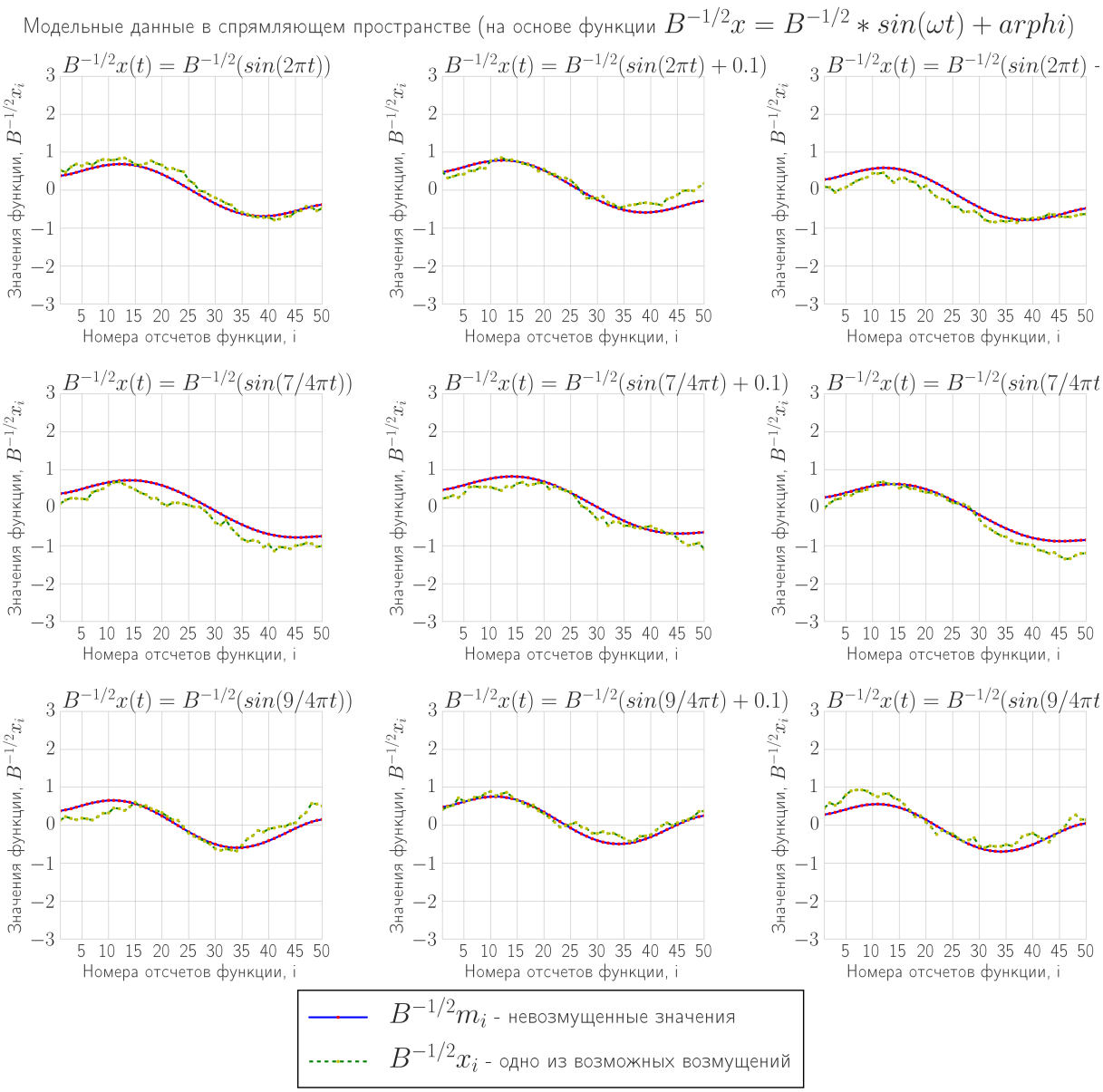


Рис. 4: Модельные данные в спрямляющем пространстве

кости и критерий разделения разделяющей полосы имели больший эффект. Точность определялась как отношение правильно классифицируемых объектов контрольной выборки к их общему количеству. Для уменьшения случайности выбора подвыборок бралось усреднение точностей, полученных на 50 различных разделений выборки на контроль и обучение.

Проводилось исследование оптимальных значений параметра гладкости α для разного соотношения между количеством объектов в выборке и размерности признакового пространства. При $\alpha = 0$ отключался критерий гладкости в применяемой

модели линейной классификации. Точности всех бинарных классификаций объектов усреднялись для получения среднего показателя точность для одного и того же параметра гладкости α.

3.2 Результаты эксперимента

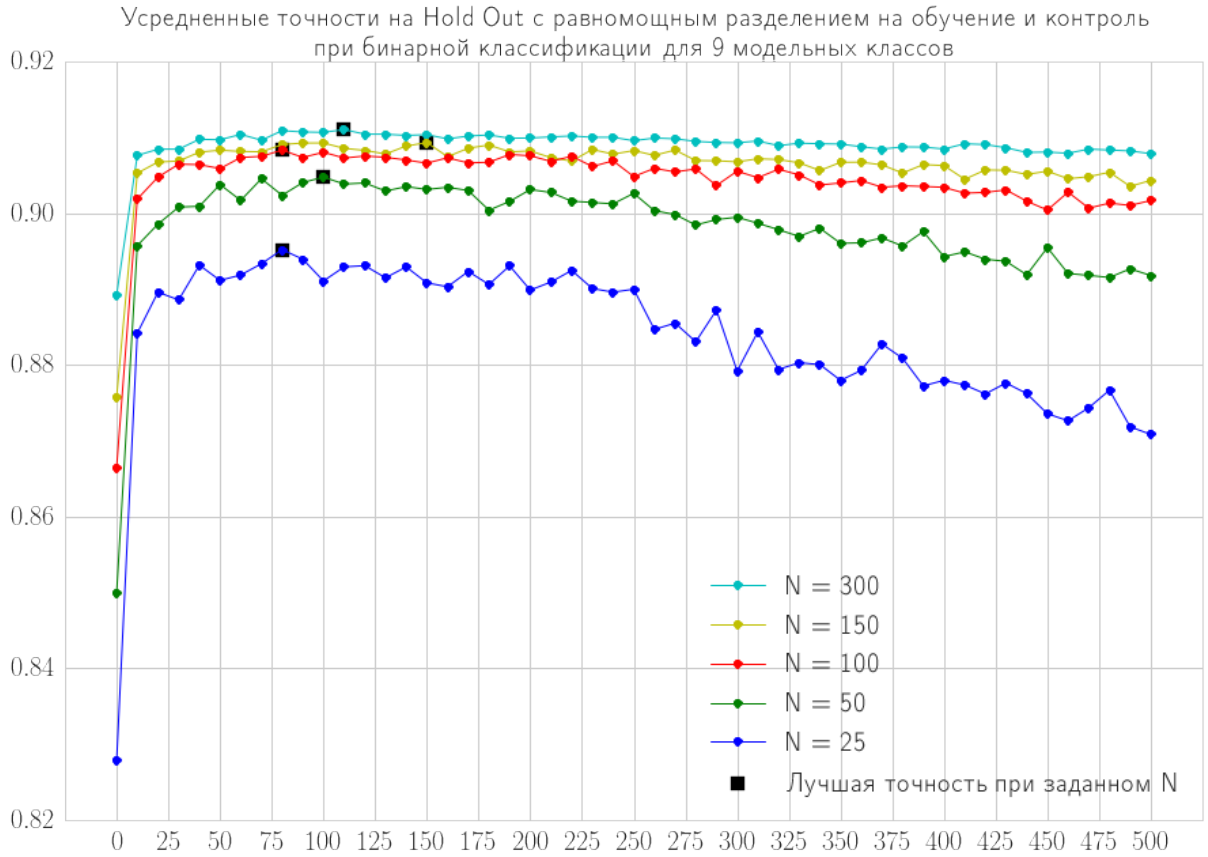


Рис. 5: Усредненная точность бинарной классификации в $d = 50$ -мерном пространстве признаков при увеличении параметра гладкости α

.

По результатам приведенным на рис. 5 хорошо видно, что даже при незначительном увеличении $\alpha = 5$ наблюдается возрастание точности классификации. Но при очень больших значениях α происходит эффект переобучения, так как критерий гладкости начинает сильно ограничивать вид нашего пространства векторов параметров $\mathbf{a}$.

Сравнивая показатели точности, для разных соотношений между количеством объектов в обучающей выборке и признаковым пространством, хорошо заметны неко-

торые зависимости. Так, критерий гладкости дает наиболее значительный прирост в точности для малонаполненных выборок, а при оптимальном подборе α в результате получается точность даже выше, чем при классификации на достаточно больших обучающих выборках, но без учета критерия гладкости. С другой стороны, момент переобучения достигается раньше на таких малонаполненных выборках, чего можно избежать применением кросс-валидации по параметру α.

3.3 Выводы

В общем, для всех размеров подвыборок оптимальными α будут значения в отрезке $[80 : 125]$, но, данный показатель может варьироваться исходя из вида, свойства и формы рассматриваемых объектов-сигналов. В случае выборки сравнимой по количеству объектов со своим признаковым пространством ($d = 50, N = 50$), вместе со значительными приростом точности в сравнении с нулевым параметром гладкости наблюдаются высокие показатели близкие к показателям на более больших выборках, т.е. критерий гладкости становится менее значимым при увеличении количества объектов.

4 Заключение

Критерий гладкости описанный в данной работе имеет схожую вероятностную интерпретацию с эвристическим подходом. Рассмотренный вид объектов-параметров модели классификации в пространстве, ограниченном свойствами гладкости дал вероятностную интерпретацию зависимости значений его компонент друг от друга, а также изменение этой зависимости при варьировании степени гладкости решающего правила.

Проведенные эксперименты на модельных данных дали численные результаты, при которых показана значимость применения критерия гладкости для задач с малым количеством объектов в обучении.

Рассмотренный вид регуляризации решающего правила дал один из способов использования априорной информации в задачах анализа данных.

В последующих работах предполагается осветить вопросы выравнивания признакового описания сигналов и использования его для работы с реальными данными.

Список литературы

[1] *Bahlmann C.* On-line Handwriting Recognition with Support Vector Machines — A Kernel Approach — 2003.

[2] *Senin P.* Dynamic Time Warping Algorithm Review — 2008.

[3] *Vapnik V. and Cortes C.* Support-Vector Networks — 1995.

[4] *M. J. Castro-Bleda, D. Llorens* Improving a DTW-based Recognition Engine for On-line Handwritten Characters by Using MLPs — 2009

[5] *B. Q. Huang, C. J. Du* A Hybrid HMM-SVM Method for Online Handwriting Symbol Recognition — 2006

[6] *Красоткина О. В.* Методы учета априорной информации в задачах распознавания образов — 2000.

[7] *Татарчук А. И., Сулимова В. В., Моттль В. В., Уиндридж Д.* Метод релевантных потенциальных функций для селективного комбинирования разнородной информации при обучении распознаванию образов на основе байесовского подхода. — Всероссийская конференция ММРО-14.М. , МАКС Пресс, 2009.С. 188–191.